Are Addicts Akratic?:
Interpreting the Neuroscience of Reward

By Gideon Yaffe

Introduction

We know that human beings like things; they find some things pleasurable and others painful and with various degrees and tones. We know that human beings value things; they take certain facts to be reasons for certain actions, and they take some facts to provide greater reasons for action than others. We know that human beings want things; they are more and less motivated to pursue some things rather than others. And we know that human beings act intentionally; they make choices and form intentions (where these mental states are understood to be distinct from desires) and they engage in bodily movements that accord with their choices and intentions. We can investigate, then, the systems that give rise to likings, valuings, wantings, and intentional actions. And we have learned a great deal about the nature of these systems. A great deal is known, that is, about the psychological states involved in such systems, about the environmental and genetic factors that influence them, about the social factors that contribute to them in various ways, and, importantly for our purposes here, about the brain structures and neural transmitters involved in them.

We even know something about how these systems interact. In a substantial class of normal, non-pathological cases, each of these systems provides an input to another and thereby influences its output. A person comes to like X; the liking of X contributes to making the person value X, or recognize a reason for acquiring X that is perhaps weightier than reasons that the person recognizes for not doing so; valuing X is part of what leads the agent to want X, or to be moved to act so as to acquire X; and this

desire plays a role in the formation of a choice to pursue X and an intentional bodily movement aimed at acquiring X. There are, to be sure, quite complicated relations among inputs and outputs of the various systems that we do not begin to capture with this brief description. Sometimes, to give just one example, it is the wanting of X that leads to the valuing it, rather than the reverse; and sometimes there are complex feedback mechanisms: valuing X, you come to want it, which in turn leads you to value it even more strongly. The wiring diagram, as it were, describing the routes and gates connecting these various systems is likely to be very complicated. In fact, it is even possible that some of these systems are components of others, or have overlapping components. Further, even in normal, non-pathological cases, an input to one system can nonetheless be overridden and lead to an output discordant with it. Sometimes, for instance, we come to value what we do not like, come to want what we neither value nor like, and even come to intentionally pursue something we neither like, value, nor want. It seems likely, that is, that all of the various combinations of match and mismatch between the outputs of these various systems are possible.

I will focus here on one particular form of possible mismatch, namely, that between intentional action, particularly wrongful intentional action, and what the agent values. One form of failure of self-control—although it is not, by any means, the only sort—involves precisely such a mismatch. In this form of failure to control oneself, the agent ends up choosing contrary to what he values most and thereby chooses wrongful conduct. Let's call such conduct "akratic" conduct, recognizing that the label is sometimes used to refer to a wider range of failures of self-control than these.

I focus on this form of mismatch for two reasons. First, it has been suggested by many people, among them Richard Holton in his terrific recent book, that addicts are often subject to such a mismatch. Holton writes,

> Standardly, if someone wants something—a clever device for peeling garlic, say—and then discovers it does not work, the want will simply evaporate. It is, as we might say, *undermined*. In contrast…in cases of addiction there must be an almost complete disconnection between judging an outcome good and wanting it, or, conversely, between judging it bad and not wanting it. (Holton (2009), pp. 108-9)

Holton here suggests that what the addict values fails to match what he desires. His implication is that the addict nonetheless enjoys a match between desire and choice; he chooses what he wants, but does not want what he values, and so does not choose what he values. In short, if Holton is right, the addict is akratic.[1]

If this is the nature of addicts' failures to control themselves, then that's an important discovery that can direct further research into the causes of wrongful, not to mention self-destructive, behavior by addicts. And such research can potentially tell us a lot about how to intervene to keep addicts from hurting themselves and others. Helping someone to align what he chooses with what he values is a very different project from the project of, say, helping him to align what he likes with what he values as we do when, for instance, we try to teach someone who values truffles to enjoy them, or when we administer the recently developed "cocaine addiction vaccine" which, purportedly, prevents cocaine consumption from producing a high.[2] If Holton is right, for instance, it might turn out to be ineffective to try to help the addicted cocaine user, desperate to quit, to stop liking cocaine. The drug that stops cocaine from being pleasant to him will help him to align what he likes with what he values; he disvalues cocaine and the drug will prevent him from liking it too. But unless he is also thereby

led to align what he chooses with what he values (perhaps by aligning what he wants with what he values), he'll keep on using, if Holton is right. The drug might not intervene directly at the crucial place.[3]

The question of whether or not addicts choose in accordance with what they value at the time of action is important for another reason, as well. If, when the addict acts badly, he chooses contrary to what he values, then this fact is of immediate relevance to some forms of moral evaluation of the addict's behavior. It is relevant, in particular, to an assessment of the addict's blameworthiness for bad behavior. At least one of the things that modulates our blame of bad action, and should modulate it, is the degree to which that action is expressive of bad attitudes on the part of the agent of the act; and at least one of the reasons that we care about the attitudes of a person when he acted is because those attitudes indicate something of importance about what facts he took at the time of action to give him reason for action, and to what degree. An agent's modes of recognition, weighing and response to reasons are deep facts about the agent which, arguably anyway, are of intrinsic moral importance. What facts agents take to be reasons, and with what weights, are facts about what the agent is like in a morally crucial respect. Wrongful akratic actions are not expressive of quite the same objectionable modes of recognizing, weighing and responding to reasons that we find in those who do wrong non-akratically, and so the fact of akrasia mitigates blameworthiness. The akratic merely seems to care more about leisure than about work, or more about his own convenience than about physical harm to others. His conduct seems to be expressive of such distortions in his conceptions of what reasons he has, or what weights to give them, but is not in fact. This is not to say that the akratic is excused from blame; he is probably still blameworthy to some degree. But akrasia nonetheless mitigates blame; the akratic is less blameworthy than the otherwise

identical agent who values what is gained through wrongdoing more than what is foregone. But then if we are to allocate blame to those who deserve it, and to the degree and in the way in which they deserve it, we need to know in which category to place the addict. Is the addict acting akratically, or not? This is not the only question that the proper allocation of blame requires us to answer about the addict, but it's one of them.

But do addicts act contrary to what they value at the time of action? Or do they, instead, at least at the time of action, value that to which they are addicted more than those things that they forego in order to use? Does the heroin addict who leaves work in the middle of the day to use, knowing he'll be fired, value heroin, or the using of it, or the high that it gives him, more, at the time of action, than he values his job, or the money it pays, or the support that it provides to his family? Does the pregnant crack addict, who smokes crack, value what crack gives her more than she values the health of her unborn child, or more than she disvalues the punishment and censure that she expects her behavior to bring? Or, when she is using, is she doing violence, herself and through her own agency, to that which she values more than that which she pursues?

These are empirical questions. And, in fact, we have empirical data from neuroscience that bears on them. The problem is that the data is difficult to interpret. This paper looks at two interpretations of some of the neuroscientific data that have been offered in the recent philosophical literature: Holton's and Timothy Schroeder's. For different reasons, although on the basis of some of the same data, Holton and Schroeder reach the conclusion that addicts are, indeed, acting akratically. The paper argues that the experiments that Holton and Schroeder mention show precisely the opposite of what Holton and Schroeder take them to show. They show, that is, that addicts ordinarily act in accordance with what they value at the time of action. This is probably often temporary—many addicts, that is, value use over abstention at the

moment they choose to use, but value abstention over use moments before and even moments after. But, still, at the time of action they value what they choose. If this is correct, then addiction influences what people do intentionally by working through, rather than against, the valuing system.[4]

As we'll see in the final section of this paper, this result has implications for how a criminal defendant's addiction ought, or rather ought not, be considered when assessing his legal responsibility for a crime. It will be argued that addiction ought to be considered in a way quite similar to the way in which duress is considered in the criminal law. This claim, it will also be suggested, is consistent with denying what should be denied, namely that addicts are under duress. They are not; but their condition bears sufficient similarity to the condition of those under duress to warrant treating them similarly under the criminal law. Like victims of duress, addicts find themselves valuing criminal conduct more than they value refraining from such conduct. And like those under duress, and unlike those with such values who are not under duress, addicts have the values they have thanks to the fact that they bear burdens that are not, themselves, reflective of morally or legally objectionable attitudes on their parts.

Valuing Defined

Before moving forward, it is important to head off a possible misunderstanding concerned with the verb "to value". To value X, for our purposes here, is not merely to say or believe that X is a good thing. It is, instead, to take oneself to have reason to do those things that promote X, or bring X about, or are believed to be necessary or even just useful to promoting X or bringing X about. To value X more than Y is to grant

greater weight to the reasons one takes for promoting X or bringing X about than one grants to the reasons, if there are any, that one takes there to be for promoting Y or bringing Y about. To value something, in the sense in which the term will be used here, then, is to engage in a mode of recognition, weighing and response to reasons. To believe that something is a good thing often, maybe even always, goes along with valuing it; but it is nonetheless distinct. In so far as it is conceptually possible to believe something to be a good thing while failing to recognize reasons for promoting or bringing it about, believing that something is good and valuing it are distinct.

One's mode of recognition, weighing and response to reasons is a function of the way in which one consciously deliberates, or of the way in which one would consciously deliberate in circumstances in which one does not. Say that one believes that a particular act would promote X; one believes, for instance, that boycotting British Petroleum, in contrast to Chevron, will promote the use of alternative sources of energy. Perhaps BP, unlike Chevron, actively lobbies against the expansion of research into alternative energy sources. To treat the fact believed as giving one reason to engage in the act is to deliberate in a way, or to be ready to deliberate in a way, which involves treating the fact that X would be promoted by the act as a reason to engage in that act. Someone who has the belief and who values the use of alternative sources of energy will deliberate, or be ready to deliberate, in a way that treats the fact that the boycott of BP would promote such use as a reason to boycott BP. It may be a reason that is outweighed by others; but it is still a factor given weight in deliberations about what to do, or would be were the agent to engage in relevant deliberations. By contrast, someone who believes that the boycott would promote the use of alternative energy, but does not value the use of alternative energy sources will not grant the fact believed any weight in his deliberations about what to do. When deciding whether to fill his

tank at BP or Chevron, he will consider and weigh a variety of reasons, but the fact that boycotting BP would promote alternative energy will not be among them.

To grant weight to certain facts in one's deliberations involves, among other things, being ready to recognize a failure to treat the fact as giving reason as involving an error in one's own deliberations. A mark of valuing, that is, is the acceptance of norms governing deliberation that would not be accepted by someone who did not value in the same way. To continue with our example, consider the person who values the use of alternative sources of energy, and is deliberating about whether to go to BP or to Chevron to fill up his car's tank. If he ignores the fact that boycotting BP will promote the use of alternative sources of energy—he deliberates as though he granted that fact no reason to go to Chevron over BP—this will be a failure in his deliberations when they are held up to his own standards. Since he values the use of alternative sources of energy, he ought to take the fact that boycotting BP will promote it as a reason to shop at Chevron rather than BP. This "ought" applies to him and not to others who do not value as he does.

It is quite possible that the verb "to value" is used here as a term of art. Nothing is invested in the claim that there is perfect, or even approximate overlap between one's modes of recognition, weighing and response to reasons and ordinary usages of the term "to value". Perhaps they do not align, even if they do overlap. One reason to think that they diverge is that in some ordinary senses of the term "valuing", what a person values cannot be a local property of him, the possession of which at one time entails nothing about his properties at other times. In some ordinary usages of the term "valuing", that is, you cannot value something at one time unless there is a substantial period of time over which you value it. There are no fleeting valuings, in this sense of the term. By contrast, nothing in the way the term is being used here implies that. You

may employ a mode of recognition, weighing and response to reasons at one particular time and not employ that same mode at any other time at all. Valuing, in the sense in which the term is used here, could be a local property. This isn't to say that it typically is local. Typically, what one values at one time, one values at many other times, too. But this is not an entailment, but, instead, a contingent fact about most people.

A question arises, however, whether, in assessing people's blameworthiness, we are concerned with the local property for it's own sake—with, merely, the person's modes of recognition, weighing and response to reasons at the moment of action—or are concerned with the local property only because we assume it to be stable over time. Perhaps, that is, we think it more blameworthy to act harmfully and wrongfully while granting little reason-giving weight to the fact that one's act would be harmful to another only because we assume that a person who failed to grant sufficient rational weight to such a consideration at the moment of action fails in this regard generally. Perhaps, that is, valuing is important to blame only because of what it says about character, where one's character is understood as involving stable tendencies to recognize and respond to reasons in a particular way.

I suspect that sometimes what an agent values at the time of action matters for its own sake and sometimes it matters because of what it tells us about the agent's character. Judgements of blameworthiness are a diverse lot and there is little reason to expect uniformity in this regard. However, the criminal law, if not morality, has a particular concern with the very moment at which a defendant chose to commit a crime, and thus with the defendant's values at that moment, to the exclusion of other times. For the purposes of the criminal law, we care about the moment of action to a greater degree than moments before and after it for several reasons. One of the most important is that we typically don't allow prosecutors to bring in evidence about the defendant at

other times unless it can be shown to say something about him at the moment of action. This is a fundamental principle of criminal law in a liberal state that underlies our practices of, for instance, excluding prior convictions from evidence except in special circumstances. We convict only for criminal conduct performed at a particular time and not for other features of the agent or his conduct that surround the conduct, but are not specifically proscribed by the state. But if the criminal law has good reasons to base its assessments of defendants on facts about them only at the time of criminal conduct, then the notion of valuing at work here, where valuing can be local, is potentially of importance to the criminal law, even if we grant that temporally distributed properties operative at the time of action are also potentially important.

In short, it is possible that the question of whether addicts act akratically is of relevance to only a subset of our judgments of blameworthiness; but, still, it is of importance to a large percentage of our judgments of criminal blameworthiness. If we are to treat addicts who commit crimes justly under the criminal law, if we are to blame them in a way and to a degree that matches their blameworthiness, we need to know if they act akratically or, instead, act in line with what they value at the time of action.

Holton on Berridge and Robinson

Holton reaches the conclusion that addicts act contrary to their values largely on the basis of the well-known experiments on rats conducted by Robinson and Berridge and colleagues. Robinson and Berridge showed that amphetamine addicted rats pursue a sugar reward far more zealously than do rats that are not addicted, despite the fact that the two groups of rats do not differ in the degree to which they like the sugar, at least by behavioral measures of liking. The addicted rats press a lever on hearing a tone

that has been associated with sugar reward four times more frequently than unaddicted rats. Thus the addiction seems to increase the desire for sugar that is prompted by the conditioned cue, and in turn increases the frequency of choices to consume sugar, despite the fact that the addiction has no effect on the degree to which sugar is enjoyed. The addicted rat receives no more pleasure from the sugar it zealously pursues than the unaddicted rat. But, still, it pursues it much more aggressively. Put in the terms in use here, addiction seems to increase desire, or want, but does not increase liking. The experiments do not show exactly how it is that addiction influences choice. But they do show that it does not do so by increasing the degree to which a substance is enjoyed. Its influence is on some system other than the system that gives rise to pleasure and pain.

In addition, the Robinson and Berridge experiments provide powerful evidence that this important behavioral change is linked to the dopamine system in rats. Amphetamines, as well as many other drugs of abuse, are known to cause immediate release of dopamine, and there is good reason to think that, after the cue has been associated with the reward, addicted rats, even when not treated with amphetamine at the time of the cue, have greater dopamine release on encountering the cue than do unaddicted rats. Thus it appears that dopamine release plays an important role in the way in which the liking system modulates the outputs of the wanting system. The addicted rat is motivated to pursue the reward <u>as if</u> he liked it much more than the unaddicted rat likes it, despite the fact that he does not, in fact, like it any more at all. Addiction seems to weaken the normal connection between the liking system and the wanting system, and it seems to do so thanks to the way in which dopamine signals in addicts differ from those in unaddicted subjects. Since drugs of abuse have both temporary and long term effects on dopamine release, they also have temporary and long term effects on the way in weakening the way in which liking modulates wanting.

What do these startling results show about human addicts? Since, as Holton notes, rats probably don't value anything—they do not recognize facts as constituting reasons in favor of certain courses of action and grant them weight in deliberation—the Berridge and Robinson experiments, as important as they are, do not speak directly to the question with which we are concerned here. To say that the wanting and the liking systems are disconnected from one another in addiction is not to say that the valuing system is disconnected from the wanting system. It is perfectly possible that while the addicted human being wants the drug much more than he likes it, he still comes to value it in a way that comports with his degree of desire for it. Holton thinks this unlikely largely because of the undeniable fact that addicts often pursue drugs in ways that directly conflict with what they judge to be good, or take themselves to have most reason to promote. The addict who prostitutes his child for drugs does not think this is a good thing. This is true. But what truth, exactly, does it register? The addict, to be sure, recoils at the prospect of selling his child for drugs when he is not craving, or is not in the presence of cues that prompt use; and he suffers powerful regret after having done this. His judgement that such behavior is an unqualified evil is real and is held by him for a much larger percentage of his time than any competing judgment. But it does not follow from this that, <u>at the time of decision</u>, he does not judge it best, overall, to sell his child; at the time of action, he may judge that to be the best of a number of bad options. The sense in which he acts contrary to what he judges to be best may just be that he acts contrary to what he <u>usually</u> judges to be best, and what he in fact judges to be best both before and after the time of action. But, still, his attitudes at the time of action matter and we have, as yet, no reason to believe that he does not, at that time, judge it less bad to sell his child than to go without the drug to which he is addicted. He may be locally, although not globally, just like the unrepentant child pimp.

How could we settle this question? How could we determine if the addict values what he chooses at the moment he chooses it, or values something else instead? In fact, we have a tool for making progress on this question already for much is known about the information that is carried by dopaminergic activity, which is what appears to be disrupted by drugs of abuse and addiction. As I will suggest, when we reflect carefully on what is known about the information carried by dopaminergic activity, we will see that disruptions of the dopaminergic system are, in human beings anyway, disruptions in the valuing system, and not just in the wanting system. If this is right, then precisely what we learn from experiments like Berridge and Robinson's is that addicts value, at the time of action anyway, precisely what they choose. I explain.

A crucial question is what the dopamine signal represents. Much of the most important work bearing on this question has been done with monkeys. In well-known experiments done by Wolfram Schultz and colleagues, for instance, dopamine is measured in monkeys at the time of a light cue, and at a time moments later when a reward is delivered.[5] There are important differences between the dopamine signal in the monkeys initially and the signal after the monkey has learned to associate the light with the reward. Initially, before the monkey has learned to expect a reward on seeing the light, the dopamine signal goes up when the reward is received. But after the monkey has learned that the light precedes reward, dopamine goes up when the light appears, and remains flat when the reward is obtained. In short, the signal increase moves from the time of the reward to the time of the cue, due to learning. (This fact by itself shows that the dopamine signal is not a measure of something like pleasure since the cue is never pleasurable and the reward always is.) Further, after the monkey has learned to associate the cue with the reward, when the light appears and no reward is given, the monkey's dopamine signal goes up initially on seeing the light, but goes

down relative to its baseline when the monkey realizes that it will not receive a reward. That is, after the increase has moved from the time of reward to the time of the cue, the monkey shows a decrease in dopamine at the ordinary time of the reward when it does not receive a reward at that time.

What do these results show? A plausible explanation, which is the explanation favored by Schultz and his colleagues, is that the dopamine signal is modulated at least in part by the monkey's expectations. Initially, when the monkey does not expect the reward on seeing the light, it is the receipt of the reward, and not the appearance of the light, that shows the monkey's condition to be better than it expected it to be; and so the dopamine signal goes up on reward receipt. Later, when the monkey has learned to associate the light and the reward, it finds on seeing the light that its condition is better than expected, and so, again, the dopamine signal goes up on seeing the light. But, since the appearance of the light resets the monkey's expectations—having seen the light it now expects the reward—the dopamine signal remains flat when it receives the reward, and its expectations are met. And the signal goes down when it does not receive the reward, and things turn out to be worse than it expects thanks to its conditioning.

The last of these results is worth highlighting. It is widely believed, and not without reason, that the dopamine signal plays a role in the generation of choices. Decisions about what to do, that is, are influenced by the dopamine signal. But it is important to see how this comes to pass. It appears that the way the dopamine signal influences choices is by influencing future expectations which then, in turn, influence choices. The dopamine signal represents something about the past; it represents the difference between how things actually came out and how they were expected to come out. But it influences future decisions by influencing expectations about the future. If

things came out less well than expected, then future expectations ought to be different from past expectations, and the dopamine signal provides a guide for determining how different they should be. The larger the dopamine signal, the more need there is for having different expectations about the future when conditions are otherwise the same as they were at the time of the last prediction. The reason the primed monkey's dopamine signal goes down relative to the baseline when it is not given the reward following the cue is that it expects the reward thanks to the dopamine signal's representation of the world as just as expected on receipt of reward on the previous trials. The dopamine signal, then, is a representation of a fact about the past which influences future decisions by altering the subject's expectations about the future. What this implies, among other things, is that in a healthy animal that learns quickly from its mistakes a large dopamine signal will result, later, in flat dopamine signals in response to exactly the same experiences. The large dopamine signal will help the subject to learn to have expectations that match reality. And when expectations match reality, the dopamine signal is flat.

An important question remains when we accept that the dopamine signal represents the difference between expectation and reality: A difference in what respect? There is a powerful pull towards answering that the dopamine signal represents a difference in the value expected and the value actually received. In fact, it is known that dopamine signals are unaffected by discrepancies in neutral differences between expectation and reality [[refs]]. When a monkey gets something different from what it expects, but of equal value, the dopamine signal remains flat. [[true?]] But we should be careful about characterizing the dopamine signal as representing a difference in value. What the experiments show is only that the dopamine signal represents a

difference in something with the following property: more of it is better than less. To see the point, consider two competing hypotheses:

> The Likability Hypothesis: The dopamine signal represents the difference between the amount that the subject expects to like something and the amount that the subject actually likes it.

> The Desirability Hypothesis: The dopamine signal represents the difference between the amount of desire satisfaction that the subject expects to receive from acquiring something and the amount that he actually receives.[6]

The data just described—under which the monkey's dopamine signal varies with its expectations—does not allow us to discriminate between these two hypotheses. If both how much organisms like things, and how much they expect to like things, on the one hand, and how much desire satisfaction they experience on acquiring something, and how much they expect to experience, on the other, influence what choices they make, then both hypotheses are consistent with the data. And, importantly for our purposes here, given subsidiary assumptions of this sort, the data is also consistent with the following hypothesis about the dopamine signal in humans:

> The Value Hypothesis: The dopamine signal represents the difference between the amount that the subject expects something to be supported by reasons and the amount that it is actually supported by reasons.

How can a person's expectations diverge from reality in this respect? The obvious way this happens is when the facts turn out differently from expected. I expect the lock to turn when I insert the key and it doesn't, so I recognize myself to have had less reason to insert it than I expected to have. I had reason to turn the lock, and falsely expected the key to turn it. But there can be a discrepancy in expectation in this respect in other ways too. Notably, there might be a shift in my standards about what counts as a reason between the moment of expectation formation and the later moment. If, after I insert the key, I come to the view that turning the lock is not worth doing, but do not update my expectation in light of this change in view, then I will be disappointed in my expectations about what reasons I have when I find that the key turns the lock. Or, there can be a shift in the way in which I weigh reasons between the moment of expectation formation and the later moment. I expect myself to have more reason to turn the lock than to wait for the door to be answered, but my ordering of these two options swaps after I have inserted the key. Assuming that the swap in my attitude does not lead me to update my expectation, perhaps because there isn't time, I will find myself to have had less reason to insert the key than I expected.

If how much a human subject expects available courses of conduct to be supported by reasons influences his choices, then the Value Hypothesis, too, is consistent with the data. We know that the dopamine signal represents the difference between what is expected and what is encountered. And we know that when the dopamine signal represents there as having been more of that, whatever it is, then the subject will more zealously pursue outcomes that are like those encountered; the signal is representing what was encountered as <u>better</u>, in some respect, than what was expected. But we don't know <u>in what respect</u> the expected and the actual world are represented as different by the dopamine signal.

What likability, desirability and support by reasons have in common is that more of them is better than less. What the data about the dopamine signals shows is only that dopamine represents the difference in expectation and reality of something more of which is better than less. Let's call that something "X". X might be likability, desirability or value; or perhaps something else entirely more of which is better than less. Further, what the Berridge and Robinson experiments show is that addicts choose as if they judged drugs to have a lot of X, under the assumption that what subjects choose is powerfully influenced by such judgements. And they do so thanks to the fact that drugs of abuse disrupt the dopamine signal.

Now, this much is clear: addicts do not act contrary to their judgements about how much X their acts promise. Precisely their problem is that they act in line with such judgements in circumstances in which those judgements themselves have been adversely affected by the disruptions of the dopamine signal from which addicts suffer. Why does the addicted rat press the lever so much more frequently than the unaddicted rat? It's not because he actually likes sugar more than the unaddicted rat. Nor is it because sugar gives him more desire satisfaction than it gives to the unaddicted rat. It's because the rat judges sugar to promise more X than the unaddicted rat judges it to promise. His judgement is different from the unaddicted rat's because his dopamine signal represents pressing the lever as far better, with respect to the amount of X it promises, than it was experienced as being last time he pressed it. But this last fact is consistent with, and in fact explains, the further fact that he represents pressing the lever as promising more X than any alternative. He judges it of his alternatives to be X-optimal, as it were, and that's why he pursues it so zealously. He seems, in fact, to judge it to be about four times better, with respect to X, than the unaddicted rats judge it to be.

To say that X must be something more of which is better than less is not to imply, all by itself, that in representing an act as promising more X than expected the organism is representing the act as better than expected. More vitamin C is better than less; but in representing the orange before me as containing more vitamin C than I expected it to contain, I am not necessarily representing it as better than expected. For that to be the case, I need to represent vitamin C as good in some respect. A dispassionate chart indicating the nutritional value of foods does not represent nutrient rich foods as better than nutrient poor foods; it merely represents how many nutrients the foods contain and leaves it to the reader of the chart to draw his own conclusions about which foods are better from the point of view of nutrition. What would allow us to distinguish between a mental representation of the orange as containing q miligrams of vitamin C and a representation of it as worthy of pursuit in virtue of the fact that it contains q milligrams of vitamin C? The answer is that the latter representation would play a role in both conscious deliberation and in the guidance of behavior that the former does not. Someone who represents q milligrams of vitamin C as worth pursuing will judge himself to have a reason to eat the orange, and will be motivated to pursue it. In short: it is thanks to its causal role in an agent's psychology that a representation of a fact as reason-giving is distinguished from a representation of that fact that does not represent it as reason-giving.

Now, a great deal is known about the causal role in our and animal's psychology played by the dopamine signal. As predicted by computational models of learning, so called "reinforcement models", biological organisms like human beings update their judgements in response to their calculations of the difference between expected and achieved X and they both deliberate and act in a way which accords with their updated judgements.[7] But, and here is the important point, the dopamine signal can play this

role only if the organism represents things that are X as worth pursuing on those grounds. The dopamine signal itself needn't represent things that are X as things that there is reason to pursue on those grounds. It may represent only the fact that a particular thing promises more (or less) X than expected; it may be like the nutritional chart that represents the amount of vitamin C in various food. But, still, a creature whose dopamine signal plays the crucial role in learning that the dopamine signals of rats, monkeys and human beings plays must also represent things as worth pursuing in virtue of the fact that they are X. Such organisms must represent X as reason-giving. Given the information that the dopamine signal carries, that is, it must be part of the valuing system.

Is the claim just made consistent with the suggestion that animals like rats and monkeys, in contrast to human beings, do not think about what reasons they have for and against particular courses of action? Notice that there is one very important difference between the dopamine signal in animals like rats and monkeys and the dopamine signal in human beings. We have no reason to believe that rats and monkeys update conscious judgements about how much reason they have to pursue particular objects or courses of conduct in response to the dopamine signal. Because their psychologies are (probably) not as rich as ours in this respect, their dopamine signals play a relatively impoverished role in their psychologies compared with ours. Our dopamine signals lead us to update both conscious judgement and action-guiding preference, while theirs, since they probably lack the kinds of conscious judgements that we have, lead only to the second kind of updating. Whether both roles are required for a representation to count as a representation of a course of conduct as supported by reasons, is a hard question. But, thankfully, it's not a question that we need to answer, for this much is clear: in human beings, conduct that makes sense in

light of the dopamine signal makes sense in light of the person's values at the time of action. The dopamine signal represents something more of which is better than less, and thanks to the role that that representation plays in human psychology, it must be part of a system that represents acts, objects and states of affairs as worth pursuing or avoiding. The result: addicts act in accord with, rather than in opposition to, their values at the time of action. Addicts do not act akratically.

Schroeder on Desire and the Reward System

It helps to understand the argument just offered to contrast the position outlined here with Tim Schroeder's position as expressed in his recent book, and in an even more recent paper on addiction.[8] At least for the purposes of argument, Schroeder holds an instrumental conception of rationality. Under such a conception, what one has reason to do is a function of what one desires. If there are no desire-independent reasons, then the question of whether addicts act in accord with their desires seems directly relevant to the question of whether addicts act akratically, or contrary to their modes of recognition and response to reasons employed at the time of action.

So, if an instrumental conception of rationality is correct, and if it turned out that addicts frequently act contrary to their desires in the sense that they act contrary to the reasons that their desires supply, then that fact would turn out to be of relevance to our discussion here. Schroeder has argued that a proper appreciation of what is known about the effect of drugs on the reward system in humans supports the claim that addicts actually act contrary to their desires. If we think through what the neuroscience of reward means, we will see, thinks Schroeder, that addicts actually want that which they choose far less than it might appear.

Schroeder's argument takes the following theory of desire as a premise:

> The Reward Theory of Desire: "[T]o desire something is for one's reward system to treat it as a reward. And for one's reward system to treat something as a reward is for the reward system to take representations of that thing as positive inputs into a calculation of how many rewards the world contains versus how many it was expected to contain." (Schroeder (2010), p. 395)

On this theory, it is not the case that everything a representation of which causes an increase in the dopamine signal is desired. Sometimes a representation of something will cause an increase in the dopamine signal even though the organism's reward system does not cause that increase. This is what Schroeder thinks takes place when one consumes a drug of abuse. The drug of abuse, not the reward system, causes an increase in dopamine. So the person who wants the drug and has a representation of it, on the one hand, and the person who does not want it but consumes it, on the other, have something in common: both have an increased dopamine signal in response to a representation of the drug. But they are also different: the first, and not the second, has a reward system that responds to a representation of the drug with an increase in the dopamine signal.

What does Schroeder take the dopamine signal itself to represent? He takes it to represent the degree to which the drug is desired. That is, where the object of desire is a thing in the world—a drug, say—that which is represented by the dopamine signal is actually a state of the organism: the state of desiring something in the world. So, when the dopamine signal is driven up by something other than the reward system's processing of a representation—when it is driven up, for instance, directly by

consumption of a drug of abuse—the dopamine signal actually misrepresents: it represents the drug as strongly desired, when, in fact, it may not be desired at all. The result, Schroeder holds, is that in so far as an addict's conduct is dictated by his dopamine signal, he frequently pursues things as though they were strongly desired when, in fact, he does not desire them at all. But if he acts contrary to his desires, and if the reasons that he has are constituted by what he wants, then it follows that he acts contrary to what, given his desires, he has most reason to do.

While this seems relevant to the question that concerns us here (namely, whether the addict acts akratically), it is not clear precisely what it implies in that regard. While Schroeder has provided us with a theory of desire, he hasn't provided us with a theory of valuing, and that's what's crucial to our question. If a person lacks a desire, but thinks he has one, and then acts as dictated by the desire he thinks he has, is he acting contrary to, or in line with, what he values? Put another way, is conscious deliberation responsive to our desires, or to what we think our desires are, or both, or neither? While I cannot claim to have an argument for this answer, I think we ought to hold that under a desire-based conception of reasons for action, a person takes himself to have a reason to A only if he has both a desire that is supportive of A-ing and a belief that he has a reason to A (which may be grounded in a belief that he has a desire that gives him a reason to A). Since, if Schroeder is right, the addict whose dopamine signal is driven up directly by the drug of abuse, and not by his reward system, lacks the desire, he also fails to value that which he pursues. He would not, that is, deliberate as though he took there to be reason to do those things that the desire he thinks he has provides. The result: if Schroeder is right, and if the accompanying theory of valuing just offered is adequate, then the addict acts akratically. He acts contrary to what he values, because

he acts contrary to what he wants, and wanting is necessary for valuing given the desire-based view of reasons for action.

It is important to note a feature of Schroeder's position that might not be obvious. In saying that the dopamine signal represents the degree to which something is desired Schroeder is not overlooking the evidence supportive of the claim that the dopamine signal represents the difference between expected and actual X associated with some object. Schroeder's idea is that the dopamine signal ordinarily covaries with two things: the difference between expected and actual X associated with an object, and the degree to which the reward system, when given a representation of the object as an input, causes the dopamine signal to go up. The dopamine signal covaries with the latter of these two things because, typically (although not when drugs of abuse are involved), the dopamine signal goes up because the reward system takes a representation of the object as an input and drives up the dopamine signal. Since the reward system's tendency to drive up the signal in response to a representation of the object is, under the Reward Theory of Desire, what it is to want something, Schroeder takes the dopamine signal to carry information about, or represent, the degree to which the object is desired.

To illustrate the point with an analogy, consider a camera that is used to take a photograph of an apple. The photograph represents the apple; but it also carries information about the camera. In particular, it will typically carry information about what kinds of representations the camera will produce when used in the normal way—namely, perspectival visual representations of those things at which its lens is pointed when the button is pressed. The photograph carries information of this sort because it is produced by the camera and has various properties—it is, in particular, a perspectival visual representation of that at which it was pointed when the button was pressed

(namely, an apple). The dopamine signal represents the difference between actual and expected X associated with some object, but it also carries information about the reward system since it is typically created by the reward system. In particular, it carries information about how the reward system responds to certain kinds of inputs; and since what it is to desire something, Schroeder thinks, is for the reward system to respond in a certain way to certain kinds of inputs, it follows that the dopamine signal carries information about the organism's desires.

The analogy to the camera helps us to understand Schroeder's position, but it also helps us to see one of the problems with it. Imagine that I typically take pictures with a Nikon camera. I think the camera needs adjusting, so I bring a stack of pictures taken with it to the camera doctor so that he can diagnose the problem. Trouble is that I mistakenly include in the stack a picture of an apple taken with a different camera, a Canon. Does the picture taken with the Canon carry information about, or represent anything about, the Nikon? Of course not. The camera doctor might think it does because he thinks it was taken with the Nikon; but he's mistaken. Since the picture was taken with the Canon, it only carries information about the Canon, no matter what the camera doctor happens to think. Now consider the dopamine signal driven up directly by some drug of abuse. That signal is not the product of the reward system. Does it carry information about the reward system? No. In the first instance, it is a representation of the difference between actual and expected X analogously to the way in which the Canon photograph is a representation of an apple. We can grant that representations often carry information about the systems that created them. But since, according to Schroeder, it is not the reward system that created the dopamine signal when it is driven up directly by a drug of abuse, the signal does not, in such cases, carry information about the reward system. And, given the Reward Theory of Desire, it

follows that the dopamine signal in such a case is not a representation of the subject's desire. But if the dopamine signal in such cases does not represent the subject as desiring its object, then nor does it misrepresent that. Even when the dopamine signal is the product of a drug of abuse, the subject probably has representations of the degree to which he desires the object the signal represents; but the signal itself is not a representation of that desire.

While I believe that this is a problem with Schroeder's position—even given the Reward Theory of Desire, the dopamine signal is not a representation of the degree to which an object is desired when it is driven up directly by a drug of abuse—it is not a problem that interferes with the conclusion that interests us, namely that addicts act akratically. So long as addicts really do act contrary to what they desire, then they act contrary to what they value; that conclusion can be reached without appeal to the claim that the dopamine signal in addicts represents the degree to which the object is desired. However, there is a more serious problem for Schroeder's position, and reflection on it suggests that his view fails to imply that addicts act akratically.

To see the problem, start by recalling that there is data supporting the idea that the effects of cues on the dopamine signal are found in heavy drug abusers even when they do not consume the drug. The amphetamine addicted rats in Robinson and Berridge's experiments, for instance, press the lever for sugar much more frequently than unaddicted rats even when they are deprived of amphetamine. In such cases, the addicted rats would appear to represent the sugar—such a representation, after all, is what their perception of the tone causes in them—and then must suffer an increase in dopamine in response. Assuming that Schroeder wants to hold that the rats in such a case do not desire the sugar as much as they are moved to get it, he must hold that the effect of the representation of the sugar on the dopamine signal bypasses the reward

system. He cannot hold, for instance, that the reward system is itself altered by heavy consumption of drugs resulting in an increase in the dopamine signal caused by the reward system when a representation of sugar is given as an input. If he were to hold that, then he would be holding that addicts actually want cued rewards more than non-addicts want them, which is just what he hopes to deny. Why is this a problem for Schroeder? The reason is that to make this assertion is to overlook the most plausible explanation for the long term effects of drugs of abuse on the dopamine signal.

We know that drugs of abuse drive up the dopamine signal directly. And, from independent experiments, we know that the dopamine signal encodes the difference between actual and expected X, where X is something more of which is better than less. But, further, we know from formal models of machine learning how a representation of the difference between actual and expected X helps an organism to learn: that representation serves as an input to the very system that generates expectations of the amount of X promised by a particular course of conduct. In other words, the representation of the degree to which the organism was mistaken last time in his expectation informs his next expectation. If he underestimated how much X a particular course of conduct promised last time, then, thanks to the fact that he has a representation of his degree of error, he will increase his estimate next time. The dopamine signal, that is, alters the way in which the reward system functions the next time it is fed a representation of the very thing that drove up the signal last time. But, if the Reward Theory of Desire is correct, it follows that the dopamine signal influences behavior by changing the organism's desires. The rat has a representation of amphetamine to come and expects it to promise a bit of pleasure. When he consumes it, the drug drives up the dopamine signal, which thereby represents the drug as actually yielding far more pleasure than it was expected to yield. Then, when the rat has a

representation of amphetamine again, he uses this earlier dopamine signal to "learn" that he had it wrong last time, he underestimated the amount of pleasure that amphetamine promised. The result is that his reward system responds with an expectation of a greater amount of pleasure. So, thanks to the way in which the drug drove up the dopamine signal, there has been a change in the way in which the reward system responds to a representation of the drug. It represents it now as more rewarding than it represented it as being prior to consuming it in the first instance. Thus, under the Reward Theory of Desire, the effect of the drug on the reward system has caused the organism to want the drug. So, when we appreciate the role that the dopamine signal plays in the reward system—it is an output now and an input tomorrow—we can see that under Schroeder's own theory of desire, the opposite conclusion from the one he reaches is the correct one. Things are just as we would have thought they were before any fancy theorizing: addicts want drugs and that's why they pursue them. Therefore, Schroeder's remarks ought not to lead us to think that addicts act akratically, even if we accept that all reasons for action derive from our desires.

Legal Implications

What lesson, if any, should the criminal law take from the fact that addicts do not act akratically? It is important, in thinking about this question, to separate the question of what should be done with addicts who commit crimes, generally, from the question of what should be done with addicts who commit the crime of drug possession. Although no shortage of addicts find themselves in trouble with the criminal law solely because they possess illegal drugs, there are also no shortage who, instead, commit other, independent crimes—sometimes violent crimes—solely so that they can feed

their addictions. Possession of drugs is almost an intrinsic feature of addiction; what drug addict has never possessed drugs? And so, it can seem that when we criminally punish an addict for possession we are, in the end, merely punishing him for being an addict. That is little different from punishment on the basis of status, a practice that is abhorrent. The point for our purposes, however, is this: whatever objections one might make to punishing addicts for possession are independent of the question of whether addicts are akratic. Criminal behavior performed in order to come into possession of drugs, or in order to make use possible, is importantly different from possession itself. Such behavior is not an intrinsic feature of addiction and punishment of it is not punishment for status. Still, it can seem as though the fact of addiction is relevant in such cases. Our question, then, is what the criminal courts should do with addicts who have committed crimes other than possession. Should the courts treat two people who do the same bad thing for the same reasons in the same circumstances differently when one, but not the other, is an addict? Compare, for instance, two people both of whom leave a young child in a very hot car for two hours so as to buy drugs from a dealer inside a nearby house. Should the fact that one is addicted, and the other merely wants the drugs for recreational use, matter to the assessment of their criminal responsibility for child endangerment? Does the fact that addicts are not akratic help us to answer questions of this kind?

To see how it helps, start with a distinction between two different kinds of reasons that might be given for less severe treatment of a defendant with feature F in comparison to a duplicate differing only in that he is not-F. First, we might claim that thanks to the fact that he is F the defendant lacked the power to comply with the law, or had a severely diminished ability to do so. A delusional and religious defendant who believes that God has ordered him to kill is capable of compliance with laws against

murder only to the degree to which he is capable of mustering the courage and fortitude to defy a divine command. Alternatively, we might cite some burden that the defendant would bear, thanks to the fact that he is F, that made compliance with the law particularly costly for the defendant. A delusional defendant who believes that God will punish him severely if he does not kill would, he believes, have to bear such punishment were he to comply with laws against murder. Or, more prosaically, a defendant who robs a bank when the mob threatens to kill his child if he does not, would, he believes, bear a severe burden—namely the loss of his child's life—were he to comply with the law. Call features of defendants that ground an argument in their favor of the first sort "Can't Factors" and features of the second sort "Can't-be-Expected Factors". To cite a Can't factor is to say that the defendant's path to compliance was blocked (or he believed it to be). To cite a Can't-be-Expected factor is to say that the defendant's path to compliance would involve severe hardship on the defendant's part (or he believed it would) of a sort that he cannot be expected to suffer. As the two examples of delusional defendants with deific visions indicate, a single factor—e.g. that the defendant has deific visions—might be a Can't factor, or might be a Can't-be-Expected factor, depending on the details. The person who cites the factor, that is, might take it to be relevant because it diminishes or eliminates the defendant's power to comply or might take it to be relevant because it results in compliance involving severe burdens that the defendant should not be expected to bear in order to comply.

One set of Can't factors produce a mismatch between conduct and the defendant's values; they produce akrasia and do so inevitably. In such cases, what the agent cannot do is to guide his conduct in accordance with his values; something else determines what he does. Consider someone, for instance, who kills someone he loves while in a rage. What he does—namely kill another—fails to align with what he values,

we can imagine, even at the moment that he does it. But his rage takes over. In this case, the wanting system controls conduct thanks to the influence of emotion, and bypasses the valuing system. Thanks to his rage, that is, the agent is motivated to kill and does so, despite the fact that he does not value killing. We would need to know much more about the case before we could know whether the defendant's rage, understood as a Can't factor, actually diminishes responsibility. If the defendant is a hothead, then it does not; if he had a good reason for being so angry, then perhaps it does diminish responsibility, as under the law of manslaughter, even if it does not eliminate it entirely.[9] But the point for our purposes isn't whether such a factor diminishes responsibility; the point is that if it does so it does so in part because it inevitably induces akrasia: it causes conduct that is in violation of the law, is not valued by the agent, and in circumstances that make it impossible for the agent to have avoided akratic conduct.

So, what we learn from the fact that addicts are not akratic is that addiction is not this kind of Can't factor. Perhaps it is another kind. Nothing said here rules out that possibility. But it seems more likely that it is, instead, a Can't-be-Expected factor. Addiction is relevant to responsibility, that is, because for addicts to comply with the law they must bear burdens that unaddicted duplicates need not bear, burdens that suffice to make it inappropriate, or less appropriate, to hold them responsible for bad behavior.

What burdens must they bear? As I want to briefly suggest, the fact that addicts are not akratic helps us to identify the particular burden that compliance with the law would require them to bear. The key is to recognize that addicts are inevitably subject to periods in which they will value violating the law over complying with it. When they are in such a state, compliance with the law would require them to act contrary to

the dictates of their valuing system. The reason that they violate the law is, precisely, that they are built, as we all are, to avoid suffering this burden; we are built to, as much as possible, guide our conduct in accordance with our conception of our reasons for action. And a substantial burden it is, for part of what it is to be a fully functioning, autonomous adult, is to act in accord with what one values. To have to give that up in order to comply with the law would be no less than to have to give up part of what makes one a citizen of a state equipped to be the target, not to mention the beneficiary, of exercises of state power.

Consider an example. D has a legal obligation to deliver his child to his former spouse by 3:00 PM on Saturday following his weekly court-approved visit. He's an addict and he's craving and he knows that if he drives himself and his child to his former spouse's home, he will come to judge that he has greater reason to stop to use than he has to deliver the child on time. The result will be that he will not deliver the child on time and will violate the law. What is he to do? Notice that he can comply with the law. He merely needs to ask a friend to drive. But if he asks a friend to drive, then during the course of the trip there will be a time when he is doing something—foregoing use—that he values less than the alternative. He can anticipate, that is, that compliance with the law will come with a price: akratic action. Now imagine that D does not ask the friend to drive and so violates the law. He stops for a hit and so delivers the child late, in line with his values at the time of use, although not in line with his values at earlier and later times. Is the fact that he is addicted relevant to his responsibility for this failure? The answer is "yes". He should be treated less harshly than a non-addict who is late because he stops for a hit. D, unlike the non-addict, would suffer the burden of engaging in akratic conduct were he to take the path to compliance that was available to him (namely, having a friend drive). Now, we can

debate how much of a break D warrants for his bad behavior given this burden—perhaps a great deal, perhaps very little—but the point is that it is that debate that we must have if we are to make progress on the question of the relevance of addiction to legal responsibility. The discovery that addicts are not akratic when they act badly helps us to see, then, what burdens they would need to bear in order to act as the law demands. Often, they would have to bear the burden of performing akratic, non-autonomous action; they would have to bear the burden of taking steps to bypass their own valuing systems and give control of their conduct over to something else. What exactly this means, in practice, about how addicts are to be treated under the criminal law is hard to say. To determine that we need a theory of the degree to which a burden associated with compliance ameliorates responsibility for non-compliance. Lacking such a theory, we must settle at this point for less: we now have a better idea of what question needs to be answered.

Conclusion

Our attitudes towards addicts are deeply ambivalent. To have an addicted friend, or family member, who (inevitably) acts badly and harmfully is to find oneself torn between the conception of his behavior as a symptom of a disease, on the one hand, and as a sign of distortions in his fundamental values, on the other. It is to be torn between pity and resentment. What has been suggested here is that both reactions are appropriate. Addicts should be resented for their bad behavior. The addicts who hurts another to feed his addiction typically cares more, when he acts, about himself than he cares about the injury he inflicts. Addiction influences behavior not by bypassing what the addict cares about, but, instead, by influencing it and shaping it, at

least over short periods of time—short, but long enough to lead to very bad behavior. This is one of the things that we learn when we recognize what the neuroscience of addiction, and particularly the influence of drugs of abuse on dopaminergic systems, means. Given what dopamine signals represent, what information they carry, we can deduce that drugs of abuse cause us to recognize greater reasons to use drugs than we recognize for promoting the things that we hold most dear most of the time. But, at the same time, to be subject to such distortions in one's values is a deep and terrible burden to bear, one that no one should have to bear in order to comply with the law. To be in such a condition is, indeed, to be worthy of pity and to be worthy also of some special consideration from the courts. We should not be ambivalent in our attitude towards addicts, vacillating between conflicting points of view. Instead we should recognize that our conflicting attitudes have an equal and legitimate basis in addiction's nature.[10]

---

[1] Holton's usage of the term "akratic" is different from that here. So, he may not endorse this way of summarizing his position. Still, in the sense in which the term is used here, Holton takes addicts to act akratically.

[2] See Kosten, T., Rosen, M., Bond, J., Settles, M., St. Clair Roberts, J., Shields, J., Jack, L., and Fox, B. "Human Therapeutic Cocaine Vaccine: Safety and Immunogenicity" in Vaccine, vol. 20, pp. 1196-1204, 2002.

[3] In fact, such drugs probably both prevent the cocaine user from liking cocaine and decrease the likelihood he will choose it; what is unclear is whether they decrease the likelihood of choice because they decrease the degree to which cocaine is liked, or for some other reason. Such drugs work by producing antibodies that bind cocaine in the bloodstream before it enters the brain and so prevent cocaine from affecting either the liking system or the choosing system. It is therefore no surprise that we see decreases in

consumption behavior in rats who have taken the vaccine. See Fox, B. et al, "Efficacy of a Therapeutic Cocaine Vaccine in Rodent Models" in Nature Medicine, vol. 2, pp. 1129-1132, 1996.

[4] One implication of this is that there is something misleading in the idea of an "unwilling addict". (The term was coined in Harry Frankfurt, "Freedom of Will and the Concept of a Person" in The Importance of What We Care About, Cambridge: Cambridge University Press, pp. 11-25, 1988.) The unwilling addict is thought to take the drug to which he is addicted despite the fact that even at the time of action he does not value taking the drug. While such a creature is possible, if the argument of this paper works we have reason to believe that human addicts, with brains that function the way ours do, are not unwilling addicts in this sense. Consistent with having a brain like ours, however, it is possible, even common, not to value consumption both moments before and moments after the time of the decision to consume.

[5] See, for instance, Schultz, W. Apicella, P. & Ljungberg, T. "Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task" in Journal of Neuroscience, vol. 13, pp. 900–913, 1993; Hollerman J. R. & Schultz, W. "Dopamine neurons report an error in the temporal prediction of reward during learning" in Nature Neuroscience, vol. 1, **pp.** 304-309, 1998; Schultz, W. "Predictive reward signal of dopamine neurons" in Journal of Neurophysiology, vol. 80, pp. 1–27, 1998.

[6] Desire satisfaction is intended to be understood here as follows: When a subject is moved to acquire something, and acquires it, he enjoys a reduction in motivation. If he gets exactly what he was moved to acquire, then his motivation is reduced to zero. If he has residual motivation left-over—he is not satisfied with what he acquired—then it is

reduced to less than zero. Desire satisfaction is the amount of reduction in motivation that is enjoyed on acquiring the object. If the subject was highly motivated, there is greater potential for desire satisfaction, although there is also greater potential for disappointment.

[7] See, for instance, Waelti, P., Dickinson, A. & Schultz, W. "Dopamine responses comply with basic assumptions of formal learning theory" in Nature, vol. 412, pp. 43–48, 2001; Bayer, H. M. & Glimcher, P. W. "Subjective estimates of objective rewards: using economic discounting to link behavior and brain" in Society of Neuroscience Abstracts, vol. 28, p. 358.6, 2002; Berridge, K. C. & Robinson, T. E. "What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience?" in Brain Research Review, vol. 28, pp. 309–369, 1998.

[8] Timothy Schroeder, Three Faces of Desire, Oxford: Oxford University Press, 2004; Timothy Schroeder, "Irrational Action and Addiction" in What is Addiction?, edited by Don Ross, Harold Kincaid, David Spurrett and Peter Collins, Cambridge: MIT Press, 2010, pp. 391-407.

[9] Under the Model Penal Code, for instance, a homicide is a manslaughter, rather than a murder if it was performed "under extreme mental or emotional disturbance for which there is reasonable explanation or excuse" (Model Penal Code §210.3(b).

[10] Thanks to Pamela Hieronymi, Neil Levy, and Walter Sinnott-Armstrong for comments on earlier drafts.