ADDICTION BETWEEN COMPULSION AND CHOICE [1]

RICHARD HOLTON AND KENT BERRIDGE

Despite a wealth of recent empirical findings, the debate on addiction remains polarized along traditional lines. In one camp stand those who see the characteristic actions of the addict as driven by something very much like a disease: a pathologically intense compulsion that they can do nothing to resist. Over a century ago William James quoted an alcoholic giving powerful expression to this approach:

> Were a keg of rum in one corner of a room, and were a cannon constantly discharging balls between me and it, I could not refrain from passing before that cannon in order to get at the rum.[2]

At the same time this understanding of addiction was finding its way into literature. Oscar Wilde described the lure of opium on Dorian Gray in very similar terms:

> Men and women at such moments lose the freedom of their will. They move to their terrible end as automatons move. Choice is taken from them, and conscience is either killed, or, if it lives at all, lives but to give rebellion its fascination and disobedience its charm.[3]

Modern expressions tend to be less dramatic, but the basic conception remains much the same. Many contemporary theorists insist that the addict is in the grip of a brain disease that removes control over their actions and so requires treatment rather than condemnation.

In the other camp stand those who see addictive behaviour as involving normal choices, and so as something that takes place within the domain of ordinary intentional action. This approach sees an addict's decision to take a drug as motivated by a standard structure of beliefs and desires and still subject to self-control. This harks back to an earlier understanding but in recent years it has received new impetus in the hands of certain economists and behavioral psychologists.[4]

---

[1] This paper derives from two independent papers presented at the Oxford conference. Versions have since been presented by RH at MIT, Yale, Reading, Bristol, Oslo, Cambridge and NYU. Thanks to all the audiences at these places; and to Anna Alexandrova, Tony Dickinson, Olav Gjelsvik, Rae Langton, Neil Levy, Hanna Pickard, Drazen Prelec, Tim Schroeder, Gabriel Segal and Nick Shea.

[2] (James, 1890) Vol. II p. 543. It is unclear from the text whether this is a real quotation, or whether James simply made it up. And the case that follows it—of an alcoholic who supposedly chopped his hand off with an axe so that he would be given brandy—is very hard to credit.

[3] (Wilde, 1891) Ch 16. In describing it this way Wilde says he is 'following what psychologists tell us'. Admittedly there is much more going on in Dorian Grey than simple opium addiction; but into those depths we do not venture.

[4] The classical understanding of alcoholics simply saw them as people who were too fond of wine; the idea that addiction involved some kind of compulsion doesn't really take hold until the 18th century. For discussion see (Sournia, 1990).

Members of the disease camp point to the extraordinarily self-destructive behavior that addicts exhibit, and to the burgeoning literature that suggests that their brains are functioning in abnormal ways. Members of the ordinary choice camp point to findings that show that addicts often respond to incentives in normal ways. For example, most succeed in getting over their addictions by their mid-30s, often with minimal help.[5] Further, many addicts beyond that age stop taking drugs if the incentives are great enough and clear enough. Anesthesiologists and airline pilots who, having been once detected in their addiction, are required to pass random and frequent drug tests on pain of dismissal, are remarkably good at giving up.

The two approaches are typically seen as quite incompatible. If addiction is a brain disease, then there is no role for willpower or self-control. To take a representative example, the book from a recent television series lists as one of the 'seven myths of addiction' the idea that 'addiction is a willpower problem', and goes on to say:

> This is an old belief, probably based upon wanting to blame addicts for using drugs to excess. This myth is reinforced by the observation that most treatments for alcoholism and addiction are behavioral (talk) therapies, which are perceived to build self-control. But addiction occurs in an area of the brain called the mesolimbic dopamine system that is not under conscious control.[6]

We agree with the last sentence here; we agree that the mesolimbic dopamine system is centrally involved in addiction, and that the workings of that system do not appear to be under direct conscious control (in the sense that there doesn't seem to be much that one can deliberately do to directly affect the workings of that system). But it is one thing to say that people cannot control their mesolimbic dopamine system; quite another to say that they cannot control how it influences their actions. In a parallel way, there isn't much that people can deliberately do to influence their perceptual system, but that doesn't mean that there is nothing they can do to control its effects on their actions.

Our aim is to present a middle path. The findings from brain science are solid enough. There is good evidence that the brain of an addict is importantly different from that of a normal non-addicted individual—indeed there is even some reason to think that the addict's brain might have started out with a vulnerability to addiction. Certainly once addiction is under way, the desire for the addictive drug takes on a life of its own, with an intensity that is particularly, perhaps uniquely, high.[7] The desire becomes insulated from factors that, in normal intentional behavior, would undermine it, and so persists even when the addict knows that acting on it would be highly damaging. The addict may recognize that taking the drug again will incur the loss of family, friends, job, and most that makes life worth living, and yet still continue to take it. More surprisingly, addicts need not even like the thing that they are addicted to: they need gain no pleasure from it, nor anticipate that they will. Nor need they be

---

[5] This point is made very forcefully in (Heyman, 2009), Ch 4. He draws his conclusion from examination of national population surveys—not just surveys of addicts. He argues that most of those who remain addicted do so because they suffer from other psychiatric illnesses.

[6](Hoffman and Froeke, 2007) p. 37 (accompanying an HBO TV series).

[7] We speak in terms of 'drugs' here as a shorthand for 'addictive substance', even though some such substances—most obviously alcohol—are not typically thought of as drugs outside the biomedical community.

motivated by a desire to avoid the horrors of withdrawal. Alcohol or heroin addicts often relapse long after withdrawal is over, and cocaine addiction is no less potent for having a relatively mild withdrawal syndrome. Addicts may relapse when they see nothing good in their drug whatsoever. They may see it as nasty, damaging and worthless in every respect. Yet they may still want their it, and want it, moreover, in a particularly immediate and intense way—perhaps more immediately and more intensely than most other people ever experience.

There is another way in which an addictive desire does not typically function like a desire to see the Pyramids or to get a paper finished before the weekend. It does not serve as an input to deliberation, something to be weighed, along with other competing desires, in deciding what to do. Instead addictive desire functions as something more like an intention: as something that, unless checked, will lead, in a rather direct way, to action. This combination of features—the insulation of addictive desires from factors that should undermine them, and their tendency to lead directly to action—means that addictive behavior is very different from ordinary behavior that results from deliberation.

Nevertheless, the intensity and power of an addictive desire does not mean that addicts are automata, standing powerless spectators as they are moved by their desires. For whilst addictive desires are very strong, the human capacity for self-control is also highly developed; much more developed, it seems, than in rats. Addicts do not actually cross into the paths of cannonballs or their equivalents, despite William James' colorful assertion. They go around or wait for a lull. Smokers on aeroplanes postpone their urge to smoke until the flight is over.

So addictive urges are not entirely uncontrollable: as these cases show, they can be controlled, at least for a short while, and sometimes for longer if the stakes are high enough and clear enough. The experience of self-control that everyone has at certain moments is a veridical one: self-control is a real phenomenon, something that can be used to control acting on addictive desires, even if at a considerable cost and, for most addicts, subject to occasional failure. We should thus not be thinking of addictive desires as things that are impossible to resist, but as things that are very difficult to resist.[8] Our moral evaluations should reflect this fact, and our scientific account of addiction should explain why resistance is difficult and why failure happens on the occasions it does.

Our aim here is to articulate such a model, one that explains why addictive desires have the distinctive features they have, but that also explains how they can be controlled. We start by outlining what we think is wrong with the two extreme positions, the pure choice model and the pure disease model.

I. ORDINARY CHOICE MODELS

We cannot hope to survey all of the different ordinary choice models here, but some brief comments will serve to show why we think that they cannot provide a complete explanation of addiction. An ordinary choice model can, of course, easily explain the behaviour of those who willingly and knowingly take addictive drugs. But addicts frequently say that they have been somehow captured by the addiction—that they wish that they were could escape it but that

[8] Could they sometimes be truly irresistible? It seems rash to rule that out, although it is hard to be sure quite what the claim means: That no incentive *would* overcome it? That no incentive *could*?

something is making it very hard for them to do so. Some listeners might dismiss these comments as disingenuous or self-deceived, but we think there is something in what they say.

How can ordinary choice models make sense of this capture? They have two approaches. One is to ascribe to addicts abnormal desires; the other is to ascribe to them mistaken beliefs. Advocates of the first approach typically see addicts as having steep temporal discount curves —they see them as having much stronger desires for the present and immediate future than for the more distant future. Since addictive drugs normally involve a nasty period of withdrawal, an already addicted agent whose focus is on the immediate future will want to avoid embarking on the suffering that such a process will involve, even if they know that the long-term effects will be beneficial. Of course they might well prefer not to have started consuming the drug in the first place—in this way advocates of this approach can make sense of the idea that they are really addicted and are not simply willing consumers—but given the state that they are in now, continuing to take the drugs is preferable to withdrawal.[9]

Advocates of the second approach typically see addicts as mistaken, at least initially, about the effects of their drugs (they believe that they will not become addicted, or that addiction will not be so bad); or they see them as failing to take into account the consequences of current consumption for their future state: by focussing only on their current options, addicts fail to see that consuming addictive drugs now will lower their overall well-being in the future.[10]

The two approaches may be combined: mistaken beliefs might explain why addicts fall into addiction, and then the steep discount curves might explain why they stay there; and elements from these approaches might be used to supplement other accounts. Indeed we ourselves are inclined to think that there are important insights to be had here. In particular, there is good evidence that ignorance has an important role in the process of acquiring an addiction. But we do not think that an ordinary choice account can provide the fundamental explanation of what is distinctive about addiction. For if they were right, then a former addict who had been through the pains of withdrawal should be the least likely to consume again. They would no longer have the cost of withdrawal to endure; and they, of all people, would be well informed of their own vulnerability to addiction, of how nasty it is, and of the cost of not looking to the future. We are not talking here of the person who really prefers to be addicted; they will just start consuming again, although such a person would be unlikely to have put themselves through the process of withdrawal in the first place. But the person who genuinely wanted to be free of the drugs should be uniquely well qualified to ensure that they remain so.

Yet that is not what we find. People who have come through withdrawal, and gained much self-knowledge in the process, are much more likely to take up drugs again than those who never started, a process that is typically triggered by cues that are associated with the previous addiction. Indeed, withdrawal seems largely irrelevant in the process of maintaining addiction. Not only do people consume again after having gone through it, cravings are experienced long before it comes in, and some highly addictive drugs—most notably, cocaine—have minimal

---

[9] The most influential presentation of this line is from Becker and Murphy; for a simplified presentation see (Skog, 1999). Becker and Murphy give no explanation of how addicts get into the state of addiction; that is left to be explained by exogenous factors.

[10] See (Loewenstein, 1999) and (Herrnstein and Prelec, 1992) for versions of the first approach; and Heyman, 2009) Ch 6, for a detailed development of the second. An addict, understood on Heyman's lines as one who fails to think about their future, will be behaviorally equivalent to the steep discounter who doesn't care about it; but this will derive from features of their beliefs rather than of their desires.

withdrawal symptoms. A pure choice model struggles to explain these features. So let us turn to the disease models that do better with them.

II. DISEASE MODELS

There are many disease models of addiction. To get some traction on the debate, we divide these into four, at the cost of some simplification. The first, exploiting classical behaviorist mechanisms, sees addiction as a habit: drug-taking actions are triggered automatically in particular situations, independently of the subject's beliefs and desires. The second sees it as involving distorted pleasure: addictive drugs 'hijack' the subject's pleasure circuits, and it is this that causes the skewed behaviour. The third, using reinforcement learning theory, sees the distortion as affecting not the pleasure itself, but the subject's *beliefs* about what will give them pleasure. The final account, which we shall endorse (whilst denying that this provides the *whole* story about addiction), involves desire: consumption of addictive drugs gives rise to pathologically intense desires or cravings, states that are largely insulated from the subject's beliefs and other desires. We start by briefly outlining those with which we disagree.

Habit accounts

In its simplest form the habit model follows the classic stimulus-response account that was laid down in the early 20th century by Thorndike, and that became the staple of behaviorist models. An agent explores its environment, gets a positive reaction to some things and an aversive response to others, and subsequently comes to repeat those behaviors that produced the positive outcomes. In its early behaviorist guise, this approach was linked with skepticism about positive/aversive mental states altogether; but such an approach has few supporters now, and we shall say nothing about it. More interesting is the idea that habits stand alongside, but independent from, the agent's beliefs and desires.[11] Contemporary versions of habit theory hold that drugs induce brain systems of action (e.g. in the neostriatum) to form the tendency in the presence of drug cues to perform particular behaviours, behaviours that have been established previous drug taking episodes—much like a shoe-tying habit but even more strongly automatic.

If addictive states were understood this way that would provide some explanation of why they are insensitive to the addict's desire to stop. But the habit account assumes that drug taking is unmotivated, and most likely to surface when the addict's attention is distracted elsewhere. That belies the intensely motivated nature of addictive urges, and turns upside down the observation that attentively thinking about drugs is the most dangerous situation for an addict—not thinking about something else. Whilst some aspects of habitual behavior might be important in addiction—reaching unthinkingly for a cigarette—the account cannot easily explain why an agent will take a drug in full awareness of what they are doing, but quite contrary to their views of what is best.[12]

---

[11] For accounts along these lines see (Wise, 2004) and (Everitt *et al.*, 2008).

[12] For work on the areas in which pure habit accounts do provide good explanations see (Wood and Neal, 2007).

Pleasure accounts

So let us move the second class of accounts, those premised on excessive pleasure.[13] Clearly many addicts do get great pleasure from the drugs they take. If drugs can 'hijack' the pleasure circuit, giving a disproportionate amount of pleasure to those who consume them, then this would give rise to a very strong learned desire for them.[14] And if the pleasure per unit decreased over time, as tolerance developed, the agent would want more and more of the drugs to compensate.[15]

This account was once thought to be bolstered by the finding that the addictive drugs have an impact on the mesolimbic dopamine system: either by stimulating the production of dopamine (in the case of amphetamine, nicotine, caffeine); by reducing the production of substances like GABA that themselves reduce the amount of dopamine released (opiates, perhaps THC); by reducing the level of substances that break down dopamine (alcohol); or by reducing the activity of the system that reabsorbs dopamine (cocaine, and perhaps amphetamine). Add the premise that the mesolimbic dopamine system is the pleasure system, and we have what looks like a compelling picture. [16]

Simple and straightforward though the pleasure account is, it doesn't fit the empirical findings. It assumes that the dopamine system is concerned with liking. But a host of findings have now shown fairly conclusively that the primary role of the dopamine system is not to do with liking. In rats, suppressing the dopamine system does not result in a lack of pleasure responses to sweet substances; we shall discuss cases of this shortly. Likewise, human subjects whose dopamine systems are suppressed artificially, or as a result of Parkinson's disease, give normal pleasure ratings to sugar. Conversely, elevated dopamine levels in rats do not result in greater pleasure. And elevated dopamine levels in human subjects do not give rise to increased subjective pleasure ratings.[17] Dopamine thus does not seem to be directly concerned with the production of liking. We will suggest that it is concerned with the creation of wanting.

This might not matter if there were nonetheless a very tight correlation between liking and wanting: if liking invariably resulted in wanting, and if wanting were invariably the result of prior liking. But the very results that show that they are distinct states also show that, whilst they might *typically* be linked by causal connections, sometimes those connections will fail. We will argue that this is crucial for understanding addiction.

---

[13] Thorndike's original account of learning was in terms of pleasure, though he later came to talk purely in terms of stimulus and response. Historically then, pleasure based accounts represent something of a reversion to an earlier idea.

[14] We speak of 'hijacking' and 'disproportionate pleasure' here, but of course accounts that think that there is no rational constraint on what gives one pleasure will find it hard to make sense of this. To that extent, this approach will lapse back into a rational choice account, in which the agent acts on desires for their strongest pleasure.

[15] See, for instance, Roy Wise's earlier work: (Wise, 1980) and (Wise, 1985).

[16] For a recent popular presentation of such an approach by a neuroscientist, see (Linden, 2011) Ch. 2. Linden writes 'Addictive drugs, by co-opting the pleasure circuitry and activating it more strongly than any natural reward, create deeply ingrained memories that are bound up in a network of associations.' p. 53.

[17] For details, see (Berridge, 2012) p. 1132.

Learning accounts

So let us move to those models that see addiction as resulting from *learning*. Admittedly in a simple behaviorist model learning is not a very contentful notion: there isn't much more to it than the idea that a subject's behavior changes as a result of what happens to them, and hardly anyone could disagree that that is true of addiction. But in more cognitivist models, the idea of learning is much more specific: it is the idea of forming predictive associations, that is, *beliefs*.[18]

These accounts see addiction as stemming, not from heightened pleasure itself, but from mistaken *belief* about pleasure. Addictive drugs hijack, not the pleasure circuits, but the circuits that *learn* about pleasure, and so they distort the memories that are used to guide future desires. One popular theory of reward learning holds that dopamine spikes indicate 'reward prediction errors': dopamine is released whenever an outcome is better than expected.[19] Applied to addiction, the idea is that dopamine-stimulating drugs cause an exaggerated prediction error: it is as though the drugs were much more pleasurable than expected.[20] Consumption of the drug itself doesn't have to be especially pleasurable—though it may be—since the effect on the dopamine system is to trigger a large prediction error *as if* it were pleasurable, with the result that the 'memory' of the pleasure can greatly exceed the actual pleasure. This in turn gives rise to the extreme desires that characterize addiction. On this approach then, the addict's fundamental desire is a desire for pleasure. Since, at some level, the addict mistakenly believes that consumption of the drug will give them pleasure, this results in a strong instrumental desire for the drugs.[21]

We think that this is mistaken. We will present instead a model—the incentive salience model—that sees addiction as driven by desires that have no essential connection with beliefs about what will be liked, or about what will be beneficial in other ways. The key idea here is that the dopamine signals are not learning signals, in the sense that they do not give rise to beliefs, predictions or memories (real or apparent) at all. Instead, they give rise to desires directly—or, more accurately, to a sensitivity to experience desires when cued with appropriate stimuli. The desire felt is not an instrumental desire, driven by an intrinsic desire for pleasure;

---

[18] At least, this is the core notion of learning, that is present when one says that a subject learns *that* something is the case, or learns *who*, or *what*, or *where*. In all such cases the subject acquires a belief. We also speak more broadly of learning *to* do something, and in this sense our account could be phrased in terms of *learning to want*. But to avoid confusion, we'll talk only about acquiring wants. A further linguistic complication: in most ordinary talk, 'learning *that*' and learning *who*' are factives: the belief that is formed must be true. We'll follow the standard psychological use and talk of learning even when the resulting belief is false.

[19] (Sutton and Barto, 1998).

[20] (Schultz *et al.*, 1997); (Redish, 2004).

[21] We say 'at some level', since many proponents of the prediction error approach insist that their account is 'model free', by which they mean that the subject does not have a full-blown representation or cognitive map of the world and of their own preferences within it. Nonetheless, we insist that if talk of 'prediction' is appropriate, there must still be beliefs, even if they are of a partial, local, or implicit form; otherwise we would simply simply have a habit account. The other possibility is to understand the relevant states as desires, the approach that we develop below. Of course treating predictions as beliefs is compatible with the idea that subjects also have more explicitly articulated beliefs about their own preferences.

instead, it is an intrinsic desire for the drug, a desire that may lead to action even in the face of contrary desires, and in the face of beliefs that consumption will have bad consequences. Whilst the incentive salience account can embrace a parallel formalism to that employed by the prediction error model—we explain how below—it uses that formalism to explain the formation of desires and not of beliefs.

Before we explain the evidence for such an account in any detail, let us get clearer on the distinctions we have just outlined: that between wanting and liking; and that between the formation of beliefs, and the acquisition of desires.

III. DISTINGUISHING WANTING AND LIKING [22]

In one sense it is obvious that wanting and liking are distinct, at least if we think of liking in terms of pleasure: wanting typically comes before one gets the thing wanted, whereas the pleasure typically (though not invariably) comes once one has got it. And liking and wanting can also come apart as a result of false beliefs. We can want something that we believe we will like, even though we won't in fact like it: perhaps we haven't tried it before, or have forgotten that we didn't like it, or we believe for some reason that our reaction will be different to last time.

For parallel reasons we can like something and not want it: we might not realize that we like it, or we might have other reasons for foregoing it. Indeed, those disciplines that have not traditionally made much of the distinction between wanting and liking—behaviorist psychology say, or revealed preference economics—have not normally identified them. Rather they have thought that they could make do with one (typically wanting) whilst discarding the other (typically liking) as illusory or scientifically intractable.

So the real issues do not concern the *identity* of wanting and liking. Instead we think that they are two-fold. One concerns the *causal* relations between wanting and liking, and their embodiment in particular brain mechanisms. The second concerns the relation of wanting to *expected* liking. We take these in turn.

Causal relations between wanting and liking

Does liking invariably cause wanting? (I.e. is liking causally *sufficient* for wanting)? Are increases in wanting always preceded by incidents of liking? (I.e. is liking causally *necessary* for wanting?) It is commonly supposed that there are some such relations here. Indeed, even so implacable opponent of hedonism as G.E. Moore wrote that he was 'ready to admit that pleasure is always, in part at least, the *cause* of desire.'[23] But that is a substantial claim. Whilst we think that brain activations that cause increases in liking *typically* cause increases in wanting

---

[22] In discussing the incentive salience model, one author (KB) has in previous writings been careful to distinguish the notions of wanting and liking that are involved from our ordinary folk notions. For that reason he has placed them in inverted commas. The other (RH) has no such scruples, and he has prevailed here. But note that the kind of wanting involved here needs to be distinguished from other kinds that have equally good claim on the term; and that certain features that might be expected—that one knows what one wants, that one judges it to be worthwhile—will often be absent.

[23] (Moore, 1905) §42

too, we think that these mechanisms are in principle separable, and that under some conditions liking can in fact be generated without wanting.

When we turn to the converse question of whether wanting is always preceded by liking, our answer is more straightforwardly 'no'. Many brain activations that cause wanting are not accompanied by increased liking; wanting without liking occurs frequently in addicts. The evidence here came originally from studies of the brain activity and behavior of rats. Since rats can't talk we need to have some non-verbal behavioral indicators of wanting and of liking. Wanting is straightforward: rats want something if they try to get it. (This is where we assume that issues of self-control will not intrude; things are more complicated with human beings as we shall see later.) Liking has been traditionally viewed as more harder to identify. But a set of results indicate that a range of evolved facial expressions—including tongue protrusions and lip sucking—are correlated with liking for the sensory pleasure of tastes across a wide range of species including rats, monkeys and human infants.[24] In the past decade, the distinction between liking and wanting has also been confirmed in a number of human studies based on ratings of their own experience of sensory pleasures, such as cocaine and other addictive drugs.[25]

Once we have distinct criteria for wanting and liking, we find that one can be induced without the other. If rats' dopamine levels are suppressed, they are no longer prepared to work to gain food rewards that they would previously have worked for. At the extreme, they will not eat pleasant foods that are freely available, even though they still display strong liking for them once the foods are placed in their mouths. Indeed, rats who had 98% of the dopamine neurons in their nucleus accumbens and neostriatum chemically destroyed would have starved to death had they not been intragastrically fed, yet their normal liking reactions indicated that pleasure in the food was unchanged. So liking is not sufficient for wanting. Conversely, by boosting rats' dopamine levels we find that their wanting can be increased without their liking being increased—we will discuss an example of this shortly.[26] So increased liking is not necessary for increased wanting. Indeed wanting can be artificially engendered in rats without any signs of liking.[27]

Relation of wanting to expected liking

The second issue concerns the relation of wanting to *expected* liking. Can subjects want something whilst believing they will not like it? And conversely, can they believe that they will like something and not want it? This is where the talk of learning fits in: can subjects come to learn that they like something, and yet not go on to form a desire for it? And conversely, can they come to learn that they dislike something, and yet go on to form (or at least maintain) a desire for it?

---

[24] (Berridge, 2000); (Berridge and Kringelbach, 2008). The reaction seems to be suppressed in humans after early infancy, but can recur in those suffering from Alzheimer's.

[25] (Leyton, 2010); (Lawrence *et al.*, 2003).

[26] See also (Berridge, 2007); (Smith *et al.*, 2011).

[27] (Peciña *et al.*, 2003); (Wyvell and Berridge, 2000); (Faure, *et al.*, 2010); (Smith *et al.*, 2011); (Tindell *et al.*, 2005); (Berridge and Valenstein, 1991).

Here again the empirical evidence suggests that wanting without expected liking is indeed possible, and so the two cannot be identified, nor are they invariably causally connected. There are two kinds of consideration. First, wants fluctuate in ways that are hard to mesh with the idea that belief is also fluctuating. Second, what we know about the formation of belief suggests that it uses quite different mechanisms to those involved in the formation of wants.

On the first point, consider a set of experiments done by Cindy Wyvell, which produced momentary pulses of elevated wanting for rewards that exceeded both liking for the rewards and learned memories and expectations of the reward's value. The initial stage was to get rats to associate a random stimulus (a noise), and an activity (lever pressing) with each other by pairing each with a sugar reward. As a result, the noise tended to trigger the lever pressing. The experiment was then to see the effect of changes in dopamine level on this triggering, even when the sugar is not present.[28]

On some days the rats were trained to press a lever to gain sugar (instrumental training). On others a conditioned stimulus was created: a sound heralded freely available sugar, which resulted in the rats associating the sound and the sugar (Pavlovian training). Their facial responses showed that they liked the sugar.

Cannulas were inserted into the rats' brains enabling their mesolimbic dopamine systems to be affected directly by microinjections of tiny droplets of drug. A control group received an inert substance through this cannula, whilst the other group received amphetamines, which greatly increase dopamine release. The effects of the action of the mesolimbic dopamine system could then be determined by observing the differences between the two groups.

Both groups continued to like the sugar. As expected, they liked it to the same degree: the amphetamine group did not show an increased facial pleasure response when given it, further evidence that dopamine does not produce pleasure. Importantly though, the amphetamines did not seem to increase *anticipated* pleasure from the sugar: when given the lever to press, the amphetamine group did not press it any more frequently than the controls when freely allowed to without any distraction.

The difference came when the rats heard the noise that they had been conditioned to associate with sugar. Now both groups increased their lever pressing. But rats in the amphetamine group pressed the lever dramatically more: more than four times as frequently as before, and more than 50% higher than the rats in the no-amphetamine group who heard the same noise. And this effect was switched on and off as the cue went on and off. [29]

What was happening here? It appears that the increased dopamine levels resulted in a massive amplification of the conditioned response that was already present. Hearing the cue signal caused the control rats to press the lever. But the presence of high levels of dopamine caused the group that is receiving the amphetamines to press it far more.

It is very hard to explain this result in terms of changes in expectation, for we have no reason to think that hearing the signal caused change in the rats' beliefs about how pleasurable the sugar would be, and since the amphetamine rats did not elevate their effort to earn reward in the absence of the particular noise. The rats were not learning anything new; and the effects fell off as soon as the tone ceased.

---

[28] See (Robinson and Berridge, 2003) pp. 41–3, for further discussion of why this feature is important.

[29] It's an interesting question why the sight of the lever didn't itself work as a cue. Clearly not all cues are created equal.

In general, learning seems to be different from wanting. For example, the rats discussed above who have lost nearly all of their mesolimbic dopamine due to neurochemical 6-OHDA lesions are still quite capable of learning new values about food rewards. When a previously liked food is made unpalatable by inducing nausea, the dopamine depleted rats will learn to react to it with signs of disgust, in just the same way as normal rats. Similarly, mice who have been genetically engineered to lack dopamine are still able to learn basic Pavlovian reward associations.[30] When learning and 'wanting' are made to diverge by manipulating a second input to 'wanting' (physiological state/trait) levels and roles of dopamine and mesolimbic brain activations all track 'wanting' outputs much more faithfully than learning inputs.[31] Learning, in the sense of the formation of new beliefs or of the formation of new behaviour does not seem to be essentially dependent on the dopamine system.

So what exactly is dopamine doing? As we saw from the Wyvell experiments, it is involved in the generation of desire at particular moments. But as those experiments showed, this is not blanket desire. Dopamine seems instead to be involved in producing specifically targeted desires in response to certain stimuli. To see what it might be doing, let us start by employing some relatively *a priori* considerations about creatures like us and about the kind of wanting system that we would need. This is not to tell an evolutionary just-so story—we think that the account we give is independently well supported by the empirical evidence. Rather we tell it because a good way of understanding a complex mechanism is by understanding its function.

IV. MODELING THE WANTING SYSTEM: SOME *A PRIORI* CONSIDERATIONS

Some creatures are tightly locked into a specific pattern of consumption: an insect that eats the leaves from a single plant species, or a koala that eats the leaves from four. Such creatures can have their tastes hard-wired. Other creatures are more opportunistic, adapting their consumption patterns to what is available. Human beings, like rats, are at the far end of this continuum. Although some of our desires are perhaps hard-wired, most are highly plastic.

Let us think in the abstract about how a creature with plastic desires will structure its consumption. We assume that it has some way of telling, when it samples a given food, how good that food is in providing it with what it needs. It might do this largely by means of a pleasure mechanism—the better the food, the more pleasure it gives—although in fact we think that, for reasons we shall discuss shortly, pleasure is not always involved. So let us just say that the creature can register how good the food is for it. Suppose then that the goal of the

---

[30] For summary see (Berridge, 2012) 1139–40.

[31] (Flagel *et al.*, 2011); (Tindell *et al.*, 2005); (Smith *et al.*, 2011); (Robinson and Berridge, 2010); (Saunders and Robinson, 2012). There are a number of highly influential papers by Schultz and others that appear to demonstrate that dopamine firing does track learning: see, for instance, (Schultz *et al.*, 1997) and (Schultz, 2002). We suggest that these finding may have been flawed by an experimental confound. Those studies allowed only the learning input to vary, while clamping the second physiological input as stable. Under those conditions, wanting and dopamine as outputs naturally track the only input that was allowed to vary: learning. The wanted output then mimics the learned input, the two signals cannot be told apart, and an observer can confuse one with the other. For further discussion see (Berridge, 2012) p.1132.

creature is to maximize its consumption of things that are good for it. How could it go about that?

One way would be for the creature simply to try each thing that it comes across to see how good it is and then consume it if it is; but obviously that would be highly inefficient, since it would involve constantly retrying things which had already been shown to be bad. A second would be to learn what is good for it, in the full sense of that term: the creature would develop beliefs about which foods are good for it, and then, given a desire to consume what is good, it would form instrumental desires for those foods.[32] A third possibility would be to avoid forming the beliefs at all. Instead the creature could directly form its desires on the basis of what it had discovered to be good. That is, it could form intrinsic desires for the good foods, without recourse to any beliefs or predictions about them.

This third possibility would have some advantages. It could be simpler and easier to implement than a belief based system, and in some ways more robust. So let us consider how it might work. The creature we will consider will need to do two things. Firstly it will need to form its desires for certain foods; and then it will to act on them in the presence of those foods. To do this it will need to make use of two systems, a desire formation system which creates intrinsic desires for foods, and a consumption system which regulates the creature's consumption in accord with those desires. We should keep these systems conceptually distinct, even if, for reasons we shall come to shortly, there may be some overlap of function. So spelling this out we have:

(i) *a desire formation system*. This will need to identify each sample as belonging to a certain food type; to determine how much goodness it gets from that type; and, on the basis of that determination, to send a signal to the consuming system that will regulate its subsequent desire for foods of that type. Let us call this the *A-signal*.

(ii) *a consumption system*. The settings of this system will be determined by the A-signal. Presented with a potential food, it will need to identify it as belonging to one food type or another, and then it will respond to this by sending out, in accordance with its setting for that food, a signal that regulates consumption of that food. Let us call this the *B-signal*.

Quite how to use the term 'desire' here is delicate (so far we have fudged it). The term could be used to label the *dispositional state* of the consuming system: its tendency to respond to a certain food in a certain way, as determined by the A-signal. Alternatively it could be used to label the state that comes about as a result of the triggering of that dispositional state: the *occurrent state* that is the B-signal or that is brought about by the B-signal.[33] Both uses have good philosophical and ordinary language pedigree. To mark the distinction, let's call the first the

---

[32] (Dickinson and Balleine, 2010).

[33] We won't broach the question of whether the B-signal will actually *be* the desire (or its neuro-physiological correlate, if that is meant to be different). We don't see any compelling arguments, other than an appeal to ontological economy, for or against that claim. But to keep things simple, and to avoid getting involved in issues of the nature of phenomenal states and so forth, we will just talk about the B-signal as the immediate cause of desire, with the thought that the identity claim may be substituted. Obviously if the B-signal *is* the desire, it is rather misleading to call it a 'signal'.

dispositional desire, and the second the *occurrent desire*.[34] The desire formation system is thus in the business of forming dispositional desires; and the consumption system is in the business of forming occurrent desires when the dispositional desires are triggered.

Let us inquire a little further into how these systems might work. Since there is likely to be considerable variation in the goodness of different samples of the same food (as we all know, whilst some strawberries are delicious, others are quite tasteless) an obvious strategy will be for the creature to sample each food several times over, to compute the mean goodness it gets from the set of samples of each food, and then proportion its desire to that mean. How would it do this?

We'll first address the issue of how it computes the mean from a given set of samples of a certain foodstuff. It could, of course, record the value of each sample taken so far, sum them, and divide the result by the number of samples. But keeping on computing the mean in that way would involve keeping track of a lot of data that aren't needed, which would be far from easy for a biological system to implement; and it would mean that there would be no provisional result until the sampling period was over.

A far simpler method, which does provide provisional results, is to keep a rolling mean and a record of the number of samples taken. Then the creature can update the mean with each new sample recorded. Suppose it has so far examined n samples, has identified the value (V) of each, and has computed their mean ($mean_n$). Now it takes a further sample, n+1, with value V_{n+1}. To compute the new mean (i.e. $mean_{n+1}$) all it needs to do is to see how much the value of the current sample differs from the existing mean, give this difference the right weight by dividing it by the total number of samples, and then add it to the existing mean. Or more concisely:

$$mean_{n+1} = mean_n + (1/(n+1)) [V_{n+1} - mean_n])^{35}$$

This update-rule, which looks at the difference between the current sample and the previous mean—or, in other words often used in the literature, at the error between the prior prediction and the current sample—provides the core of most reinforcement learning methods; and certainly a mechanism based upon it can be used to acquire new beliefs. These are the learning systems that we mentioned above when discussing the learning interpretations of the role of dopamine. But such a rule need not be tied to a learning system. The very same rule can be used in this framework to form new desires directly.

To employ this rule the desire formation system will (i) need to recognize the sample it is encountering as belonging to a certain foodstuff; it will (ii) need to retrieve the mean level of goodness for that foodstuff ($mean_n$), and the previous number of samples (n); it will (iii) need to register the level of goodness gained from the current sample (V_{n+1}); and then it will (iv) need to perform the computation. In fact the desire formation system doesn't need to send the mean on to the consumption system. All it needs to do is to send the new information gained

---

[34] As we shall see, dispositional desires understood in this way are rather different to dispositional desires as they have normally been understood in philosophy, since they have a very specific cue. Without the cue they have no force. In that sense then they are more like dispositions to have desires than like dispositional desires as traditionally conceived.

[35] See (Sutton and Barto, 1998) pp. 36–7 for the simple proof that this is equivalent to the more normal way of defining the mean.

from each new sample, so that the consumption system can modify its dispositional desires in the light of this. To do this it need only to send a signal corresponding to $(1/(n+1))\ [V_{n+1} - mean_n]$). In the learning literature this is commonly termed an 'error signal' or a 'learning signal', but since we are working with a desire based system, rather than a belief or prediction based system, and hence are not concerned with learning in the strict sense, we will stick with our earlier stipulation and just call it the *A-signal*.

The consuming system will now employ the A-signal to regulate its activity by setting its dispositional desires. How will it do this? One possibility would be for the desire formation system to send, in addition to the A-signal, information about the identity of the thing that had been sampled. But sending such information would be a complicated business, and, besides, it would be largely redundant. Since in order to employ the information the consumption system will anyway need to be able to categorize potential food as belonging to one of the food types, it could instead use a basic associative mechanism. The consumption system would thus only receive information given by the A-signal. It would in turn impute this information to whichever food it identified as being currently consumed. This of course will mean that it is vulnerable to a certain kind of error: if it receives the A-signal at the same time that it identifies a given food as being consumed, it will impute the information to that food even if it is not in fact the source of the signal, or not the source in the standard way. This fact will be crucial to our account of addiction.

Let us summarize then: the consumption system will set its dispositions—its dispositional desires—on the basis of two inputs, the strength of the A-signal and its own identification of what is being consumed at the time it gets the A-signal. On the basis of these dispositions it will send out an appropriate B-signal whenever it recognizes a food as belonging to a certain group. That B-signal will in turn determine the pattern of consumption.

The model that we have presented sounds much like a simple version of an actor-critic model: the consumption system is the actor, and the desire formation system the critic.[36] Rival models combine the two roles. Such an approach might seem attractively economical. In particular, couldn't a single signal serve both to lay down dispositional desires, and to induce occurrent desires? We don't mean to rule out such an approach, but we will persist with the two system approach for two reasons. First, as we mentioned before, even if the systems are not realized distinctly, it is useful to keep them conceptually distinct. Second, there is some good evidence that rats beings employ an actor-critic model; we will come to this in the next section. Third, there is a real advantage to a creature in keeping the two systems—and hence the A-signal and the B-signal—distinct.

To see this consider the role of appetite. No creature will gain by going on consuming even when it no longer needs to. Once it has eaten to capacity it should stop. Better still, it could regulate its appetite depending on its current needs. If it has plenty of one nutrient and not enough of another it should increase its desire for the latter relative to the former. The advantage of the two system model is that it enables the creature to do this, whilst still gaining useful information. A simple—doubtless over-simple—example will make the point. Suppose that our creature's sugar needs are sated, but that it still needs water. It samples a new foodstuff and finds it rich in sugar but quite without water. It shouldn't form an occurrent desire for the food; it should go on searching. But it would be good if the information that this food is rich in sugar could have some impact on its dispositional desire for it; after all, next time it comes

[36] See (Sutton and Barto, 1998) pp. 151ff.

across it it may be short on sugar. Having a system that keeps separate the A-signal and the B-signal enables it to make just this distinction. The idea is that appetite is regulated at the level of the consumption system, i.e. by controlling the B-signal. So when the creature in the water deprived state encounters the sugar-rich substance, its A-signal can still fire, laying down the dispositional desire. But because it is not hungry for sugar, the B-signal will not fire, and so there will be no occurrent desire. In contrast, if both dispositional and occurrent desire were regulated by a single signal, it is hard to see how one could be triggered without the other.

But there is a further issue that is potentially important. The model that we have presented is good, if a little slow, at responding to relatively stable changes—a food getting better or worse, and staying that way for some time. In addition though the world is likely to exhibit some unstable changes—different samples of the same food fluctuating, sometimes wildly, in their goodness (a phenomenon most of us are all too familiar with from tomatoes and strawberries). Thus suppose that a creature has tried a certain foodstuff many times in the past, and found it middling good: worth consuming if there is really nothing else available, but not so good that the creature should keep consuming it rather than exploring elsewhere. Now it tries the same food again and finds it extremely good. How should it respond? Clearly it should increase its desire for future samples of the foodstuff, though not precipitously: its prior experience suggests that this will be an anomaly, and that future samples will not be as good as this. But at the same time it would be crazy not to make the most of this exceptionally good sample. It should consume all it can of it here and now, rather than moving on to explore elsewhere, or to consume a different food that it has found on average to be slightly better, but that is likely to be considerably worse than this sample.

This requires then that the creature have a two-part response—a large but short lived burst in desire for this particular sample, and a smaller but more stable increase in desire for future samples of the same type. But if, for the reasons we have discussed, occurrent desire is regulated by the B-signal and not by the A-signal, how can this be achieved?

We see one possibility. Even if the A-signals do not directly cause occurrent desires, they might nevertheless boost the effectiveness of the signals that do. So even if only a B-signal can cause an occurrent desire, the degree to which it does is regulated by the strength of the current A-signal. The higher the current A-signal, the higher the effectiveness of the B-signal. Call this the 'accelerator' approach. It might be implemented in different ways. One particularly simple way would be to make the change imparted to the dispositional desire by the A-signal into a temporally stepped one. Suppose that the initial impact of the A-signal on the dispositional desire is comparatively large, but that this decays rapidly to a reach a new lower equilibrium after a short while. That would give us the accelerator approach, and would be quite compatible with standard Hebbian models.[37]

We have gone about as far as we can go using *a priori* considerations. But before we return to consider the empirical findings, let us summarize. A creature with a flexible wanting system can change its intrinsic desires so that they are focussed on the things that have given it pleasure, or have otherwise benefitted it, in the past. If this system employs two separate signals, an A-signal that forms the dispositional desires, and a B-signal that gives rise to occurrent desires when these are triggered, it will be able to change its dispositional desires

[37] (McClure *et al.*, 2003) also give dopamine two different roles, but the second is not the same as that suggested here. Rather than serving to regulate dispositional desires, it is involved in 'learning to predict future rewards'.

even when it is not hungry. And if the A-signal can also boost the power of the B-signal, it can work to avail itself of an unrepresentatively good sample of a food, without changing the dispositional desire disproportionately. We don't want to be tied to the details of this account. But the empirical evidence suggests that something along these lines is correct.

V. EMPIRICAL EVIDENCE FOR THE WANTING SYSTEM

We start by returning to the findings of Cindy Wyvell and to some related experimental results. As discussed above, she found that injection of amphetamine into rats' mesolimbic dopamine systems caused, in the presence of the reward cue, huge increases in short-term wanting. But this was not all. She also ran a parallel set of experiments on rats who had received earlier amphetamine injections, rather than infusions into their brains at the time of the stimulus. She found that this sensitized their brains in an apparently permanent way. Despite being free of the drug for ten days, the conditioned stimulus of the sound still elicited twice the frequency of lever pressing from these rats as it did from a control group who had not received the sensitizing injections. This behaviour could not have resulted from the elevated dopamine levels caused directly by amphetamines, since the rats received no amphetamines when they heard the sound; it looks instead that it was caused by the structural changes produced by the earlier administration of amphetamines.[38]

So here we have exactly the evidence of exactly the kind of model that we suggested would be beneficial. Injection of amphetamine does not cause increased occurrent wanting on its own. Rather, it increases dispositional wanting. i.e. increases long run sensitivity to the cues; and it increases occurrent wanting when the cue is also present. This is what we suggested would be the function of the A-signal. Amphetamines are thus boosting the naturally occurring A-signal; and since what they are boosting is dopamine, we conclude dopamine is the A-signal.[39]

What of the B-signal? The evidence here is far from over-whelming, but we suggest that the most likely candidates are phasic glutamate corticolimbic signals that reach the nucleus accumbens. These signals come to the nucleus accumbens from the prefrontal cortex, and from

---

[38] (Wyvell and Berridge, 2001). See also (Tindell *et al.*, 2005); (Smith *et al.*, 2011).

[39] There is some very striking evidence that creatures can gain evidence that has an effect on subsequent consumption even when it does not like or need that thing. A rat who is exposed to highly concentrated saline solution, akin to swallowing a mouthful of the Dead Sea, that it neither wants nor likes, will, if subsequently deprived of salt, show both wanting and liking for the solution without needing to do any further exploration; see (Tindell *et al.*, 2009); (Berridge, 2012); (Robinson and Berridge, 2010). One of the current authors (RH) is tempted to understand this as the formation of a dispositional desire at the time when the rat is first exposed to the saline, a disposition that is only triggered by the later salt deprivation. The other (KB) thinks it far more plausible that the rat's initial exposure to the saline causes it merely to learn that the solution gives a salty sensation (as well as a nasty one), a case of sensory-based learning that gives rise to a dispositional desire only when the rat is salt deprived. Resolution of this would require finding out whether the initial exposure did elicit the A-signal. What is not controversial is that creatures can form dispositional desires for things that they enjoy even though they will not currently consume them because they are sated.

the basolateral amygdala, the thalamus and the hippocampus. It is well established that these signals interact with dopamine[40]; our suggestion is that the acceleration effect results from the dopamine magnifying the impact of the glutamate signals on the nucleus accumbens neurons. The glutamate signals are the primary cause of the occurrent cravings; but the level at which they do this is determined by the level of dopamine present.[41]

An accelerator effect along these lines also explains how foods can prime for their own consumption—the familiar cocktail party effect, where taking just one peanut can lead you to take many more. The dopamine release from the initial consumption in turn boosts the effectiveness of the glutamate signal and hence of the occurrent desire. This effect is equally familiar to those working with laboratory animals, where often a free taste of a reward is necessary to get the animals working again. Such effects can be very specific: human subjects who have already eaten a full sandwich lunch will be induced to consume further pizza by being primed with a small sample of pizza, or to consume further ice cream by being primed with a small sample of ice cream.[42] Importantly for our topic, such effects are generated by addictive drugs: a small dose of cocaine will increase future cocaine craving, a dose of alcohol will increase alcohol consumption.[43]

Whilst the desires generated by cues are specific in this way—a specific cue gives rise to a specific desire—there is some tendency for desires to generalize in the other direction—that is, there is a tendency to develop intrinsic dispositional desires for any of the cues that heralded a particular dopamine release. This is just what we would expect if the formation of the dispositional desires result from an associative mechanism as described above. Thus, for instance, pigeons will come to peck an illuminated piece of plastic that has heralded the delivery of food or of drink, and will do it, moreover, with the distinctive forms of pecks that correspond to what has previously been delivered, a eating-peck or a drinking-peck.[44] Likewise human cocaine addicts will 'chase ghosts', scrabbling for white specks that are only sugar grains or pebbles, and some smokers will prefer to puff on nicotine-free cigarettes rather than receive intravenous nicotine.[45]

There is much more evidence that we could draw on here. But rather than pursuing it, let us apply the model to an account of what it is that goes wrong in cases of addiction.

---

[40] See for instance (Kalivas *et al.*, 2009).

[41] We suggest that this is what is happening in the famous Schultz electrophysiology experiments (Schultz 1998). Unlike the Wyvell experiments, which involve elevated tonic (i.e. relatively enduring) dopamine levels, Schultz found that cues could elicit phasic (momentary) dopamine neuronal firing (and presumably firing-induced release), which has been widely interpreted as the phasic cause of the occurrent wanting. But here too the brain could be releasing phasic dopamine which amplifies the phasic glutamate signal triggered by the cue, thereby amplifying the cue-triggered 'wanting' engendered. Of course if both dopamine and glutamate signals are necessary for occurrent wanting, then it won't do to speak of the glutamate as itself the cause; but the point will remain that the dopamine is not sufficient on its own.

[42] (Cornell *et al.*, 1989).

[43] (Jaffe *et al.*, 1989); (de Wit and Chutuape, 1993).

[44] (Jenkins and Moore, 1973); (Allan and Zeigler, 1994).

[45] (Rosse *et al.*, 1993); (Rosse *et al.*, 1994); (Rose *et al.*, 2010).

VI. ADDICTION AS MALFUNCTION OF THE WANTING SYSTEM

Now that we have the model in place, our account of addiction can be quick. Let us assume then that when it is functioning properly dopamine works as the A-signal. What would happen if a subject consumed a substance that caused an artificial boost in that signal? The effect on the subject would be two-fold. First, it would likely experience a large boost in occurrent desire for the substance. Second, there would be a large boost in its dispositional desire for the substance. Given the associative nature of the system, that desire would be cued by the substance itself, or by other cues that were around at the time that the substance was consumed. If the dopamine signal was strong enough, the ongoing sensitization could be very great, potentially persisting indefinitely.

Our claim is that this is just what happens in cases of addiction. Since the addictive drugs artificially stimulate the dopamine system so powerfully they give rise to long lasting dispositional desires. The dispositional desires are triggered by cues surrounding the consumption of the drugs: the drugs themselves, but also, given the associative nature of the process, the places in which they are consumed, the paraphernalia surrounding their consumption, and so on. Since these are intrinsic and not instrumental desires, they are not undermined by the belief that consumption of the drugs will not be pleasurable, or that it will be harmful in some other way. These dispositional desires may persist long after the subject has stopped taking the drugs, and has gone through any associated withdrawal. A cue provided by seeing the drug, or the environment in which it was once taken, or even by imagining it, may provoke a powerful occurrent desire for it; and if this results in further consumption, the whole pattern will be repeated.

This seems to fit the facts very well. Or at least, it fits some of the facts very well, the pathological facts, those concerning the way that addiction differs from ordinary behaviour. But it might seem that this has taken us too far. For what are we to make of those aspects of addiction that make it seem very much like ordinary behaviour? Can we preserve the idea that addicts are nonetheless sensitive to standard incentives?

The crucial point here is that, in human beings, the incentive salience process that we have sketched does not necessarily lead directly to behaviour.[46] It typically leads instead to cravings: to powerful desires that tend to crowd out other considerations.[47] Many philosophers make a sharp contrast between desires and intentions. Desires are the inputs to deliberation; it is quite rational to have many that conflict. Intentions are the outputs of deliberation; they are insulated from reconsideration and lead directly to action, and so they need to be consistent. Cravings seem to come somewhere between the two. Whilst they have many of the features of standard desires, they are not easily thought of as inputs to deliberation. Rather, they lead directly to action unless something stops them. Stopping them requires self-control; to this we now turn.

---

[46] This is not to deny that incentive salience effects can work unconsciously in a way that takes them fairly directly to behavior. See (Winkielman and Berridge, 2004). But such behavior is still susceptible to self-control; it is just that the subject doesn't see the need to exert it.

[47] See (Loewenstein, 1999) for a good discussion of how cravings tend to narrow one's focus.

VII. SELF-CONTROL

Both philosophers and psychologists tend to view desires as a fundamentally uniform class. Roughly, they are the states that move an agent to action. In contrast we think that they are heterogeneous. So far we have focused just on one kind, the desires, or cravings, that result from the process of incentive salience. As we mentioned at the beginning, we also have other, more rationally tractable desires: a desire to take a holiday in St Petersburg, say, or to be healthy, or to treat a particular person well. And many of these are intimately connected with our beliefs. If we come to think that St Petersburg is too Western to be worthy of a visit, and that Moscow would be a better destination, then our desire to visit will be undermined. In contrast, the cravings that result from the incentive salience process are not typically undermined by the belief that they are harmful.[48]

But if we have at least two different sorts of desires—together perhaps with other factors that also influence our behavior, like our habits—then the question arises of what it is that will determines what we will do. This is a difficult and complex question that we cannot hope to fully answer here. But one thing that we think has become clear in recent years is that it is not fully determined by the relative strength of the different sorts of desires. We also need to factor in a more active control on the part of the agent.

A wealth of psychological research supports the idea that self-control should be taken seriously. Self-control develops in children after the development of desires; it is effortful; it is depleted by various factors including stress, fatigue, and its prior exercise; and it can be developed and deployed more or less successfully.[49] A failure to behave a certain way might indicate by a lack of desire to behave that way. Alternatively it might indicate that a desire, even the kind of craving that results from addiction, is being held in check by self-control.

To say that self-control is real is not to deny that its exercise is sensitive to the agent's beliefs and desires. Agents can be well motivated to employ it, if they think that there is something to be gained from it, and that its employment will be successful. Alternatively, if they think that it will bring little benefit, or that the benefits can be gained more easily another way, or that it is unlikely to succeed, they will be far less likely to employ it, and even if they do initially employ it, given that it is effortful, they will be far more likely to give up.

As we have seen, the pathology of addiction means that addicts will experience strong cue-driven cravings long after withdrawal is over, especially at particular moments such as when a drug cue is encountered in a moment of stress or emotional excitement. But this is not the end of the story. Whilst there is some evidence that addictive drugs can diminish self-control by damaging the prefrontal cortex,[50] there is no reason to think that addicts lose it altogether. Indeed, the fact that addicts can get themselves off their addictions is strong evidence that it is not. Controlling cravings may be tremendously hard work, but that it is not to say that it is impossible. Understanding when it is that addicts will continue to consume and when they will

---

[48] For an excellent discussion of such desires see (Railton, 2012). Many actual desires may combine an element of both types; indeed the very case that Railton uses as illustrative of the more cognitive desire —a desire for an espresso—is very plausibly a case in point.

[49] For general discussion of the evidence and of some of the mechanisms involved, see (Holton, 2009).

[50] See for instance, (Volkow *et al.*, 2004).

not thus requires an understanding of how their cravings interact with their self-control. Whilst we do not have even the beginnings of a real account here, we identify the following factors as very likely to be relevant to the pattern of activity that we remarked on at the outset, in particular the responsiveness of addicts to incentives, and their tendency to escape their addictions in their late twenties or early thirties.

(i) the strength of the self-control system
There is evidence that self-control, regulated primarily by the pre-frontal cortex, continues to develop in strength into the mid-twenties, typically maturing rather earlier in women than in men.[51]

(ii) the efficiency with which the self-control system is employed
A great deal of research indicates that there are techniques that enable agents to better deploy their self control. Forming prior intentions and then acting on them without reopening the question of what to do seems important. Similarly, mindfulness techniques can enable agents to stand back from their desires in ways that make their self-control more effective. It is still an open question how effective such techniques can be against the kinds of cravings engendered by addiction, but initial research indicates that they can make a difference.[52] Again, skill in using the self-control system is something that we might expect to increase with age.

(iii) the role of desires
Addicts who have strong motivations for giving up rather than continuing are more likely to employ their self-control to overcome cravings. And it does seem likely that the concerns about partners, families, and careers will become more pressing as people reach their late twenties and early thirties. Conversely, since dopamine levels start to fall from the teenage years onwards, the power of the cravings may themselves diminish.

(iv) the role of belief
If addicts think that there is little reason to give up today, since giving up tomorrow will be just as good, there will be little motivation to employ self-control. Vague concerns about health and well-being are often of that form; there can be a sense that whilst giving up is something that needs to be done at some point, one more dose won't hurt. In contrast, the incentives that have been shown to work well—for instance the knowledge that certain dismissal from a much valued job will follow a single positive drug test—guarantee a immediate cost or benefit. We suspect that much the same is true of a price rise; whilst it is true that paying the higher price just one more time is probably within the addict's reach, there is no escaping the fact that a higher price is being paid. The other set of relevant beliefs concern the efficacy of exerting self-control. If the addict is convinced that they will succumb despite their best efforts—if not today, then surely soon—the motivation to try will be much reduced. And here, presumably, the addict's own theory of addiction will have a part to play. If they think of the addiction as

---

[51] See, for instance, (Luna and Sweezy, 2004); (Goldstein *et al.*, 2009); and, for a popular review, (Sabbagh, 2006).

[52] (Prestwich *et al.*, 2006); (Kober *et al.*, 2010).

resulting in behaviour that is quite outside their control, they will be far less motivated to try to control it.[53]

VIII. THE EXTENT OF ADDICTION, AND ITS RATIONALITY

We have talked about addictions that are caused by drugs—by substances that interfere directly with the dopamine system, and gain their incentive salience effect from that interference. But what of the many other kinds of behavioral addictions—addictions to gambling, shopping, sex or the internet—that feature so prominently in current discussion. Can we give an account of them? Or is the theory we have given bound to say that they these are not really addictions?[54]

Clearly our account is bound to say that there is an important difference between substance and behavioral addictions. The latter do not, so far as we know, involve mechanisms that short-circuit the dopamine system in the way the former do. Nevertheless, there is good reason to think that they too work through the incentive salience system, and so that they too can result in cue driven cravings that are relatively insulated from other desires and from beliefs about what is good. Of course, if the dopamine system has not been short-circuited, then these behavioral addictions must have originated from behavior that was pleasurable, or was in some other way recognized by the agent's dopamine system as being beneficial. But the assessment of dopamine system might be at odds with the agent's more cognitive beliefs about the value of the activity; and even if it is not, once the intrinsic desires have been established, they will tend to persist through changes in the agent's assessment at any level. Even if the agent stops liking the thing concerned, a well-established incentive salience desire will degrade very slowly. The result can be behavior that looks very like the addiction engendered by drugs.[55]

This brings us finally to an issue that we have largely skirted up till now, that of the rationality of addicts. Most ordinary choice models see addicts as quite rational, though working with unusual desires or false beliefs (perhaps there is some irrationality in how they arrived at those beliefs, but that doesn't affect the rationality of how they act upon them). Most disease models see the addict as largely arational: addictive actions hardly count as intentional actions at all, and so fall outside the scope of rationality. In contrast, the account that we have developed here sees the addict as potentially irrational in two ways. One is familiar: if considered views about what would be best diverge from action, then both substance addicts

---

[53] A point that has been noted many times by Albert Bandura; see for instance (Bandura, 1999).

[54] We have made the traditional division between substance addiction and behavioral addiction, but it could be that some substances give rise to addiction-like behavior without hijacking the dopamine system in the way we have discussed, and so should be grouped with the behavioral addictions. Sugar might be like that, and perhaps, though here the findings are controversial, cannabis. So a more careful distinction would be between the dopamine-hijacking addictions, and those that are not. But we will stick with the more traditional terminology.

[55] Further evidence that drug and behavioral addictions have much in common comes from the cases of Parkinson's patients who respond to their dopamine supplement by developing addictive behavior. See (O'Sullivan *et al.*, 2009). We leave open the question of whether other behaviors that look rather like chemical addictions—those resulting from obsessive compulsive disorder, for instance—should also be understood in the same way.

and behavioral addicts will frequently be akratic, in ways that have at least a prima facie claim to irrationality. The second is rather less familiar. If what we have said is right, then something goes badly wrong with the process by which substance addicts (but not behavioral addicts) form their desires: substances come to be desired independently of any pleasure or other benefits that they bring. There has been much discussion in philosophy of whether intrinsic desires can be irrational. What we are suggesting is that substance addiction results from the malfunctioning of a normally rational system for creating intrinsic desires. This seems to us as clear a case of an irrational intrinsic desire as one is ever likely to find.

IX. CONCLUSION

We started by stressing the need to find a middle path. Our attempt to find one has involved exploring the interaction between two different systems: one that regulates our desires, and one that controls which desires we act on. Addiction results from the malfunction of the first; insofar as it does not result in a complete loss of agency, that is thanks to the second. In a sense then, both the disease model and the choice model are describing something real; but each gives a picture that is partial. We hope that we have gone some way to putting them together.

BIBLIOGRAPHY

Allan, R.W. and H.P. Zeigler 1994: 'Autoshaping the pigeon's gape response: acquisition and topography as a function of reinforcer type and magnitude' *Journal of the Experimental Analysis of Behavior*, 62, 201–223.

Bandura, Albert 1999: 'A Sociocognitive Analysis of Substance Abuse', *Psychological Science* 10, 214–7.

Berridge, Kent 2000: 'Taste reactivity: Measuring hedonic impact in infants and animals' *Neuroscience and Biobehavioral Reviews* 24 173–98.

— 2007: 'The debate over dopamine's role in reward: the case for incentive salience', *Psychopharmacology*, 191, 391–431.

— 2012: 'From prediction error to incentive salience: mesolimbic computation of reward motivation', *European Journal of Neuroscience*, 35, 1124–43.

— and Morten Kringelbach 2008: 'Affective Neuroscience of Pleasure: Reward in Humans and Animals', *Psychopharmacology* 199, 457–80.

— and E. Valenstein 1991: 'What psychological process mediates feeding evoked by electrical stimulation of the lateral hypothalamus?' *Behavior Neuroscience* 105, 3–14.

Cornell, C.E., J. Rodin, and H. Weingarten 1989: 'Stimulus-induced eating when satiated' *Physiology and Behavior*, 45, 695–704.

Dickinson A. and B. Balleine 2010: 'Hedonics: The Cognitive-Motivational Interface' in (Kringelbach and Berridge 2010) pp. 74–84.

Elster, Jon and Ole-Jørgen Skog (eds.) 1999: *Getting Hooked* (Cambridge: Cambridge University Press).

Everitt, B., D. Belin, D. Economidou, Y. Pelloux, J. Dalley, and T. Robbins 2008: 'Neural mechanisms underlying the vulnerability to develop compulsive drug-seeking habits and addiction', *Philosophical Transactions of the Royal Society of London B Biological Science* 363, 3125–35.

Faure, A., J. Richard, and K. Berridge 2010: 'Desire and dread from the nucleus accumbens: cortical glutamate and subcortical GABA differentially generate motivation and hedonic impact in the rat', *PLoS One*, 5, e11223.

Flagel, S., J. Clark, T. Robinson, L. Mayo, A. Czuj, I. Willuhn, C. Akers, S. Clinton, P. Phillips, and H. Akil 2011: 'A selective role for dopamine in stimulus-reward learning', *Nature*, 469, 53–7.

Goldstein, R. A. Craig, A. Bechara, H. Garavan, A. Childress, M. Paulus, and N. Volkow 2009: 'The neurocircuitry of impaired insight in drug addiction', *Trends in Cognitive Science*, 13, 372–80.

Herrnstein, Richard and Drazen Prelec 1992: 'A theory of addiction,' in (Loewenstein and Elster 1992).

Heyman, Gene, 2009: *Addiction: A Disorder of Choice* (Cambridge MA: Harvard University Press).

Hoffman, J. and S. Froeke (eds.) 2007: *Addiction: Why can't they just stop?* (Emmaus, PA: Rodale Press).

Holton, Richard 2009: *Willing, Wanting, Waiting* (Oxford: Clarendon Press).

Jaffe, J., N. Cascella, K. Kumor, and M. Sherer 1989: 'Cocaine-induced cocaine craving' *Psychopharmacology* 97, 59–64.

James, William, 1890: *Principles of Psychology* (New York: Henry Holt).

Jenkins, H. and B. Moore(1973): 'The form of the auto-shaped response with food or water reinforcers' *Journal of the Experimental Analysis of Behavior* 20, 163–81.

Kalivas, P.W., R.T. Lalumiere, L. Knackstedt, and H. Shen 2009: 'Glutamate transmission in addiction', *Neuropharmacology*, 56 Supplement 1, 169–73.

Kober, Hedy, Ethan Kross, Walter Mischel, Carl Hart, and Kevin Ochsner 2010: 'Regulation of craving by cognitive strategies in cigarette smokers', *Drug and Alcohol Dependence*, 106, 52–5.

Kringelbach, Morten and Kent Berridge (eds.) 2010: *Pleasures of the Brain* (Oxford: Oxford University Press).

Lawrence, A.D., A.H. Evans, and A.J. Lees, 2003: 'Compulsive use of dopamine replacement therapy in Parkinson's disease: reward systems gone awry?' *Lancet Neurology*, 2, 595–604.

Leyton, M. 2010: 'The neurobiology of desire: dopamine and the regulation of mood and motivational states in humans', in (Kringelbach and Berridge 2010) pp. 222–43.

Linden, David, 2011: *The Compass of Pleasure*, (New York: Viking).

Loewenstein, George 1999: 'A Visceral Account of Addiction,' in (Elster and Skog 1999) pp. 235–64.

— and Jon Elster (eds.), 1992: *Choice Over Time* (New York: Russell Sage Press).

Luna B. and J. Sweezy 2004: 'The Emergence of Collaborative Brain Function: fMRI Studies of the Development of Response Inhibition', *Annals of the New York Academy of Science* 1021, 296–309.

McClure, S.M., Daw, N.D. & Read Montague, P. 2003: 'A computational substrate for incentive salience.' *Trends in Neuroscience*, 26, 423–8.

Moore, G. E. 1905: *Principia Ethica* (Cambridge: Cambridge University Press).

O'Sullivan, S, A. Evans, and A. Lees 2009: 'Dopamine Dysregulation Syndrome: An Overview of its Epidemiology, Mechanisms and Management' *CNS Drugs* 23, 157–70.

Peciña, S., B. Cagniard, K. Berridge, J. Aldridge, and X. Zhuang, 2003: 'Hyperdopaminergic mutant mice have higher "wanting" but not "liking" for sweet rewards' *Journal of Neuroscience*, 23, 9395–9402.

Prestwich, Andrew, Mark Conner and Rebecca Lawton 2006: 'Implementation Intentions: Can They Be Used to Prevent and Treat Addiction?' in (Wiers and Stacy 2006).

Railton, Peter 2012: 'That Obscure Object, Desire' *Proceedings of the American Philosophical Association*.

Redish, A. D. 2004: 'Addiction as a computational process gone awry', *Science*, 306 1944–7.

Robinson Mike and Kent Berridge 2010: 'Instant incentive salience: Dynamic transformation of an aversive salt cue into a "wanted" motivational magnet', Abstract, *Society for Neuroscience*, San Diego.

Robinson, Terry, and Kent Berridge 2003: 'Addiction' *Annual Review of Psychology* 54 25–53.

Rose, J., A. Salley, F. Behm, J. Bates, and E. Westman 2010: 'Reinforcing effects of nicotine and non-nicotine components of cigarette smoke', *Psychopharmacology*, 210, 1–12.

Rosse, R., M. Fay-McCarthy, J. Collins,T. Alim, and S. Deutsch 1994: 'The relationship between cocaine-induced paranoia and compulsive foraging: a preliminary report', *Addiction*, 89, 1097–1104.

— M. Fay-McCarthy, J. Collins, D. Risher-Flowers, T. Alim, and S. Deutsch 1993: ' Transient compulsive foraging behavior associated with crack cocaine use', *American Journal of Psychiatry*, 150, 155–6.

Sabbagh, L. 2006: 'The Teen Brain, Hard at Work', *Scientific American Mind*, August/September, 20–25.

Saunders B., and T. Robinson 2012: 'The role of dopamine in the accumbens core in the expression of Pavlovian-conditioned responses' *European Journal of Neuroscience*, online first.

Schultz, Wolfgang 1998: 'Predictive reward signal of dopamine neurons' *Journal of Neurophysiology* 80, 1–27.

— 2002: 'Getting formal with dopamine and reward', *Neuron*, 36, 241–63.

— Peter Dayan, and Reid Montague, 1997: 'A neural substrate of prediction and reward', *Science*, 275, 1593–9.

Skog, Ole-Jørgen 1999: 'Rationality, Irrationality and Addiction—Notes on Becker's and Murphy's Theory of Addiction' in (Elster and Skog 1999) pp. 173–207.

Smith, Kyle, Kent Berridge and J. Wayne Aldridge, 2011: 'Disentangling pleasure from incentive salience and learning signals in brain reward circuitry' *Proceedings of the National Academy of Science*, 108 E255–E264.

Sournia, Jean Charles, 1990: *A History of Alcoholism* (Oxford: Blackwell).

Sutton, Richard and Andrew Barto, 1998: *Reinforcement Learning* (Cambridge MA: MIT Press).

Tindell, Amy, Kyle Smith, Kent Berridge and J. Wayne Aldridge, 2009: 'Dynamic computation of incentive salience: "wanting" what was never "liked"', *Journal of Neuroscience* 29(39) 12220–8.

— K. Berridge, J. Zhang, S. Peciña and J. Aldridge 2005: 'Ventral pallidal neurons code incentive motivation: amplification by mesolimbic sensitization and amphetamine' *European Journal of Neuroscience*, 22, 2617–34.

Volkow, Nora, Joanna Fowler, and Gene-Jack Wang 2004: 'The addicted human brain viewed in the light of imaging studies' *Neuropharmacology* 47, 3–13.

Wiers R. W. and A.W. Stacy (eds.) 2006: *Handbook of implicit cognition and addiction* (Thousand Oaks, CA: Sage).

Wilde, Oscar 1891: *The Picture of Dorian Gray* (London: Ward Locke and Co.).

Winkielman P. and K. Berridge 2004: 'Unconscious emotion', *Current Directions in Psychology*, 13, 120–3.

Wise, Roy, 1980: 'The dopamine synapse and the notion of 'pleasure centers' in the brain', *Trends in Neuroscience*, 3, 91–5.

— 1985: 'The anhedonia hypothesis: Mark III' *Behaviour and Brain Science*, 8, 178–86.

— 2004: 'Dopamine, learning and motivation', *National Review of Neuroscience*, 5, 483–94.

de Wit, H. and M. Chutuape 1993: 'Increased ethanol choice in social drinkers following ethanol preload' *Behavioral. Pharmacology*, 4, 29–36.

Wood, Wendy and David Neal, 2007: 'A new look at habits and the habit-goal interface'. *Psychological Review, 114*, 843–63.

Wyvell, Cindy and Kent Berridge 2000: 'Intra-accumbens amphetamine increases the conditioned incentive salience of sucrose reward' *Journal of Neuroscience*, 20, 8122–30.

— 2001: 'Incentive Sensitization by Previous Amphetamine Exposure: Increased Cue-Triggered "Wanting" for Sucrose Reward', *The Journal of Neuroscience*, 21, 7831–40