ADDICTION, SELF-SIGNALLING, AND THE DEEP SELF[1]

RICHARD HOLTON

1. INTRODUCTION

It is easy to see addiction as destroying intentional action altogether; easy to see addicts as swept away by a welter of compulsive behaviors. But the empirical evidence has increasingly indicated that such a picture is not right.[2] While addiction does involve plenty of involuntary states—most centrally, cue-driven cravings—these do not in themselves amount to behaviour. Behaviour comes from the interaction of these involuntary states with more familiar deliberative processes. So in trying to understand why addicts act as they do, we need to understand these processes, and the beliefs, desires, intentions, and other mental states that are involved.

In this context then, it will matter whether addicts think that it is worth giving up, and whether they think that they are able to do so. What information will they have? They will doubtless have been told many things. But they will also be observers of their own behaviour, hypothesizing about their situation and capacities on the basis of what they have done. Here they need not just be passive self-observers. They may be acting partly in order to gain relevant information; that is, they may be self-signalling.

---

[1] Versions of this paper were presented at the *Addiction, Self and Self-Knowledge* workshop run by *Mind and Language* in London, March 2015; and at the *Self-deception, Self-signaling, and Self-control* workshop at the IAST Toulouse, June 2015. Many thanks to the audiences there, and to Rae Langton, Adam Bates and the referees for this journal, for comments and discussion. Special thanks to Drazen Prelec for first making me think about self-signalling.

[2] The evidence for this is now very strong and can be found in many writings. Kent Berridge and I reviewed some of it in an earlier piece (Holton and Berridge 2013), but for a fuller and more recent summary, see (Pickard, 2016).

Why is self-signalling important in addition? Addicts who are trying to give up typically have conflicting motivations: at the same time as wanting to stop, they still want the drugs. Resisting the desire for drugs is hard work. That means that it can be easily undermined by any consideration that suggests that it not worthwhile, even if the rational credentials of such a consideration are slim. In many cases of temptation a failure to resist is underpinned by a judgment shift: agents come to think that the tempting object is better, relative to the alternatives, than they had thought when they resolved to resist. But simple judgement shift is hard to maintain in cases of serious addiction. It is hard for addicts to convince themselves that it is genuinely better to persist in their addiction. Self-signalling, I suggest, provides them with an alternative story. It is a wonderfully flexible device. If giving up seems to be easy, self-signalling can be used to provide evidence that they are not addicted and hence do not need to give up. If giving up seems to be hard, it can provide evidence that they are so badly addicted that giving up is bound to fail. So it is not, I will argue, a very helpful method to employ.

In the next section I will present the central features of self-signalling, and related phenomena, stressing the distinction between those cases where it provides good evidence about the state of the agent, and those where it is deceptive. In the third I provide a background account of how I am understanding addiction. In the fourth section I bring these two together to discuss the role of self-signalling in addiction. In the fifth I ask how the corrupting role of self-signalling might be escaped.

2. SELF-SIGNALLING

This seems like a rather general truth: it is only worth trying to do something if you believe that you will succeed. Of course there are exceptions. It may be that you simply need to show, whether to yourself or to others, that you have tried. Or it may be that all the other options are

so terrible that trying for the remaining one makes perfect sense, even though the odds of success are slim; trying is the best of a bad bunch. For the most part though, coming to think that you're going to fail in a task is likely to greatly reduce your motivation to try, an idea that has been much developed by psychologists under the banner of 'self-efficacy'.[3]

Here though is a complicating factor. Sometimes the information you receive from trying will give evidence of how likely you are to succeed. This means that there can be reason to try in the absence of belief in success; in fact, there may be reason to try even if you believe you will fail. For even if you believe you will fail, you may still be looking for evidence that you will not; and trying may be a good way of obtaining such evidence.

This may appear exotic, but it is part of a very familiar phenomenon that extends beyond just a concern with my likely success. Should I go skiing? That rather depends on whether I will enjoy it. But I don't know whether I will. I can imagine a wonderful week revelling in the exhilaration of flying down crisp mountain sides. I can also imagine a week of abject terror. 'Try it!' urge my friends; 'Only then will you know if you like it!' If I do it will be my desire for knowledge of my preferences—whether I like skiing or not—that will move me. If we think, moreover, that I seek such knowledge because I aim to satisfy my preferences, then this too is a case of finding out what I can do—how I can have a good time—by trying.

In the skiing case, the reason to try is increased by the reflection that the knowledge will be useful in the long run. If I enjoy the holiday, then, money permitting, a lifetime of winter breaks in the mountains beckons. If I don't, I write it off as useful experience, and know to avoid it in the future. If we only have one shot at something, then this is not relevant. But even in the case of the single shot, it is unusual that success will be achieved by a single act.

[3] The founding work here is by Albert Bandura. For a good overview see (Bandura 1992).

Achieving most ends require some persistence over time; and here the evidence that comes from starting the attempt may well affect whether one persists.

In some cases we can imagine that the resulting evidence will bolster the attempt. If I find that I manage to get up early every morning for a week to go for a run, that will increase my confidence that I am able to keep the schedule, and that it turn will increase my motivation to do so. Indeed it may be that simply finding I have the motivation to try is enough to convince me that I have the motivation to persevere. And if I believe I am going to persevere, that makes each individual run better motivated, since it will gain its point from being part of a virtuous pattern.

In other cases though, the experiences will be less conducive to success. Finding how hard the attempt is can be seen as evidence that failure is likely. Having failed to run most days this week, Friday's attempt will be poorly motivated. Even if I were to manage it, it might seem pointless, since my confidence that it would be part of a worthwhile pattern would be so low.

A third set of cases involve vacillation. These have had plenty of discussion in the philosophical literature, though the examples tend to the fantastical. Here is one from Andy Egan.[4] Imagine a 'Psychopath Button': pushing it will kill all psychopaths. Imagine too someone who is contemplating pushing it: 'A world without psychopaths', they reason, 'would be a better place'. But as they approach the button they hesitate. The thought occurs to them that anyone who would do such a thing would be a psychopath. Unwilling to risk self destruction to achieve a world without psychopaths, they back off. But this in turn provides them with evidence that they are not a psychopath. So then they are caught in a cycle: as they decide to push the button, pushing it seems like a bad idea; as they decide not to, the likelihood that they are a psychopath recedes, and it seems like a good idea again. (To rephrase this in terms

[4] (Egan, 2007); I have made my subject a little less certain than Egan's.

of an attempt to gain evidence of what they can achieve, make their goal a world that includes them and no (other) psychopaths; the question is whether they can achieve this by pushing the button.)

Following recent discussions in economics and psychology, think of all these cases as involving the possibility of *self-signalling*. Sometimes that term is used just to cover epistemically perverse behaviour: cases in which agents want to convince themselves of something, whether or not it is true.[5] I suggest that we should initially cast our gaze more broadly, and then we can turn to the perverse cases as a particular application. So I define active self-signalling behaviour as the behaviour of agents that is motivated, at least partially, by an aim to obtain evidence about some relatively hidden feature of themselves—the *target feature* as I shall call it. The notion of self thus plays a triple role in self-signalling: activity by the self is undertaken to provide evidence about the self to the self. (Seen this way we immediately recognize a family of related behaviours that we can perform: those that provide evidence about ourselves to others; those that provide evidence about others to ourselves; those that provide evidence about others to others.) Not all of the behaviours we have looked at are obviously cases of active self-signalling in this sense: in the case of the psychopath button, at least initially, our imaginary agent was simply motivated by a desire to eliminate psychopaths. Evidence of their own standing came as a bonus along the way. Think of this as *passive* self-signalling. Active and passive can easily merge into each other. However, once our agent realizes that evidence is available through their behaviour, then further behaviour can be motivated by a desire to obtain such evidence. Wanting to know whether they are a psychopath, they look to see how tempted they are to push the button.

[5] Bodner and Prelec (2002) define it as 'an action chosen partly to secure good news about one's traits or abilities, even when the action has no causal impact on these traits and abilities.'

What then of the perverse cases that have been the focus of much of the discussion? If self-signalling is going to be accurate, the signal (the behaviour) must be a reliable guide to the target feature. This may be because the signal constitutes the target feature; because the target feature causes the signal; because the signal causes the target feature; because both have some common cause; or perhaps for some other reason. But if the signal is to be a reliable indicator, it must not be the case that it would have been equally likely even in the absence of the target feature.

We get perverse cases when the self-signalling involves a signal that is known not to be a reliable indicator. Since Weber we have been familiar with the idea that Calvinists, who thought that their fate was sealed by God's prior decree, might nevertheless be motivated to do good works by a desire for assurance that they were among the elect. This certainly wasn't an unwarranted imputation on Weber's part: such ideas were much discussed by 17th century puritans.[6] Nevertheless, this behaviour looks as though it may be perverse. The performance of good works is under an agent's control. The elect can perform them; but so can the reprobate. That wouldn't devalue good works as a signal if the elect were more likely to perform them than the reprobate. But if the primary motivation to perform good works is to gain the reassuring belief that one is among the elect, and both groups are equally keen on that, then performing good works will not be a reliable signal. Strikingly then, it is exactly when assurance of election is proposed as a *motivation* for performing good works that the performance of good works loses its efficacy as assurance.[7]

[6] See Bozeman for an excellent discussion of the concerns about self-knowledge and its connection to behaviour, that permeated much Puritan thought.

[7] Oddly, Weber doesn't seem to be concerned about this tension, perhaps because he takes it that the kind of worldly success that bring evidence of election requires cooperation between Man and God—'God helps those who help themselves' p. 69—and so could not achieved by the reprobate. At other times though he says that fatalism is 'the only logical consequence of predestination', whereas 'the psychological result was precisely the opposite' p. 192, n66. This suggests that the 'psychological results' were irrational, though he never says as much.

In the Calvinist case, we might think that those concerned were culpable for their failure to realize that the signals they were getting were unreliable. To this extent, their actions were irrational. But in other cases things are less clear. If agents don't know that they are being motivated by the desire for a good signal, it doesn't seem right to accuse them of irrationality. Consider a much discussed experiment on self-deception from Quattrone and Tversky (1985). One group of subjects are told that being able to hold their arms for longer in cold water after exercise is a sign of a strong heart and a long life; a second group are told that it is a sign of a weak heart and a short life. The former tend to keep their arms in the water for longer. Clearly this is self-signalling: each group wants good news, and each modifies their behaviour accordingly. If they knew what they were doing, they would certainly be irrational. But until they see the data they may have no way of knowing that: three quarters claimed that they were not trying to influence the outcome. So here it seems more like simple mistake: the subjects mistakenly think that they are not influencing the outcome when they are.

Further cases of apparently perverse self-signalling need not involve mistake; indeed they may be quite rational. The awareness that the attempt may give evidence about the likelihood of success can itself be a reason to avoid the attempt, or to do something to interfere with the attempt so that it provides no evidence: if you fear that the evidence will be bad, you may not want it. This seems to be part of the motivation involved in cases of self-handicapping.[8] A student might deliberately fail to revise for an examination; worse, they might get horribly drunk the night before, staggering into the exam hall late and hung-over. The behaviour looks irrational yet there might be purpose. They might be trying to avoid evidence that they would still do badly even if they worked as hard as they could. If there is any mistake here it is indirect. They may be protecting a mistaken belief about how good they could be if they worked; or they may be mistaken in their assessment of why they failed to revise or got drunk. I suspect that it is these features that make the behaviour seem perverse. However, neither of

[8] Here I follow Bodner and Prelec's (2002) discussion.

these is essential to the case. If there is full realization of what is going on, then, while their preference for ignorance in a case like this may be unusual, it seems perfectly rational.

What is essential to the cases that do involve mistake? The crucial thing is that the signal and the target feature need to be distinct. The state of being elect is quite distinct from that of performing good works; that is how good works can fail to provide evidence for election. Conversely, if the signal actually constitutes the target feature mistake is not possible. If someone's consistently brave actions are motivated, in part or entirely, by a desire for assurance that they are brave, then that doesn't seem to disqualify them as brave, since there is nothing more to bravery than being brave. We might unfavourably contrast such self-conscious bravery with a purer kind; but it is still bravery. Other cases are more controversial. Consider apparently moral behaviour that is motivated by the desire to know that one is moral. Is that moral? Kant would have said no, since the motivation is heteronomous. But it is far from obvious that he is right. A desire to look good in one's own eyes can be a powerful motivator; we might hesitate before saying that it is bound to be self-defeating.[9]

The state of being elect is radically distinct from the performance of good deeds. But there is a lesser distinction that might be drawn between signal and target that is still enough for mistake. Even if the signal constitutes the current target behaviour, it might be being used as evidence that the behaviour will continue. Brave behaviour now may be taken as evidence of brave behaviour in the future. So once again the possibility of error opens up. Brave behaviour now might be no indicator of brave behaviour in the future, especially if the motivation to prove that one is brave is unlikely to endure.

[9] For an account that understands moral motivation as largely derived from self-signalling, see (Bénabou and Tirole 2012).

3. ADDICTION

We turn now to addiction. The view that I will be assuming, based on the work of Berridge and Robinson, takes addiction to involve a disruption of the wanting system. The background idea, at least as I understand it, is this.[10] Human beings have evolved to have a system that enables us to form intrinsic desires for things that we have liked, or otherwise benefitted from, in the past. So, having sampled a certain foodstuff, and having found it tasty, or perhaps nutritious in some other way, we now have an intrinsic desire for it, triggered by the sight or smell of it, or some other cue that we have associated with consuming it in the past. Crucially these intrinsic desires work independently of our beliefs and other desires. In addition to the intrinsic desire for the food, we might have a belief that eating it will give us pleasure, and a desire for pleasure, and hence an instrumental desire to eat it. But we do not need to have this in order to eat it; the intrinsic desire, triggered by the cue, will do that on its own.

These acquired intrinsic desires are controlled by the mesolimbic dopamine system. Addictive drugs act directly on this system, by boosting the release of dopamine, inhibiting its re-uptake, or something similar. The effect is that consumption of the drug, at least in individuals who are susceptible, will give rise to an intrinsic desire for it, independently of any pleasure that it might bring; this will tend to give rise to increased consumption, which will further strengthen the desire, and a vicious cycle will have been started.

If that were all there were to the story then things would look grim for addicts. Moved by their intrinsic desires, they would tend to go on consuming unless they succeeded in generating an even stronger desire to resist. Generating such a desire would be unlikely. It is true that when the addictive desires are not stimulated by the relevant cue, an addict might have a stronger desire to give up. But once they are in the presence of the cue, the addictive desire, whose

[10] For a much fuller account, along with references to the original research, see (Holton and Berridge 2013).

strength will have been boosted by each previous act of consumption, will attain the status of a craving, crowding out rival desires.

Luckily though this is not all there to the story. We have also evolved a capacity for self-control. Quite how this works—how much it is like a muscle—is currently a matter of some controversy. All I need here is the idea that self control can work to rein in the kinds of intrinsic desires that are generated in addiction.

There is a great deal of empirical support for this claim. The use of such self-control is hard work though. Once an addictive desire has got going—the best thing is to avoid them in the first place by avoiding the cues, or by generating habitual patterns of activity that serve to block them—then resisting it by the force of self-control is physiologically arousing and aversive. It requires effort and focus; it is easily by undermined by distraction, stress and fatigue. So if an agent is going to deploy their self-control, they need to be very motivated to do so: they need to think that the rewards of success will be great; and that the chance of success will be substantial.

It seems that much of the effort that is needed here involves a certain kind of mental control. Studies on delayed gratification—where an agent is offered something fairly good straightaway, but something much better if they wait—suggest that subjects do best if they succeed in keeping their minds from reopening the question. Children in Walter Mischel's famous marshmallow experiments—where they are offered one marshmallow now, or two if they wait; or, better for our purposes, less preferred chewing gum now, or marshmallow if they wait—do better if they can see neither the reward for succumbing nor the reward for waiting. Seeing either, and hence being nudged again into opening up the question of whether waiting is worthwhile, makes them more likely to succumb. Likewise, dieters who use resolutions to successfully resist eating, find that the effectiveness of these resolutions is completely

undermined if they reflect on why they are resisting at the moment of temptation, and hence open the question of whether they need to resist in this particular case (Wieber et al. 2014). Other research has shed some light on one of the mechanisms at work here. It transpires that, as the period of waiting goes on, so the children's evaluation of the thing that they are waiting for goes down; so by the time they succumb they think it barely worth waiting for. Call this a *judgment shift*.[11]

There are two ways of thinking about agents' judgment shifts. The first is as a response to a prediction that they will succumb. Here it works as a cognitive dissonance mechanism. Anticipating failure, agents tell themselves a story that makes their behaviour look better. The obvious thing to say is that the gap between the more preferred option and the less preferred is not so great after all. The second way of thinking about judgment shift gives it a more active role in the behaviour. It is not that desire directly drives the agent to succumb, and then belief falls into place to accommodate it. It is rather than desire works through the belief: it changes the belief, and it is because of that that the agent succumbs.

It would not be an easy task to tell these two apart. In fact they may well both be present in many cases. We know in general that beliefs that arise from a cognitive dissonance mechanism typically go on to have behavioural effects: aware of their own behaviour, agents tell stories to make themselves look better, and then go on to act in accord with those stories. But the same can presumably happen before any action takes place: thinking that they might succumb, the agent preemptively increases the value placed on succumbing by a little bit; this in turn increases their estimate that they will succumb, which causes the value to rise a little further, and so on, until succumb they do.

[11] For more discussion, including references to the original studies, see (Holton 2009) Ch. 5.

Could it be that addicted agents are vulnerable to judgment shift of this kind? It is possible, but it is unlikely that it happens in the most straightforward way. The empirical evidence suggests that judgment shift only happens when the gap between the two goods is relatively small. It is easy to see why. It is easy enough for children to convince themselves that a marshmallow is not really any better than a piece of chewing gum. But suppose that the comparison were between a piece of chewing gum and a whole box of chocolates; or a piece of chewing gum and a bicycle. That would be so implausible that the method of judgment shift would not be available.

I suspect that the same is true of addicts. It is possible that some addicts, tempted by their drug, would conclude that, contrary to their earlier judgments, a life of continued addiction, with all of the damage to self and loved ones that goes with it, is actually better than a life off drugs. Denial of various forms is widespread in addiction—witness Hanna Pickard's discussion—and this may seem to be part of that package. But I doubt that it is very common. It is certainly not what one typically hears from the addicts themselves. Much more plausible is a more subtle strategy. Giving up is indeed better than addiction, *but*. And then the *but* gets filled in various different ways: *but* they are not in fact addicted, so can go on consuming in a controlled way that will not be damaging; *but* they do not need to give up now, since they can give up tomorrow; *but* the addiction is so powerful that there is nothing they can do about it. These strategies will be much more plausible if some evidence can be given for them. It is here that self-signalling comes in.

4. ADDICTION AND SELF-SIGNALLING

We saw earlier that self-signalling involves a signal and a target feature; and that the possibility of mistake, and of the perverse use of signalling, can open up when these two are distinct. So let's start by looking at the states that play these roles in addiction.

On the signal side there are two obvious candidates. First, addicts look to their behaviour: do they manage to resist on certain occasions, or do they succumb? Second, they look to how they feel during this behaviour: is resistance hard work, or is it easy?

Which target features are being looked for? Could the target be entirely constituted by the signal itself? Maybe sometimes that will be the case. A smoker might want to renounce one particular cigarette just to show that they can renounce that very cigarette. But that is unlikely to be terribly interesting in itself. More interesting is the idea that renouncing one cigarette shows something about their future behaviour—that they could renounce any future cigarette —or, more interesting still, it shows that they are not addicted to cigarettes.

This last idea requires a conception of addiction which holds it to be a feature which explains, but is distinct from, the addictive behaviour: a *deep-self* feature, as we shall say. At one extreme, Alcoholics Anonymous teaches that alcoholics will remain alcoholics, even if they cease drinking. A deep-self view of addiction needn't be as strong as that. It may simply be the view that if you are an addict, something has changed in you which makes it hard to give up—harder than if you liked the drugs to the same degree but were not an addict—and that this feature explains your behaviour. Such a view is entailed by the Berridge and Robinson account of addiction that was discussed in the last section. Indeed, this view may be combined with the thought that the addiction is in fact in tension with some other aspects of the deep self. Much anti-smoking therapy exploits the idea that those who smoke may not really be, according to their own self image, smokers. Subjects can simultaneously think that some deep feature—addiction—is leading them to act against other deep features of their natures.[12]

[12] I am thus rather liberal here on what will count as features of the deep self: I certainly don't require that they involve only what the agent judges best, or wants to want, or cares most about. It may be that we need a rather more restricted notion along these latter lines to do other work. For discussion see (Sripada unpublished).

Indeed, some deep-self view is almost universal across different accounts of the nature of addiction; the only accounts which may reject it are very pure rational choice accounts, which see addiction as driven solely by desire, and then understand desire in terms of revealed preference. On such an account, it is hard to see that there is anything about which the behaviour could be misleading. But on pretty much any other view, the possibility of a perverse use of signalling opens up. How might this happen?

Let us imagine an agent who is starting to be concerned about their drug use. They might simply be concerned that the drugs are damaging them to an extent that outweighs any benefit that they bring; this may be enough to motivate them to give up. But their concern might be more complicated than that. They might be concerned, first and foremost, that they are addicted. Why should this matter? It could be intrinsically worrisome. Addiction, the agent might think, brings a loss of autonomy and self-control, which would be bad in itself. Or it could be instrumentally worrisome. Addiction, they might think, entails that if they ever do need to reduce their drug usage, they will be unable to do so. So even if they are not concerned about their current drug use, they might be motivated to stop using the drugs in order to break the addiction.

So let us suppose then that our agent becomes motivated to try to give up. An initial question is: When? Part of the problem of addiction is that, while any benefits are immediate, the costs typically accumulate gradually. There is rarely a pressing need to give up now; tomorrow is soon enough. This, I suspect, is one of the biggest obstacles for successful quitting; as evidence of that, note how much better people do when given an artificial incentive to give up immediately—a frequent urine test on which one's job, or even just a small payment, depends.[13] In the absence of this, a serious attempt to give up will often involve something that

[13] For the first of these see (Heyman 2009, p. 86–7); for evidence of effective contingency management policies that pay former addicts with a small monetary reward that increases each time they give a clean sample, but returns to the baseline whenever the sample shows drug use, see (Schumacher *et al.* 2007) and (Petry *et al.* 2011).

marks a certain time as special: a sudden jolt of realization of the state into which things have drifted; a jarring comment from someone whose opinion matters; or perhaps a self-imposed deadline for the implementation of a resolution, such New Year, or a birthday.

Suppose then that our agent is motivated to give up now. But suppose that they are also sensitive to the evidence that the attempt to give up offers; indeed, they may well be actively self-signalling, their actions partly motivated by a desire to gain evidence about their addiction. Giving up is not a single event. It requires multiple acts of abstaining (and perhaps some instances of backsliding) spread over a considerable time. There is plenty of opportunity to get evidence of the nature of the deep self. What could the outcome be? In the light of the earlier discussion we can identify various possibilities.

The first possibility is that the agent finds the attempt remarkably hard. Perhaps they have already failed several times to avoid taking the drugs, and even when they were successful the process was horrible. On the basis of this information they conclude that long-term abstinence is unobtainable. The best they will manage is a temporary miserable halt; and if they are going to start consuming again, that will be pointless.

The second possibility is that the agent finds the attempt to give up remarkably successful. They manage to avoid taking the drugs, and they find this easier than they envisaged. We might expect that here at least the information given by their behaviour and their feelings would bolster the attempt, bringing confirmation that they have the capacity to take it through to completion. That is what we might expect given similar feedback at the start of an exercise programme; why should there be any difference here?

The difference, of course, is that in most cases our addict still has a powerful desire for the drug whenever it is cued. The attempt to give up may have been more successful than envisaged, but

it is still aversive and demanding. So they will be looking for something that gives them a reason to abandon it. And there are two obvious possibilities. The first is to conclude that the very success of the initial attempt showed that they are not really addicted. Perhaps they have been taking slightly more than was wise, but this is not a worry, since they now know that they can control their consumption—there is no deep feature driving it—and give up whenever they want. As one often hears smoker say: 'I know I'm not addicted, since I have given up many times in the past'. The second possibility, not terribly different to the first, is to accept that they are addicted, and that the this is something that needs to be overcome, but to think that the new evidence shows that this doesn't have to happen yet since it can happen easily enough later on. This loses the idea that the behaviour is currently under control, but it maintains the thought that it could be brought under control whenever needed: I am addicted, but I can give up whenever I want.

A third possibility follows what happened with the psychopath button. Our addict might infer from the difficulty of the initial struggle that they are addicted; might work hard to try to overcome the addiction; might interpret their success as evidence that they are not in fact addicted; might therefore allow themselves to go back to start consuming again; and might then interpret this as evidence of addiction all over again. This is not the steady state consumption of the earlier outcomes, but a move in and out of consumption that might happen on different timescales depending on the individual.

In short then, if an addict has a motivation to continue consuming, they can read pretty much any signal as an indication that it would be right to do so. The situation seems utterly hopeless.

5. WHERE FROM HERE?

Of course, this conclusion is too hopeless. For we know that people do overcome addiction. They cannot always be caught in a trap of misreading the signals. So what happens when they are successful?

It could be that sometimes people read the signals more accurately: difficulty is read simply as a signal that hard work is required; ease simply as a signal that the addiction can be overcome straightforwardly. My suspicion though—certainly one in need of empirical support—is that this is not so common. I suspect that escape more often comes, not from reinterpreting, but from circumventing the self-signalling.

One way of doing that would be to give up on the idea that addiction is a deep feature. If there is nothing to be signalled other than the behaviour itself and the experiences that go with it, then no conclusions can be inferred about why it is futile or unnecessary. Admittedly there may be concern that if the process of giving up is difficult, the addict will conclude that they are unlikely to stick with it; that requires no deep feature. But if the addict is succeeding in resisting now, that should count as good evidence that they will have what is needed to resist in the future.

Nevertheless, I doubt that many addicts think in quite this way. As we have seen, addiction is widely held to be a deep explanatory feature: popular culture points that way, as does the scientific consensus. Even if it would benefit addicts to come to believe that addiction is not a deep feature that they possess, it seems unlikely that they would manage to do so. So if they are to avoid the influence of corrupt self-signalling, they will need another approach.

Given what we know about how to resist temptation in other contexts, I suggest that the best approach is simply to ignore any signals. Just as the children maintained their resolve to hold

out for marshmallows by refusing to reconsider their resolutions, so addicts successfully give up by refusing to reconsider theirs, despite any apparent information that they may receive. Self-signalling is likely to be a corrupt source, though it will not seem that way under the power of temptation. Addicts should resolve to ignore it from the outset.

For some idea of how this might work, let us look again to those Christians who accept predetermination of election. Through the late 16th and early 17th centuries there was increasing concern amongst English Puritans as to whether the elect could gain assurance of their election. Such assurance had been a central tenet for both Luther and Calvin, and had been a key feature in their break from the Catholic church. But as the Puritans became more aware of the possibilities of self-deception, so they became more concerned about the practical realities of achieving it. As Daniel Dyke puts the position in *The Mystery of Selfe-Deceiving* (1615):

> Onely God of himselfe exactly knoweth the secrets of the heart. There is a great mingle-mangle and confusion of thoughts, even as there is of drosse and good metall in silver and gold, which lie so confused together, that to the eye of man the drosse is not discernable. (402)

The Reformed churches never went so far as to deny that the elect could achieve assurance from careful introspection. Dyke listed a set of tricks that the flesh—the 'olde foxe'— was likely to use to cover its tracks; by being aware of these, and by plowing with 'God's Heifer', assurance could be gained, although it was not guaranteed (Dyke 1615, p. 4; p.399). The Puritan position is well summarized in the *Westminster Confession*:

> This infallible assurance doth not so belong to the essence of faith but that a true believer may wait long and conflict with many difficulties before he be partaker of it: yet, being enabled by the Spirit to know the things which are freely given him of God, he may, without extraordinary revelation, in the right use of ordinary means, attain thereunto. (XVIII 3)

Given the importance attached to assurance by the Reformed churches in opposition to Rome, we can see why this was never questioned. But if we look to those outside that tradition, we can see, along with the same focus on self-examination, a rather different response to the

problem that was posed by self-deception. Pierre Nicole was a Jansenist, who like the Puritans, believed in predestination. He too held that our motivations will often be 'obscured by the clouds of concupiscence' in ways that cannot be penetrated, and he too was concerned with the kind of vigilance that is required if we are to achieve any kind of self-knowledge. But when it came to how to act, he took a rather different tack. Our actions should not be guided by an attempt to fathom our own motivations, or by responding to our internal states. Rather, our gaze should be outward. We should simply aim to perform the outward actions, which we can recognize and regulate, in accord with God's will, with the hope that our internal motivations will come into line. [14] This is not an irrationalist response. Nicole thinks that we have good grounds on which to discern how it is that God wants us to act, and excellent reasons to follow him. It is simply that, by not trying to gain information from sources that are not open to us, sources that are likely to mislead, we will be more able to do what is right.

The case with addiction is not quite parallel. Here there may be times when looking inwards is revelatory—just not when in the grip of temptation. But in times of temptation the approach that Nicole advocated seems just what is required. If the aim is to escape the addiction, the focus should not be on internal states and what can be gleaned from them; it should not be on self-signalling. It should instead be on the behaviour that is needed.

REFERENCES

Bandura, Albert 1992: Exercise of Personal Agency Through the Self-Efficacy Mechanism. In R. Schwarzer (ed.), *Self-Efficacy*, 3–38. Bristol PA: Taylor and Francis.

[14] (Nicole 1715) I.7, pp. 75–8; see also (Moriarty 2006) pp. 385–6. I am very grateful to Prof. Moriaty for discussion here.

Bénabou, Roland, and Tirole, Jean 2011: Identity, Morals, and Taboos: Beliefs as Assets. *Quarterly Journal of Economics* 126, 805–55.

Bodner, Ronit and Prelec, Drazen 2002: Self-signaling and diagnostic utility in everyday decision making. In Isabelle Brocas, and Juan D. Carrillo, eds. *Collected essays in psychology and economics,* 105–126. Oxford: Oxford University Press.

Bozeman, Theodore 2004: *The Precisionist Strain—Disciplinary Religion and Antinomian Backlash in Puritanism to 1638*. Chapel Hill: UNC Press.

Dyke, Daniel 1615: *The Mystery of Selfe-Deceiving: or A Discourse and Discovery of the Deceitfulnesse of Mans Heart* (page references to the 1633 edition). London: William Stansby.

Egan, Andy 2007: Some Counterexamples to Causal Decision Theory. *The Philosophical Review* 116, 93–114.

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

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

Holton, Richard and Berridge, Kent 2013: Addiction Between Compulsion and Choice. In Neil Levy (ed.) *Addiction and Self-Control,* 239–68. Oxford: Oxford University Press.

Moriaty, Michael 2006: *Fallen Nature, Fallen Selves*. Oxford: Clarendon Press.

Nicole, Pierre 1715: *Essais de Morale* Vol I. Paris: Guillaume Desprez.

Petry, N. M., Alessi, S. M. and Rash, C.J. (2011): Contingency management treatment of drug and alcohol use disorders. In J. Poland and G. Graham (eds.) *Addiction and Responsibility* 225–245. Cambridge MA: MIT Press.

Pickard, Hanna 2016: Denial in Addiction. *Mind and Language*, forthcoming.

Quattrone, G. A., and Tversky, A. 1984: Causal versus diagnostic contingencies. *Journal of Personality and Social Psychology*, 46, 237–48.

Schumacher J. *et al* 2007: Meta-analysis of day treatment and contingency-management dismantling research. *Journal of Consulting and Clinical Psychology* 75, 823-8.

Sripada, C. unpublished: At the Center of Agency, the Deep Self.

Weber, Max 1904/5: *The Protestant Ethic and the Spirit of Capitalism*, trans Talcott Parsons. London: Routledge, 1992

Wieber, F. , Sezer, L.A. , & Gollwitzer, P. M. 2014: Asking 'why' helps action control by goals but not plans. *Motivation and Emotion*, 38, 65-78.