Normal and Addictive Desires

David Papineau and Patrick Butlin

1. Introduction

Our aim in this chapter is to understand drug addiction within the context of an overall theory of human behavior.

In our view, behavior is to be understood as influenced by a number of different control systems. We shall focus in particular on what we will call the *habit* system, the *desire* system, and the *planning* system. These three systems are of crucial importance for determining almost all of our choices, addictive and otherwise. It seems likely that the first two of these systems are shared with other animals, but that the planning system is peculiar to humans. By describing these three systems, we aim to provide a framework that will clarify the mechanisms of addiction and the loss of control they involve.

In this context, one possible theory of the mechanism of drug addiction stands out as offering a particularly simple and elegant explanation of how addicts feel and behave. According to this theory, which we call the *standing desire theory*, addictive drugs cause long-term changes to the desire system, which in turn cause addicts to experience abnormally strong *occurrent desires* to take drugs.[1] These occurrent desires can be strong enough to prevent addicts from acting on the intentions formed by their planning systems. We will compare this theory to some other prominent proposals, and assess its current prospects.

From any perspective, a distinctive feature of addiction is that addicts are characteristically unsuccessful at sticking to long-term intentions to abstain. Such undermining of intentions is not, of course, peculiar to addiction. It occurs whenever somebody forms a resolution and then later gives in to temptation. To this extent we view addictive behavior as similar to normal behavior. Just like normal agents, addicts are perfectly capable of forming rational intentions, and in particular intentions about indulging or refraining from their addictions. And again, just like normal agents, they sometimes fail to stick to these intentions under the influence of contrary later impulses.

What distinguishes addicts, according to the standing desire theory, is that their occurrent desires to indulge their addiction are often *abnormally* strong. Addicts are subject to unusual influences that accentuate their cravings and hamper their plans to abstain.

The standing desire theory, like other modern accounts of drug addiction, attributes these unusual influences to the role of *dopamine*. The ingestion of addictive substances results in unusually high levels of dopamine in the brain, and this leads to

[1] Standing desires are (roughly) persisting mental states in virtue of which we prefer or value certain outcomes, and occurrent desires are manifestations of these states at particular times that play a direct role in motivating us to pursue the outcomes concerned. We discuss this distinction in detail in section 2.4.

a standing disposition to experience strong desires for those substances in future, in ways discussed in detail below.

As far as a definition of ‘addiction’ goes, we shall restrict this term in what follows to the state of being disposed to experience abnormally strong cravings, *due to abnormal effects involving dopamine*. A less restrictive definition would allow the term also to encompass abnormal cravings that are due to unusual effects not involving dopamine. It is a moot point whether there are any other such non-dopaminergic exceptional causes for abnormal cravings. In any case, we shall not consider any such phenomena in this chapter. Given this, it will be more convenient to stick to the more restrictive definition and understand ‘addiction’ as implying ‘dopamine-caused’ in what follows.

If some condition involves abnormal desires, whether caused by dopamine or something else, does that make it a *disease*? We shall not spend any time on this question. Indeed it is not obvious that it is a good question. There is no agreed understanding of the term ‘disease’, and in our view it is doubtful that it tracks any natural kind, as opposed to some conventional category tied to the historical idiosyncracies of the medical profession. Nor is there any agreement on the moral or practical consequences of something being a disease. It is unclear, to say the least, whether or not having a disease means sufferers are not responsible for their actions, are not to blame for their condition, should be treated rather than punished, and so on.

Of course, issues of responsibility, blame and punishment are real and pressing, and particularly so with respect to addicts. But they are best addressed directly, without a detour into the issue of disease. We can simply ask straight off about the responsibility, blame and punishment of addicts, without also worrying about whether addicts are *ill* or not.

While these practical and moral questions are not themselves within the scope of this chapter, we take our arguments to be directly relevant to them. We may not need to know whether addiction is a *disease* to know how to deal with it. But we will be hard pressed to work out the right way to respond to addicts if we don’t have a proper understanding of the mechanisms behind their plight, and specifically of how they relate to the normal mechanisms of action-control. To this we now turn.

2. Systems for Behavioral Control

2.1 Three Systems

The three systems for behavioral[2] control which we will describe are the *habit system*, the *desire system*, and the *planning system*. The habit system is likely to exist in many animals; the desire system has been extensively studied in rats and mice, and can therefore be expected to exist in most mammals; and the planning system may well be present only in humans, at least in its full form. Although these three systems are of

[2] Philosophers are much concerned with definitions of such terms as ‘action’ and ‘intentional action’. We shall by-pass these issues, and simply use the neutral term ‘behaviour’ as a general term for all forms of purposive motor activity (that is, motor activity that can be explained teleologically in terms of its furthering some end). So our talk of ‘behavior’ should not be thought to exclude items normally called ‘actions’.

particular interest, they are not responsible for all human behavior; one type of exception is behavior caused by reflexes, and there may be others.

We will argue that humans use all three of the systems we shall describe. How exactly these systems interact is a complex question which will be of relevance below. At first pass, we can suppose that, in cases where the systems conflict, the desire system will dominate the habit system, and that the planning system will dominate both the others. But as we shall see, this domination is not absolute in either case.

2.2 The Habit System

In the habit system, behavior is controlled by learnt associations between *stimuli* and *responses*. So the resulting patterns of behavior, or habits, are sometimes called *S-R associations*, and are the products of *S-R learning*. Stimuli are features of the person's or animal's circumstances, and responses are behaviors that they perform in those circumstances. S-R learning takes place when stimuli are followed by responses, which are in turn followed by either *rewards* or *punishments* that between them amount to a level of reward different from what is expected in that situation. Habits get stronger when responses are followed by more reward than expected, and weaker when they are followed by less reward than expected. What makes a situation rewarding varies between species and between individuals, but typical rewards include getting food or sex.

Experiments on rodents suggest that internal states, including states of physiological need like hunger or thirst, influence the habit system by acting as stimuli. Niv and colleagues (2006) found that rats that had been trained when hungry to press a lever for sugar solution reduced responding when they were thirsty but not hungry. Since sugar solution is good for both hunger and thirst, these results suggest that habits are triggered specifically by those physiological needs that were active when they were acquired.

One crucial question about the habit system is how it interacts with the other systems of behavior control. On one picture of the habit system, it only affects human behavior when we let it – when we are performing familiar routines, and don't concentrate on what we are doing. If this view is correct, then it seems that addiction could not be explained by drug-taking habits, because these habits would be easy to resist. Also, addicts perform complex sequences of apparently pre-planned behaviors in order to source drugs, which again would seem hard to explain if addictive behavior were solely habitual.

2.3 The Desire System

The desire system is defined by responsiveness to information about the values of outcomes. The existence of the desire system in rodents was established in the 1980s, primarily by *outcome devaluation* experiments, which produce results that are hard to explain if we assume that rats are only capable of habitual control (Adams and Dickinson 1981). In these experiments, rats are first given the opportunity to press a lever, and receive a food reward when they do so–peanuts, for instance. They remain in this environment for long enough to learn to press the lever. Then the rats are taken away from the cage with the lever, and some of them undergo 'outcome devaluation':

they are fed with peanuts, and then given a lithium chloride injection that makes them sick. The other rats in the control group receive both the food and the injection, but at different times. When the rats are returned to the cage with the lever, those that went through the outcome devaluation press the lever less than those in the control group, even though no rewards are delivered to either group. It seems that the habit system cannot explain this result, because the only difference in the experiences of the rats in the two groups concerns the outcome – it does not involve them being presented with the stimuli, or performing the responses, that are relevant to the behavior that changes. Apparently, the rats were storing information *both* about the value of the outcome–getting peanuts–*and* about the relationship between the action and the outcome–between pressing the lever and getting peanuts.

Another finding of outcome devaluation experiments is that they don't work if the rats are given too much initial training on the lever; in this case, they do not reduce responding as a result of outcome devaluation, instead behaving as though their behavior was controlled by a habit (Adams 1981). So rats are thought to use both the habit system and the desire system in parallel, with habits being formed and modified relatively slowly. Further evidence for this claim comes from studies by Yin and colleagues (2004, 2005, 2006), who selectively destroyed parts of the striatum in rats, and thus identified distinct brain areas that are apparently responsible for habitual and desire-system control. The idea that rats use both systems, and that outcome devaluation is a good way to distinguish them, is now central to an extensive research programme on behavior control and reward learning (Balleine and O'Doherty 2010).

Unlike the habit system, the desire system uses states that track the values of outcomes to control behavior. It combines these with its learned information about the relationships between behaviors and rewards to calculate which available behavior is expected to bring about maximum reward. In effect, the choice of behavior depends both on how strongly desired outcomes are, and on how likely experience has shown those outcomes to be consequent on given behaviors. To this extent, the desire system works something like the economists' model of a utility-maximizer, selecting the behavior with maximum expected utility.

Both the habit system and the desire system seem to exist in humans as well as in rats. One striking piece of evidence for this is that outcome devaluation experiments work in a similar way with human participants, including the fact that overtraining leads to habitual behavior and the loss of sensitivity to devaluation (Tricomi et al. 2009; for a very similar experiment on rats, see Balleine and Dickinson 1998). In a neat demonstration of this latter point, researchers found that people who regularly ate popcorn at the cinema would eat roughly equal amounts whether the popcorn was fresh or stale, but that those who ate it less frequently were more discriminating (Neal et al. 2011). Also, imaging studies have found that the brain areas in humans involved in the two systems correspond to those that the lesion studies of Yin and colleagues found in rats (Balleine and O'Doherty 2010).

It is worth noting that desires seem to be involved, not only in determining how we behave, but also in generating the reward signals responsible for the formation of new habits and desires.[3] Evidence for this claim comes from the phenomenon of secondary

[3] This idea is central to the theory of desire proposed by Schroeder (2004).

reinforcement, in which behaviors are learnt when they lead to outcomes that animals have been trained to find rewarding, such as lights and tones, rather than outcomes they might be expected to find rewarding regardless of past experience, such as food (Skinner 1938; Hull 1943). This is an important point, because it helps to explain how humans come to have the wide range of–sometimes unlikely–desires that they do; the idea would be that these desires can all in principle be connected to 'primary rewards' like food or apparent social advancement via long chains of associations.

2.4 Standing and Occurrent Desires

It is important to distinguish between the roles of *standing* and *occurent* desires in controlling behavior. We can think of standing desires as relatively stable features of our characters which persist over time; occurrent desires, by contrast, are the active manifestations by means of which these stable features make direct contributions to the control of behavior.

Some obvious features of our behavior provide ample initial reason for making this distinction. On one hand, if we did not have standing desires, but instead occurrent desire arose spontaneously and dissolved when satisfied, then it would be hard to explain why we consistently make similar choices in similar situations–for instance, why some people almost always choose chocolate ice-cream, even when lots of other flavors are available. On the other hand, if we did not have occurrent desires, but instead all of our desires were equally poised to influence behavior at all times, it would be hard to explain why we are sometimes highly motivated to eat ice-cream, but at other times indifferent to doing so, even when we know it is readily available and we have no other very pressing tasks.

A deeper reason for recognizing the distinction is that, where standing desires enable us to store permanent knowledge about the values of outcomes, occurrent desires enable us to register the ways in which the values of outcomes sometimes change dramatically with circumstances and in particular with our physiological needs. This need for two levels of responsiveness to the values of outcomes is particularly clear when we consider that outcomes that are sorely needed in some circumstances may be dangerous in others. For example, when an animal is very salt-deprived, it may be crucial to survival that it experiences strong desires for salty foods; but if this boost to the desires persisted in the long term, the risk to the animal's health could be considerable.

The distinction between standing and occurrent desires is supported by our understanding of how the brain works. There is some evidence for identifying standing desires with relatively stable, gradually evolving neural structures in the orbitofrontal cortex (Balleine and O'Doherty 2010; Plassman et al. 2007). Occurrent desires, on the other hand, are naturally equated with current patterns of neural activity, in line with imaging results. Given this picture, we can view behavior-selection by the desire system in terms of competing coalitions of activity which represent associations between currently valued outcomes and salient available actions (cf. Cisek 2007).

It is no doubt the proximal role of occurrent desires in behavior control that renders the task faced by the desire system computationally tractable. If all standing desires needed to be taken into account when selecting behaviors, the desire system would continually need to carry out unmanageably complex calculations. However, if it works solely with occurrent desires which are activated only when perceptual experiences or ongoing thoughts bring their objects to mind, then it will only need to deal with a limited number of outcomes. This picture fits with the introspectively plausible idea that the strength of an occurrent desire is affected by the degree of attention that its object attracts (Hare et al. 2011). We want things less when they are more distant from our thoughts, and more when they are closer. Given that that our attention is also guided by our desires, it seems that a positive feedback loop may operate here, which may explain why addicts sometimes find themselves constantly thinking about the things they are addicted to.

2.5 The Planning System

The two systems we have discussed so far are both for deciding what to do *now*. However, we humans also have the ability to think about what to do in the future, to form plans, intentions and resolutions, and indeed the ability to reflect in a more general way on our own motives, purposes and concerns. These abilities, discussed by Michael Bratman (1987) and Richard Holton (2009), make up the *planning system*.

The planning system does not always fix how we will behave in the future–we do not always stick to our plans–but it certainly has a significant influence. Conflicts frequently arise between what we have planned in advance, and what the desire system or the habit system recommend when the time comes. Sometimes the older habit and desire systems dominate in such cases of conflict, but this is by no means the general rule.

The other two systems we have discussed so far both promote rewards, and so are relatively easy to understand as adaptations. Both habits and desires are learnt through the association of their objects–behaviors or outcomes–with reward, and reward signals are prompted in the first instance by outcomes that are relatively easy to detect and conducive to survival and reproduction, such as food, water, sex and (in many animals, including humans) apparent social success. The habit system allows animals to learn new rewarding behaviors, and the desire system adds to the sophistication with which reward is pursued. While there is room for debate about the nature of some of the advantages brought by the desire system over the habit system, one clear advantage is that it allows relevant information learnt in a wider variety of ways to be brought to bear in deciding how to behave–as in the outcome devaluation experiments. However, in the case of the planning system it may be less obvious why the system is adaptive.

There is much to be said on this topic. Here we shall focus on three putative advantages which come with the ability to form long-term plans. First, the planning system helps us to co-ordinate our actions more effectively, not only with other people, but also with our successive selves over time, so to speak. To appreciate the latter point, consider what may be viewed as an *intrapersonal prisoner's dilemma*: the question of whether or not to take exercise today (Ainslie 2001; Gold 2014). For many people, this question is like the prisoner's dilemma. From their present perspective, things will go best for them if they do not exercise today, but do on many other days; things will go fairly well if they exercise today and on many other days;

they will go fairly badly if they neither exercise today nor on other days; and they will go very badly if they exercise today but do little otherwise. So it looks as though on any given day, it will be rational to choose not to exercise, since this will mean things go better whether or not they exercise on other days. However, the planning system allows us to adopt a longer-term perspective, and recognise that in the face of this situation the only way to get the best overall outcome is to come up with a training plan, and stick to it.

A second advantage is that the planning system allows us to draw on a far wider range of information about means to valued ends than either the habit or desire systems. We have seen that the desire system outstrips the habit system in forming desires on the basis of new information about which outcomes are valuable. But this is consistent with the desire system being highly limited in its access to information about which behaviors will be effective means to those valued ends. For all that has been said so far, the desire system may be limited in this respect to means-end information that comes directly from personal experience of given behaviors leading to valued outcomes. If this is right, then a system that allows careful and prolonged deliberation before fixing on a plan will have the benefit of being able to draw on far wider informational resources, including crucially the testimony of others and bodies of cultural tradition.

Finally, there is the point that the planning system allows us to reflect on our desires, and privilege the pursuit of ones that we judge more worthy. This is in contrast to the desire system, in which the desires that influence us most strongly are just those that happen to be most strongly aroused by currently encountered cues. We can mark this contrast by saying that the planning system is orientated to what we *judge* to be genuinely *valuable*, where the desire system simply pursues those outcomes that are currently desired. Reflection may lead us to judge that some things we desire are of little value–cigarettes, say–or to judge that things that we do not desire much are of significant value–like spending time exercising outdoors. This does not necessarily commit us to the view that our judgements of value are radically independent of our desires. Some philosophical theories of value do have this consequence, and they may well be right. But for present purposes it will make no difference if judgements of value are simply the upshot of reflection about the ways in which different desires facilitate, enhance, obstruct or diminish one another. For example, you might currently desire both good health and cigarettes. But reflection will tell you that your good health is something that it is also conducive to satisfying many of your other desires, whereas cigarette smoking is not, and will thus indicate that the former desire is more worthy that the latter. Note that, even on this deflationary understanding of the difference between value and desire, a planning system that focuses on values will have a clear adaptive advantage over one that is restricted to pursuing currently desired outcomes.

2.6 Systems in Conflict

So far we have said little about the way the three systems of behavioral control interact, and in particular about the way the human planning system relates to the habit and desire systems we share with other animals. There are two possible models here.

On a first *parallel systems* model, the three systems each operate independently, fixing on their preferred behavioral response at any one time, and then fighting it out, so to speak, as to which one prevails. On this model, the planning system will presumably manage to dominate the other two in normal cases, but there will be special circumstances where it is outfought.

On a second *nested systems* model, the planning system will rather seek to control behavior, when it does so seek, not directly on its own behalf, but by controlling the processes of the older behavioral control systems, perhaps by somehow increasing the strength of desires for certain outcomes, and reducing desires for other outcomes. On this model, the planning system can again be expected to determine behavior in most cases, but will fail to do so if it loses control of the desire system. This might happen, for example, if more immediate influences on the formation of desires are strong enough to eclipse the modulation of desires stemming from the planning system.

Holton (2009, ch. 6) argues for a view of the relationship between the planning and desire systems that amounts to the parallel systems model, on the grounds that resisting temptation can be effortful, and that this implies a struggle between the two systems. But we see no reason to commit to this model. Even on the nested systems model, resisting temptation will involve a process in which different influences–the planning system and current cues–compete to generate stronger occurrent desires. These influences may be mediated by effects on attention, since this is a major factor in determining the strength of occurrent desires, and the conscious control of attention is typically effortful.

On either model, then, the planning system will on occasion fail to determine behavior, either because its directive is countermanded by an independently operating desire system, or because it loses control of the desire system that it is normally able to manipulate. So, on either model, there will be a threat to the planning system at any time when the independent influences that direct the desire system push strongly for a course of action other than chosen by the planning system.

Our discussion so far has indicated a number of ways in which divergence between the planning system and the desire system can arise. For a start, we have seen how the planning system will use a wide range of testimonial and cultural information to choose means to its ends, where the desire system is arguably limited to information derived from the organism's experience. Moreover, the planning system normally chooses strategies for possible *future* circumstances, and so will often fail to take into account changes in circumstances that occur in the interim and influence the later desire system. And finally the planning system will seek to pursue considered *values*, where the desire system is directed by unmodulated occurrent *desires*, which are largely determined by what is salient to us at any given time, with the result that at least in modern environments we often experience strong temptations that fluctuate rapidly. Given that the habit system is also capable of affecting how we feel inclined to behave, and will also sometimes conflict with the other two systems, it is not surprising that we often fail to stick to our resolutions, or behave inconsistently in other ways.

3. A Theory of Drug Addiction

3.1 Dopamine and Addiction

We now turn to the standing desire theory of drug addiction, which has been most clearly advocated by Richard Holton and Kent Berridge (2014). The theory claims that many addictive drugs boost dopamine levels in the brain abnormally, which causes the formation of excessively strong desires for those drugs. This theory is concerned with what it is about the drugs that makes them addictive, and why addiction to these drugs generates the patterns of motivation and behavior that it does, rather than with other issues, such as why some drug users become addicted and others do not. It does not incorporate any claims about the extent to which addicts are in control of their actions, unlike proposals by Frankfurt (1971) and Foddy and Savulescu (2010), but does provide empirical foundations from which such claims could be developed. It also illustrates the value of the framework for understanding behavior that we have just described. As we will explain, this theory is elegant and promising, although it is in tension with some prominent accounts of the function of dopamine. We start by describing the role of dopamine in the theory.

Dopamine is a neurotransmitter that is thought to have a small number of specific cognitive functions, although there is considerable debate about exactly what those functions are (we will discuss aspects of the debate in section 3.4). However, one point which is well-established is that many addictive, psychoactive drugs boost levels of dopamine in the brain, by a variety of mechanisms. These include amphetamines, alcohol, nicotine, opiates, cocaine, cannabis and benzodiazepines (Nestler 2005; Tan et al. 2010). Given that in other respects these drugs have variable effects, a natural hypothesis is that it is this effect on dopamine which explains why they are addictive, and this claim now forms part of many different theories of the mechanism of drug addiction.

Even if this is correct, it is likely that the details of the ways in which different drugs act are important for understanding the various specific symptoms that addiction can incorporate, but we will concentrate on the wider picture. Also, it is entirely possible that some psychoactive drugs are addictive and abused despite not having this effect, or primarily in virtue of other effects, but for the sake of convenience we will ignore this possibility when using terms like ‘drugs’ and ‘drug addiction’.

A further reason to view dopamine as central to drug addiction is that, despite the uncertainty about the exact functions of dopamine, it is known to be centrally involved in processes relating to reward and behavior selection. In particular, there is considerable (although perhaps not conclusive) evidence that *phasic* dopamine is some form of reward signal. Dopamine is released continuously by the *midbrain dopamine neurons*, at what is called a ‘tonic’ level, but its release is also frequently punctuated by brief bursts and pauses which are believed to be signals of some kind, and are known as ‘phasic’ dopamine signals. Some of the most influential studies suggesting that phasic dopamine signals represent reward levels were conducted by Wolfram Schultz (e.g. 1998), who found that bursts of dopamine release occurred in monkeys’ brains when they were given unexpected rewards. Further to this, Schultz also found that when the monkeys were trained to expect the rewards subsequent to a cue, the bursts stopped occurring when the reward was delivered, and instead came to coincide with the cue. When the cue was delivered but the reward omitted, a phasic pause in dopamine release followed the omission, interpreted as a negative signal.

As a result of these studies, phasic dopamine has come to be thought of as a *reward prediction error* signal, representing the difference between the level of reward received and the level expected at that time. So on this hypothesis, drugs of abuse which cause artificially high dopamine levels would–provided they boosted phasic rather than tonic dopamine–lead the brain to treat them as always much better than expected.

3.2 The Standing Desire Theory

The theory we are interested in further proposes, in keeping with this idea, that the phasic dopamine signal is used for updating the strength of standing desires (see esp. Holton and Berridge 2014, sect. IV). This point, rather than the claim that addiction is caused by abnormal dopamine signals, is what distinguishes the standing desire theory. On this view, positive dopamine signals at a given time strengthen standing desires for those outcomes that were salient at that time, while negative dopamine signals cause such desires to be weakened. So the standing desire theory proposes that addiction is the result of dysfunctional, exceptionally strong standing desires to take drugs, which are produced as a result of artificially-boosted dopamine signals. We will first discuss some features of this theory, then describe how it relates to existing accounts of drug addiction.

A first point is that the standing desire theory denies that drug addiction can be explained purely in terms of strong habits. It attributes addiction to acquired desires, and accordingly sees drug-seeking behavior as an adoption of behavioral means to desired ends, as directed by the desire system and perhaps also the planning system. In support of this perspective, Berridge and his co-authors have emphasised how addicts often perform complex sequences of carefully-planned behaviors in order to get the opportunity to take drugs (Berridge and Robinson 2011; Holton and Berridge 2014). They argue that this counts against the thought that drug addictions are literally habits, because the habit system involves no capacity for means-ends reasoning.

Further, the standing desire theory yields an immediate explanation of why addicts are strongly attracted to drugs: strong standing desires generate strong occurrent desires, when appropriately cued. On this point, it is an advantage of the standing desire theory that it makes motivation to take drugs depend on addicts' circumstances, to about the right degree. For addicts who are trying to quit or cut down, avoiding drug cues is a valuable–perhaps crucial–strategy, and the theory can explain this, because standing desires motivate us only when they become occurrent, and occurrent desires are strongest when they are most strongly cued. This also explains the possibility of relapse even in addicts who describe themselves as having lost their desire for drugs, since standing desires alone may be difficult to recognise in oneself. Other addicts, however, may spend considerable amounts of time seeking and taking drugs, and the theory can also explain this, because it predicts that people who spend a lot of time engaging with drugs and in circumstances they associate with drugs will be almost continuously motivated to remain in those circumstances, and take more drugs. Notably, the standing desire theory predicts not only that addicts will form strong standing desires for drugs, but also for other outcomes that are salient when they take drugs, because these desires will also be artificially boosted by the drugs' effect on dopamine. Also, because standing desires persist over time, the theory can explain why even addicts who have stopped taking drugs remain at risk of relapse much later in life.

Finally, the standing desire theory can also explain why addicts are so prone to succumb to temptation even when they have resolved to abstain, Along with everybody else, addicts will have many other standing desires, and many of these will conflict with their desires to take drugs, especially if their addictions become harmful. The planning system means that addicts are likely to recognise this and to form plans, intentions or resolutions which involve reducing or ending their drug use. However, the role of dopamine in the formation of the desires to consume drugs means that their occurrent desires for drugs will on occasion be exceptionally strong. This can influence the desire system to operate in opposition to the planning system, along the lines discussed in section 2.6 above, with the result that the addicts find themselves unable to stick to their resolution to quit.

Overall, then, the standing desire theory provides a particularly straightforward account of addiction, which explains all of its most striking features. By focusing on the desire system, the theory avoids the difficulties that face habit-based accounts of addiction, and also leaves room for the planning system–and the desire system, in circumstances that do not evoke drug-related occurrent desires–to motivate addicts to quit. So crucially, the standing desire theory in the context of the three-system view accounts for the conflict in addiction. The theory does not imply that this conflict is a necessary feature of addiction, but only that it will arise if addicts come to believe that their addictions are damaging.

The standing desire theory also has the advantage of making very limited claims about the neurobiology of addiction. It claims only that all addictive drugs boost dopamine levels, and the dopamine has the function of updating standing desires. Addictive drugs certainly have many effects on the brain, which may contribute in a variety of ways to the behavior and experiences of those who consume them, but it is an advantage for a theory to make minimal claims about these matters, since we are looking for the common mechanism behind drug addiction in general.

3.3 Comparison with Existing Theories

In this section, we shall briefly discuss how the standing desire theory relates to some other existing accounts of drug addiction.

One well-known account which also has strong affinities to the standing desire theory is by Hyman (2005). Hyman argues that drug addiction is caused by distorted dopamine signals, that phasic dopamine is a reward prediction error, and that it explains drug addiction because it causes long-term changes to the motivational effects of drug cues and to dispositions to perform drug-seeking behaviors. However, Hyman’s account is less specific about the effects of dopamine than is the standing desire theory. On his view addiction seems to involve the distortion of habits, desires, and possibly some other states involved in behavioral control as well.

Arpaly and Schroeder (2014) also give an account of addiction which focuses on the role of dopamine. However, they do not see this as generating artificially strong standing desires, but rather as affecting habits–including habits of thought–by accentuating the strength of positive and negative reward signals. They also claim that these signals would produce very pleasant and very unpleasant conscious experiences, respectively, which would also influence behavior. From the perspective we have adopted, their theory seems unnecessarily complex; it is unclear why they resist the straightforward suggestion that dopamine updates standing desires.

Other recent accounts have tended to emphasise the degree to which addictive choice is like ordinary choice. Foddy and Savulescu (2010) argue that this follows from the claim that addicts are motivated by unusually strong desires to take drugs, and write that 'addictive desires are just strong, regular appetitive desires' (2010, p. 14). Pickard (2012) acknowledges that addictive desires are strong and 'habitual', but argues that addiction is explained in large part by the genuine benefits drugs bring to some addicts with comorbid psychiatric disorders, and also by the poor quality of life that addicts reasonably expect to experience on quitting. Both of these accounts therefore emphasize the way that taking drugs is rational for addicts, given their strong desires. Up to a point, we agree with this perspective. From our point of view, addictive behavior should indeed be understood in terms of the systems of behavior control that addicts share with everybody else, and in particular in terms of the operation of the addicts' desires within these systems. But on its own this observation leaves out the crucial point that addicts' desires are abnormally and artificially strong, due to the special effects addictive drugs have on dopamine, and that this explains why addicts find it so hard to stick to their resolutions.

In addition to these, a further prominent account has been put forward by Berridge and his collaborators, which is based on his view of dopamine function. This view does not sit entirely comfortably with the standing desire theory, so we attempt to disentangle these issues in the following section. In our view, it is best to think of Berridge's *incentive sensitization theory* as related to, but distinct from, Holton and Berridge's standing desire theory.

3.4 The Incentive Salience Theory

While we find the standing desire theory plausible and explanatorily attractive, we do not take it to be mandated by the evidence. As we have explained, what makes the standing desire theory distinctive is not the claim that drug addiction is caused by drugs' effects on dopamine, but the further claim that the function of dopamine signals is to update standing desires. However, this is not the only hypothesis about dopamine function that is defended in the current literature. Two prominent alternatives are the *incentive salience* view, and *reward learning* accounts.

The incentive salience theory is advocated by Berridge (2007, 2012). According to this theory, the function of phasic dopamine signals is to motivate animals to act at times when they encounter either rewards or cues that predict reward, and to pursue the perceived or cued reward. As Berridge puts it, dopamine is a 'wanting' signal–it makes animals *occurrently* 'want' perceived or cued rewards. Dopamine does not, according to this view, have the function of producing lasting effects, either by contributing to learning new actions, or by updating stored information about reward. Berridge argues that dopamine is neither necessary nor sufficient for habit learning, since mice with almost no dopamine can learn to perform new actions (Hnasko et al

2005), and giving mice extra dopamine does not seem to enhance learning (Cagniard et al 2005).

If the incentive salience theory described dopamine's only effect, then drug addiction would be hard to explain. The incentive salience theory predicts that, when we take drugs, we will become unusually strongly motivated to take more of the same drugs, because of the artificially high levels of dopamine release that this causes. So the incentive salience theory is particularly well-suited to explaining drug binges. However, the incentive salience theory does not itself predict that drugs will cause any longer-term changes, or that addicts will feel strongly motivated to take drugs at any time other than once they have already begun a session of use. So the incentive salience theory of dopamine function certainly needs to be supplemented with further claims about dopamine's effects, in order to contribute to a theory of drug addiction. (Without such claims, one might even argue that the phenomenon of drug addiction undermines the incentive salience theory of dopamine function. After all, the phenomenon of being abnormally attracted to drug use even before starting a session seems to be common to a range of dopamine-involving drugs. So we would expect the role of dopamine to explain this phenomenon, and not just the continuation of binges once they have started.)

Berridge and his colleagues are not unaware of this difficulty, and deal with it by adding a further element to their theory. They argue that repeated exposure to drugs causes *sensitization* of the incentive salience system. As Berridge and Robinson (2011) present things, addiction occurs because the incentive salience system becomes hyperreactive to drug cues, after those cues have been repeatedly followed by unusually strong dopamine signals. Since this is a long-lasting effect, it can potentially explain addiction. The idea would be that exposure to cues will activate the addicts' incentive salience system, even before they have ingested the drug. This would then explain the addicts' abnormal tendency to succumb to temptation in the first place, and not just their disposition to carry on bingeing once they have succumbed.

It is worth emphasizing the differences between the standing desire theory and the incentive salience(-plus-sensitization) theory. For a start, the standing desire theory thinks of what it is to desire something as, roughly, to value it for the purposes of goal-directed control, as identified by outcome devaluation experiments. The incentive salience(-plus-sensitization) theory, in contrast, thinks of 'wanting' something as, roughly, the state of being attracted to it when one encounters it. 'Wanting' in this form is possible in creatures that lack the desire system, because they can respond to 'wanted' objects by unlearnt behaviors such as approaching and consuming them. Perhaps the same kind of state does track the values of outcomes and objects for these two purposes in creatures that do have desire systems, but this is not to be taken for granted. A further point is that Berridge and Robinson (2011) distinguish incentive salience 'wants' from what they call 'cognitive incentives' or 'cognitive desires', which they say are involved in 'goal-directed planning' and responsible for subjective ratings of desire. It is not clear that they are here thinking of behavioral control in quite the same tripartite way as we do, but in any case their 'wantings' seem distinct from our desires.

3.5 Reward Learning Accounts

Reward learning accounts of dopamine function are probably the most popular current class of theories (e.g. Wise 2004; Balleine et al. 2008; Glimcher 2011). These theories support the standing desire theory of addiction to the extent that they claim that phasic dopamine is a reward prediction error signal, with the function of producing long-term changes to motivational systems. But like the incentive salience theory, they do not back up the crucial claim for the standing desire theory, which is that phasic dopamine has the function of updating standing desires. Instead, they argue that dopamine signals update *habits* (and perhaps also *Pavlovian values*, which are associations between stimuli and innate behaviors).

This does not mean that they agree with Berridge's incentive salience account; Berridge claims that dopamine bursts signal that a stimulus with high incentive salience is present, and generate enhanced effects only in the moment, whereas reward learning accounts claim that dopamine bursts teach the incentive salience system to come to treat certain stimuli as action-guiding even when they are encountered in the future. So reward learning accounts of dopamine function, unlike the incentive salience account, predict that drugs would cause the formation of strong habits controlling drug consumption. These accounts therefore fit well with Hyman's and Arpaly and Schroeder's accounts of drug addiction.

In principle, it could be that dopamine signals update habits and Pavlovian values as well as desires–or, for that matter, that they both update desires and have the incentive salience function ascribed by Berridge. If this were so, it would be a partial vindication of the standing desire theory. But at the same time it would mean that this theory would not be the whole story about dopamine and addiction.

4. Conclusion

Our strategy in this chapter has been to understand addiction within the framework of three interlocking systems of behavioral control—the habit, desire and planning systems. We take the analysis we have offered to illustrate the virtues of this approach.

In the first instance, our framework allows a clear specification of what is distinctive about addicts. They experience abnormally strong occurrent desires to consume the relevant drugs. As a result, they will characteristically—though not necessarily—fail to stick to any long-term abstention plans they may form. As we have explained, while the planning system will normally function to override any contrary behavior prompted by conflicting occurrent desires, this is by no means inevitable, and is crucially sensitive to the strength of the conflicting desires.

Moreover, our framework allows a plausible and attractive explanation for why addicts experience abnormally strong cravings. We attribute this to the subversion of the desire system by the dopamine that is produced by the ingestion of drugs. In our view, this leads to permanent changes in addicts, in the form of strong standing desires for drugs.

Other theories also attribute addiction to abnormal effects resulting from the production of dopamine, but not via the formation of standing desires. We have explained above how these theories face certain empirical difficulties. In addition, it is not always clear exactly how these theories view the overall psychological effects produced by dopamine, for lack of any explicit discussion of mechanisms of behavior control, along the lines of our three-system model.

Given this, we recommend the standing desire theory as the most elegant explanation currently available of the phenomenon of drug addiction. It explains why addicts experience lasting attraction to drugs, which nonetheless is sensitive to a significant degree to their circumstances; it leaves space for addicts to have a wide range of attitudes to their own addictions, depending on the judgements of value made by the planning system; and it can explain readily why it is hard (but possible) to quit and why relapse remains possible. Also, it requires only a single, straightforward claim about the function of dopamine–that it constitutes the reward signal that is required for desire-updating.

References

Adams, C. D. (1981). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, 34B, 77-98.

Adams, C. D. and Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, 33B, 109-22.

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

Arpaly, N. and Schroeder, T. (2014). *In Praise of Desire*. New York: Oxford University Press.

Balleine, B., Daw, N and O'Doherty, J. P. (2008). Multiple forms of value learning and the function of dopamine. In: P. Glimcher (ed.) *Neuroeconomics: Decision Making and the Brain*. London: Academic Press.

Balleine, B. W. and Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37, 407-19.

Balleine, B. W. and O'Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology Reviews*, 35, 48-69.

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology*, 191, 391-431.

Berridge, K. C. (2012). From prediction error to incentive salience: mesolimbic computation of reward motivation. *European Journal of Neuroscience*, 35, 1124-43.

Berridge, K. C. and Robinson, T. E. (2011). Drug addiction as incentive sensitization. In: J. Poland and G. Graham (eds.) *Addiction and Responsibility*. Cambridge (MA): MIT Press.

Bratman, M. (1987). *Intention, Plans and Practical Reason*. Cambridge (MA): Harvard University Press.

Cagniard, B., Balsam, P. D.,Brunner, D. and Zhuang, X. (2005). Mice with chronically elevated dopamine exhibit enhanced motivation, but not learning, for a food reward. *Neuropsychopharmacology*, 31, 1362-70.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society B*, 362, 1585-99.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, 68, 5-20.

Foddy, B. and Savulescu, J. (2010). A liberal account of addiction. *Philosophy, Psychiatry and Psychology*, 17, 1-22.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108, 15647-54.

Gold, N. (2014). Team reasoning, framing, and self-control: An Aristotelian account. In: N. Levy (ed.) *Addiction and Self-Control: Perspectives from Philosophy, Psychology and Neuroscience*. New York: Oxford University Press.

Hare, T. A., Malmaud, J., and Rangel, A. (2011). Focusing attention on the health aspects of food changes value signals in the vmPFC and improves dietary choice. *Journal of Neuroscience*, 31, 11077-87.

Hnasko, T. S., Sotak, B. N. and Palmiter, R. D. (2005). Morphine reward in dopamine-deficient mice. *Nature*, 438, 854-7.

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

Holton, R. and Berridge, K. (2014). Addiction between compulsion and choice. In: N. Levy (ed.) *Addiction and Self-Control: Perspectives from Philosophy, Psychology and Neuroscience*. New York: Oxford University Press.

Hull, C. L. (1943). *Principles of Behavior: An Introduction to Behavior Theory*. Oxford: Appleton-Century.

Hyman, S. E. (2005). Addiction: a disease of learning and memory. *American Journal of Psychiatry*, 162, 1414-22.

Neal, D., Wood, W., Wu, M. and Kurlander, D. (2011). The pull of the past: when do habits persist despite conflicts with motives? *Personality and Social Psychology Bulletin*, 37, 1428-37.

Nestler, E. J. (2005). Is there a common molecular pathway for addiction? *Nature Neuroscience*, 8, 1445-9.

Niv, Y., Dayan, P. and Joel, D. (2006). The effects of motivation on extensively trained behavior. *Leibniz Technical Report*, Hebrew University 2006-6.

Pickard, H. (2012). The purpose in chronic addiction. *AJOB Neuroscience*, 3(2), 40-9.

Plassmann, H., O'Doherty, J. P., and Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience*, 27, 9984–8.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80, 1-27.

Skinner, B. F. (1938). *The Behavior of Organisms*. New York: Appleton-Century.

Tan, K. R., Brown, M., Labouèbe, G. et al. (2010). Neural bases for addictive properties of benzodiazepines. *Nature*, 463, 769-74.

Tricomi, E., Balleine, B. W. and O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, 29, 2225-32.

Wise, R. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience*, 5, 483-94.

Yin, H. H., Knowlton, B. J. and Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, 19, 181-9.

Yin, H. H., Knowlton, B. J. and Balleine, B. W. (2006). Reversible inactivation of dorsolateral striatum enhances sensitivity to changes in action-outcome contingency in instrumental conditioning. *Behavioural Brain Research*, 66, 189-96.

Yin, H. H., Ostlund, S. B., Knowlton, B. J. and Balleine, B. W. (2005). The role of dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, 22, 513-23.