https://doi.org/10.15446/ideasyvalores.v66n3Supl.65651

EXPLAINING IRRATIONAL ACTIONS

LA EXPLICACIÓN DE LAS ACCIONES IRRACIONALES

JESSE S. SUMMERS*
Duke University - Durham - Estados Unidos

* *j.s.summers@gmail.com*

Cómo citar este artículo:

MLA: Summers, J. S. "Explaining Irrational Actions." *Ideas y Valores* 66. Sup. N.°3 (2017): 81-96.

APA: Summers, J. S. (2017). Explaining Irrational Actions. *Ideas y Valores*, 66 (Sup. N.°3), 81-96.

CHICAGO: Jesse S. Summers. "Explaining Irrational Actions." *Ideas y Valores* 66, Sup. N.°3 (2017): 81-96.

ABSTRACT

We sometimes want to understand irrational action, or actions a person undertakes given that their acting that way conflicts with their beliefs, their (other) desires, or their (other) goals. What is puzzling about all explanations of such irrational actions is this: if we explain the action by offering the agent's reasons for the action, the action no longer seems irrational, but only (at most) a bad decision. If we explain the action mechanistically, without offering the agent's reasons for it, then the explanation fails to explain the behavior as an action at all. I focus on cases that result from compulsion or irresistible desire, especially addiction, and show that this problem of explaining irrational actions may be insurmountable because, given the constraints on action explanations, we cannot explain irrational actions both as irrational and as actions.

Keywords: action, addiction, explanation, irrational.

RESUMEN

Algunas veces queremos entender las acciones irracionales, o las acciones llevadas a cabo por alguien de manera tal que entran en conflicto con sus creencias, sus (otros) deseos, o sus (otros) objetivos. Lo desconcertante de las explicaciones de tales acciones irracionales es que si explicamos la acción mediante las razones que tuvo el agente para ejecutarla, la acción ya no parece ser irracional, sino a lo sumo una mala decisión. Si explicamos la acción de manera mecanicista, sin ofrecer las razones del agente para ejecutarla, entonces la explicación no logra dar cuenta del comportamiento como una acción. Este artículo se enfoca en casos que son el resultado de compulsiones o deseos irresistibles –especialmente la adicción–, y muestra que el problema de explicar las acciones irracionales puede ser infranqueable, puesto que, dadas las restricciones que caracterizan las explicaciones de las acciones, no podemos explicar las acciones irracionales como irracionales y como acciones al mismo tiempo.

Palabras clave: acción, adicción, explicación, irracional.

There are many types of irrational action, from the expressions of mental illness to smoking to eating a slice of cake I know I'm going to regret immediately afterwards. We often ask why people act irrationally. Sometimes we're just asking why people did something we don't understand. But other times we think we do understand yet still want an explanation. This is the case for irrational actions, actions a person undertakes *given* that their acting that way conflicts with their beliefs, their (other) desires, or their (other) goals. Why, you ask an agoraphobic, won't you leave the house to see your family? Why, you wonder of the smoker, does she continue to smoke knowing the health and financial consequences? Why, I ask myself, did I just eat a dessert that was large enough for four people? These are the conflicts that are irrational and difficult to explain.

Different actions are explained differently, and the very same action may have many compatible explanations. But there is something puzzling about *all* explanations of irrational action. The puzzle is this: if we explain the action by offering the agent's reasons for it, the action no longer seems irrational, but only –at most– a bad decision. If we explain the action without offering the agent's reasons for it, that other explanation fails to explain the behavior as an *action* at all.

My intention here is to show that this is a genuine problem we face in trying to explain irrational actions. Perhaps it's even an insurmountable problem. Given the constraints on action explanations, perhaps we simply cannot genuinely explain irrational actions as both irrational and as actions. More optimistically, once we understand the limitations of our action explanations, perhaps the need to explain irrational actions as such will simply seem misguided and unnecessary. First, though, I'll make the problem clearer.

What Is to Be Explained

People perform all kinds of actions that are hard to explain. I don't understand why people collect stamps or go birdwatching, read complicated Russian novels, or learn to speak Elvish. But, when I say "I don't understand" these actions, what I mean is that I don't understand why the person enjoys them or finds them fulfilling; what I *do* understand is that the person does them *because* she finds them enjoyable or fulfilling. And perhaps I even understand that I would find them fulfilling, too, if circumstances were different (perhaps even if I just tried them).

Irrational actions –as I use the term here– aren't like that. Given that I don't enjoy Russian novels, it would be a bad decision for me to bring only *War and Peace* to read on my vacation. It's a bad decision because, although I might want to give this classic a chance, I know –or have enough information to know– that I'll quickly want to read

something else. But, it would be irrational for me to bring on vacation only the Russian edition of the book, which I can't read, given that I want a book to read. It would be irrational to bring two identical copies of it so that I can read the book twice, given limited packing space. These actions are not best understood as coming from strange but understandable preferences, but as resulting from decisions that can't be made consistent with some of the person's other, immediately salient beliefs and desires. They're irrational actions.

There is a worthwhile debate to have about why an action is irrational instead of merely the result of a bad decision.[1] Why are actions based on delusions irrational, but actions based on false beliefs –of the sort we all have– mere mistakes? Are impulsive actions, of the sort that often illustrate discussions of "akrasia" and weakness of will, genuinely irrational in this sense, or do they reflect rapidly oscillating preferences, or something else entirely? Such questions are worth answering, but here I'm going to stick with clear cases of irrationality and a simple distinction between bad decisions and irrationality. What is true of irrational actions, but needn't be true of bad decisions in general, is that there is some significant internal conflict between the intentional state that causes the action and the person's other intentional states, in particular her beliefs, (other) desires, or (other) goals. This internal conflict is how I'll characterize the irrationality.[2]

To make things even simpler, I'll focus exclusively on irrational actions that result from something like a compulsion or an irresistible desire. If the account doesn't generalize to all irrationality, it will still be an interesting case study in how to explain compulsions, though I'll talk here as if the account generalizes.

Difficulties in Explaining Irrational Actions

Let's return to the opening cases, then, in order to see what the difficulty is in explaining irrational actions. One way that you can try to explain why I had the giant dessert is by citing my reason(s). If you say I had a reason for eating the cake, namely that it tasted so good, then this

1 I have in mind both the contemporary debate about why to be rational (*cf.* Broome 2007; Kolodny 2005), as well as an earlier debate about forms of irrationality (*cf.* Davidson 2004a, Davidson 2004b).

2 Donald Davidson (2004a) similarly starts with such cases of "subjective" irrationality in order to determine what exactly makes "objective" cases of irrationality –like a failure to respond to evidence– irrational. It's not just for reasons of space that I don't follow Davidson in discussing more objective cases; I'm also not sure that my criticisms will apply so strongly to those cases. The subjective and objective senses of "irrationality" may be less connected than Davidson realizes, though, and this is a larger topic than I can address here.

now sounds like –at most– an ordinary bad decision that leads to a suboptimal result: I really should have considered how much I'd be eating and what I'd feel like afterwards, and I was wrong to think taste would trump this gluttonous feeling I now have (of course, citing my reason has now raised the possibility that the results weren't even suboptimal, but just that my decision had some unpleasant consequences: I feel terrible, but it was totally worth it). The point is that, by citing one's reasons, we explain the action as the result of a decision, probably a bad decision or a decision that leads to suboptimal results; but bad decisions aren't *eo ipso* irrational.

On the other hand, you might try to explain my action in another way, without citing my reasons. Consider the other ways you could explain my action. My "sugar addiction" made me do it, my body "needed" the calories after my long workout today, the craving was "overwhelming" or "irresistible", or my impaired dorsolateral prefrontal cortex couldn't stop my limbic system from causing me to eat. All of these explanations would explain the behavior of my feeding myself cake that I recognize I have overwhelming reason not to eat. But explaining my behavior (*i.e.* why my body moved in the way it did) isn't clearly the same thing as explaining my action (*i.e.* why I acted as I did).

To see this, consider an extreme case: I'm unconscious, but I'm hooked up to a machine that makes me sit up. That is, it physically moves my body into a sitting position. If you ask why I sat up (*i.e.* for an explanation of my behavior) the best answer is: the machine sat me up. Nothing about me, about my desires or goals, about what I thought were good reasons for sitting up, or about what I thought would be good about sitting up, explains my sitting up. It could be true that I'd been wanting to sit up before I fell asleep, and may even have been dreaming about sitting up. But the mental states don't explain my sitting up: the machine best explains my sitting up.

By contrast, consider an ordinary case in which I'm lying down and decide I want to get some water, sit up, then go to the kitchen to get water. In this case, the best explanation of why I sat up was (something like) my wanting some water and sitting up as a way of moving towards getting the water. Some mental states, like wanting water, can themselves be the best explanation of my action. Without getting into subtleties, the content of my mental state –wanting the water– is the primary or distinctive part of the explanation of why I sat up.[3]

3 I'm avoiding the issue of whether such mental states are all reasons. Nothing here depends on that question itself, and, while some of the issues surrounding what it means to act for reasons can also arise, they needn't be resolved in order to make these more general points.

Moreover, we can't simply sort behaviors into those done for reasons and those not done for reasons, or "actions" and "mere behaviors", respectively. Not only is this because it would require a notion of reason clear enough to illuminate difficult cases, but also

Consider now an intermediate case. I am not asleep and tied to a machine, so the machine doesn't best explain my sitting up, but other circumstances lead you to suspect that my desire to get some water isn't the best explanation of my action. Perhaps I didn't first have the thought, and my body moved without my forming the conscious desire. In that case, we might say that my body –but not I– "wanted" to sit up. Perhaps it's a reflex of some sort. Maybe the best explanation of my sitting up refers only to my central nervous system, to some errant chemical and electrical impulses. My desire to sit up –or to get water, which requires sitting up– isn't the best explanation of my action because I don't even have such a desire.

Consider a harder intermediate case: I actually have some desires or other mental states that would explain the action, but those mental states are not the *best* explanation of my action. I'm hypnotized to want to sit up whenever I hear a bell. I hear the bell, but I don't think, "I'm sitting up because the bell just rang". Instead I think, "Hm… I now feel like sitting up. I'll get some water". In which case my desire to sit up does explain my action, and I might even think I'm sitting up to get some water, but the *best* explanation of my sitting up isn't my proximate desire to sit up but is instead the hypnosis that brought about that proximate desire. If you explained why I sat up by referring only to my proximate desires and without mentioning my hypnosis, you would have left out the most important part of the explanation, the part that distinguishes my hypnotic sitting up from ordinary sitting up.

More controversially, but also more commonly, if I'm addicted to a drug but have resolved to lie here and not use the drug, but I nevertheless succumb to my desire and sit up to get more of the drug, then what explains why I sat up? While I do have a desire to sit up and even a desire to go get the drug, the *best* explanation of this addictive behavior isn't simply the proximate desire to get the drug but the neurological reward system, diminished executive control, and other features that, together, neurologically and psychologically comprise my addiction, all of which themselves explain –as the hypnosis did– why I come to have that proximate desire. If you explain my action by citing my proximate desire to get the drug without also mentioning the underlying addiction, you explain my action in a way that doesn't distinguish it from a non-addictive whim, a "mere" desire to get the drug.

because in many cases of irrational action, a person does have reasons for acting; but we suspect, as I'll discuss below, that the reasons are not the best explanation of the person's action. Therefore, this simple distinction –absent significant development– will fail to capture the interesting cases of irrational actions.

Of course, you might not think that's the best explanation of my action. A theoretical distinction between the best explanation of one's actions and other explanations of one's actions will not resolve which explanation is best in any particular case or how to determine it. You might therefore agree that there are neurological and psychological changes that come about from previous drug use, but you deny that those best explain my present action, and you insist that the best explanation of my present action is that I simply wanted to use the drug. Or you may have an entirely different "best" explanation of that desire. Those debates are all subsequent to the distinction between the best explanation of one's action and other possible explanations (*cf.* Summers 2017).

Now, what is hard to explain in cases of purportedly irrational action is why a person acts as she does given that she has conflicting beliefs, desires, etc. that are themselves sufficient for her not to act that way. It's not hard to give *some* explanation of the action: one can often cite at least the proximate beliefs and desires, or the physical forces that acted on (or in) the body. Why did I eat the fourth piece of cake? It tasted good. Why did he use heroin? It felt good. But such explanations risk being shallow and uninformative, like explaining why someone ran naked across a highway by saying that, well, apparently he had the desire to run naked across the highway. That may be true, but what we're after is an explanation that gives some deeper desire or that explains why that desire wasn't in fact the operative one.

The reason it's so hard to explain irrational actions, then, is that the offered explanations fall into two categories, neither of which is satisfying. The first is that the best explanation cites some reason for the action that makes the action just a case of a suboptimal or otherwise bad choice. I ate the fourth piece of cake because it was good, and I just didn't care about how I would feel afterwards.

The second kind of explanation avoids citing my reasons or intentions entirely. I ate the fourth piece of cake because my brain made me do it or even because some mental state made me do it: my desire to do it was literally irresistible, so I did it despite wholeheartedly and desperately trying not to.

Neither kind of explanation explains why the person acts intentionally in one way given that her relevant intentional states overwhelmingly support a different action. What is so hard to explain in cases of irrational action is why the person acts (or behaves) in some way *given* that the action (or behavior) conflicts with his beliefs, (other) desires, or (other) goals. It's this conflict that makes the action so difficult to understand and hence to explain. I can easily explain why I eat tasty cake; what I can't easily explain is why I eat tasty cake long after I've resolved that I shouldn't eat any more.

The problem goes beyond gluttony to more serious cases of irrationality. If the *reason* the agoraphobic doesn't leave the house is that she is afraid of having a panic attack, then this makes sense of the action, and we can then discuss whether this is a reasonable decision, given the unfortunate risk, or a bad decision, perhaps because she overestimates the risk or seriousness of having a panic attack. However, if we don't cite her reason in explaining her action, then we explain the agoraphobic's action in some other way, probably by citing the (non-reason) causes of her action: her phobia, her overactive amygdala, etc.

Likewise, the smoker's reason for smoking may be to calm her nerves or otherwise "self-medicate" (*cf.* Darke 2013; Khantzian 1985; Pickard 2013). If she's considering such reasons to smoke weighed against the reasons not to smoke, then she might make a bad decision, but that's not obviously irrational. Perhaps having calm nerves is worth the health risks. If it's not, then she made a bad choice. Or, again, if we don't cite her reasons, then we need to explain her action in another way, but if we cite the addiction, or directly cite her neurobiology –her ratio of D2 to D4 dopamine receptors and impaired DLPFC– then we have explained her behavior by ignoring the conflict.

Nothing is wrong with these explanations of gluttony, phobias, or addictions that I've gestured towards, except –and this is the concern here– they are not explanations of an irrational action that is opposed by one's other intentional states. In all of these cases of putatively irrational action, the explanation either cites a reason for the action, in which case the action may be the result of a bad decision, but is not clearly irrational since it is not opposed overwhelmingly by one's other mental states; or, the explanation does not cite a reason, in which case we have not explained the action as intentional at all. This apparent dichotomy is worth more attention.

Explaining Actions by Reasons

If all I've done so far is establish a dichotomy between more-or-less explicit, rational decisions and arational behavior, that would be a problem, since that dichotomy is false. Many of our actions can be explained both as the result of some rational decisions and as the result of arational causal processes: when I decide to do something and then do it, my brain also causes some behavior. We don't have to deny that both explanations are true or settle how they are related when we insist that, in a given situation or for some purposes, one of the explanations is better. So it's worth a digression to show the range of cases in which we cite reasons to explain actions: both cases in which arational processes are less good as explanations and cases in which there is no explicit, rational deliberation. In both cases, we explain actions by citing reasons.

As Donald Davidson puts it, to explain an action is to give the person's reasons for the action as they appeared to her, to state what about the action appeared good to the person (*cf.* Davidson 2001 3-20). And that often is the way we explain actions: you ask me to explain why she left the party so early, and I tell you that she wanted to stop by the store before it closed. That is the reason that she would sincerely give and that did in fact play a deciding role in her decision to leave early.

We use the same model of citing reasons even when those reasons were not reasons the person explicitly represented to himself. "I think the real reason she left the party was because it was the only way she could get out of that boring conversation, even though she's too nice to admit that to anyone, or maybe even to herself".

We may even cite one's reasons when we know that a person did not and *could* not explicitly consider those reasons at the time: "Why is he dating her? He would never admit this is the reason, but it's because she reminds him of his mother". Of course, if it ever occurred to him that he was dating her because she reminded him of his mother, he would break up with her immediately, so clearly he doesn't know that's his reason. Yet we do coherently and ordinarily talk about such reasons, reasons that a person doesn't, perhaps even couldn't, explicitly recognize in acting.

When we talk about reasons that the person does not explicit represent in acting, we have various ways of indicating that the person does not explicitly represent to himself the reason. We speak, for example, of his "subconscious" reasons or "real" reasons or "deep" reasons. We might not use the word "reason", but the explanation cites features of the situation as they appeared to the actor –the way she and his mother look and act– and as they would have figured in some ("subconscious") reasoning ("they look and act alike, and I love my mother, so I will love her, too"), regardless of whether one would explicitly recognize or affirm that those features were reasons for doing what one did, or even that they could be good reasons to act on (*cf.* Horgan and Timmons 279-295).

Further, we talk about reasons for acting without regard for whether there is some underlying neurological or other arational or non-rational explanation of one's action. When I tell you I went to the store to buy peanut butter, I don't thereby deny all of the physical causes that also (partially) explain my action. My neurobiology played a role, as did hormonal changes brought about by hunger, and no doubt thousands of small factors each of which played some small role in explaining my action. I don't deny these when I cite my reason: I'm instead explaining my action by emphasizing one particular factor, namely the content of my reason for acting. We can, as always, ask what makes this the best

explanation of any particular action, but we can offer the explanation without denying that there are other explanations.

I'm using "reason" in this very broad sense, encompassing all reason-based explanations, from implicit, subconscious, real, deep reasons to explicit reasons used in explicit reasoning, and assuming that there may always be other available explanations. I'm including as a reason what some would call "mere" desires, as long as they can figure in one's (implicit, subconscious, etc.) reasoning: "Why did she go for a walk? She just wanted to". The reason is that she wanted to. And I'm even including reasons in cases in which they're explicitly denied: "Why did she go for a walk? No reason: she just felt like it".[4] The reason is her feeling, her inclination, and what is probably denied here is that she had any reason greater than that (any "good reason"). My point in talking about reasons so broadly is not to take a stand on what reasons are, but to be able to draw a contrast between reasons on the one hand and mechanistic explanations on the other, so I'm ignoring differences among reasons to emphasize this other contrast.

One way of explaining actions, then, is to cite the reasons –understood broadly– that led to that action, even if, as just suggested, those reasons were not explicitly taken as reasons.[5] That is, when we say, "He decided to dance on the table for no reason whatsoever", we may still explain the action as the result of some intentional states (he felt like dancing, he wanted "deep down" to impress or embarrass his companions, etc.) that could have figured in his reasoning, just as when we say: "He decided to dance on the table to show to all around him that he flouts convention". (Obviously, these aren't the *same* explanation.) These are the explanations I have in mind when I say we explain an action by citing the agent's reasons.

4 What I'm including here are many intentional states, like considerations that seem to the agent to count in favor, but also including desires, goals, and even representations of one's own emotions. This is in part because the same action can often be explained as caused by a reason or by some other intentional state: "Why did she take up bird-watching? She was bored". "Why did he yell at the passing car? He was angry". These explanations may be a shorthand for an explicit deliberation (*e.g.* "She was bored and wanted to relieve the boredom, so she looked for something that would relieve the boredom"), or they may express a "subconscious" decision, or they may be a declaration that there was no decision, not even a "subconscious" decision, in the case. These points are far subtler than I can address here.

5 It's not always true that we explain an action by citing the decision to act in that way, since actions may go awry. The explanation of my throwing the ball through the window may be that I decided to throw the ball to my friend who was in front of the window.

Explaining Actions Mechanistically

The contrast to this first way of explaining actions is to explain actions by citing something that caused the movement that we are trying to explain. "Why did he kick the doctor? The doctor tapped his patellar tendon with a reflex hammer". "Why did she drive off the road? She had a seizure". The contrast with the first way of explaining action is that explanations of this second type do not cite, or assume, that there was a decision or choice that better explains the action. Even if he didn't make a decision, it is still possible that the person had *reasons* to perform the action –the doctor with the reflex hammer may have been just the kind of jerk who deserves a good kick– but those reasons wouldn't be a good explanation in the absence of a decision or choice.[6]

Notice, now, that subtler versions of these mechanistic explanations are sometimes present in neurological explanations of behavior. Consider for example an explanation of addiction as a "hijacking" of the neurological reward-system. There are many such accounts, as well as more general accounts of addiction as a "brain disease", so I'll use Timothy Schroeder's account to illustrate (*cf.* Schroeder 2010 391-407; Arpaly and Schroeder 274-289). These accounts do not explain addictive action by citing a person's reasons but by citing something neurological that explains the person's behavior.

Remember, what we often want to explain about addictive action isn't why people use drugs –it seems obvious enough why people would use pleasurable drugs– but why people use drugs when the drawbacks of use are serious or the alternatives to use are good. Schroeder's account tries to answer the question about irrational actions neurologically. In summary, Schroeder's proposal is that the addict uses a drug because the addict's dopaminergic system, the neurological system that encodes which actions and objects are most rewarding, is modified by the drug in a way that makes the drug neurologically more rewarding than it otherwise would be (*cf.* Schroeder 2010 391-407). That neurological process does not change how rewarding the person would *tell* you that the drug is, though. As a result, the person responds to the drug as if it's more rewarding than many other things in his life, but he does not believe that he finds it more rewarding. To use a distinction from the addiction literature: the addict "wants" the drug more without also "liking" it more (*cf.* Berridge and Robinson 1995).

Although this general proposal ignores the various effects of specific drugs, and we should quibble with whether it does justice to the range of

6 Some may want to draw this distinction using a distinction between "explanatory" and "justificatory" reasons, though there may be several such distinctions using these same terms. I'm not committed to any one such distinction.

neurological effects that contribute to addiction (*e.g.* it says nothing about impulse control or other changes in executive control of one's actions, which are not simply changes in the reward system), it nevertheless illustrates how a neurobiological explanation leaves unexplained the irrational action. To see this, consider two ways to understand how Schroeder's proposal would explain the irrational action of drug taking.

The first way of understanding Schroeder's proposal is that modifying the reward system causes a person to have reasons to act that they did not initially have, *e.g.* it makes withdrawal from cocaine cripplingly unpleasant. Now using cocaine seems like a decision to stave off withdrawal, which may be a bad decision, and one may have many reasons not to do it as well, but that's not an irrational action either. Perhaps the addict should –by his own lights or all-things-considered– accept withdrawal instead of using, but there is no inherent conflict in using in order to avoid withdrawal despite side effects. Actions and medications both have side-effects, and, on this view, taking cocaine would look a lot like any other action or medication.

Another way in which we can understand this proposal is that the addictive behavior is caused by changes in the reward system. Those causes don't affect what the person takes to be good reasons for and against taking drugs, but, regardless of what they take to be reasons, they are caused to act to use drugs. That proposal, even if it were plausible, bypasses what was confusing about the addictive action, and what made it irrational: why do addicts use when they *know* they should stop? This explanation is that the reward system causes them to act *regardless* of their beliefs, (other) desires, and (other) goals, so it explains why they behave as they do, but not why their action is irrational. It is not irrational for me to sit up despite my strongest desire not to sit up if it's the bed I'm strapped to that sits me up. My sitting up isn't irrational because the behavior is caused by something other than the mental states involved in my making a decision about what to do. My actions don't match my desires, intentions, etc., but those mental states are all in accord (I don't want to sit up; I don't want to take drugs), and they are simply overridden by something physical.

A modified version of this second proposal is that the person isn't moved by something physical but is moved by something mental, by a mental state that functions in the same way that something physically or neurologically irresistible would. The addict has a desire that moves the person to take the drug regardless of any other desires or beliefs. She could be, as it were, kicking and screaming, begging her psychology not to let her take the drug, but an irresistible desire is just that, and it would do no good. Of course, almost no one would ever think of a mental state as literally irresistible in this way, so most talk about

"overwhelming" cravings and desires are probably best understood as instances of the first way of understanding this proposal, *viz.* that the desire gives the person a very strong reason to act in a particular way (*e.g.* that it would be very psychologically painful to resist acting in that way).

On neither way of construing the proposal, then, do we have an explanation of the irrational action of taking some drug, if what we wanted to explain was why someone takes a drug as an intentional action, one done for reasons –broadly understood–, despite one's reasons clearly counting against taking the drug. What we have instead are these two general strategies for explaining irrational actions.

One way to explain action is to cite the reasons behind the action. But this makes the action an intentional action, like any other, which may be a bad decision destined to yield a suboptimal outcome, but that is no longer clearly irrational.

The other way is to cite some probably physical mechanism that causes the behavior, but this seems not to explain the conflict of the irrational action at all. The behavior is no longer explained as issuing from intentional states that are in conflict with one's other intentional states; instead, one's action issues from mechanisms that bypass the mental states entirely.

Neither way of explaining an irrational action succeeds in explaining the action as both an action –*i.e.* issuing from intentional states– and as irrational –*i.e.* issuing from intentional states that are in strong contrast to one's other intentional states–.

Conclusion

Where does that leave us? Perhaps this just is the conclusion: irrational actions cannot be explained in a way that embraces the conflict that makes them irrational while also explaining them as actions. Perhaps the best we can do is explain the action in such a way that it no longer appears irrational or no longer appears to be anything more than mechanistically caused behavior.

Interestingly, this conclusion, while perhaps unsatisfying, would account for the regular recurrence of the argument that no desires are literally irresistible (*cf.* Feinberg 1970; Korsgaard 1997; Pickard 2012; Watson 2004). The common argument goes something like this: if a desire were literally irresistible, then there could be no cases in which the person resists the desire successfully. There *are* cases in which the person resists the desire successfully. Therefore, the desire is not literally irresistible. Desires that *seemed* literally irresistible must in fact have just been very hard to resist.

The argument is simple and (arguably)[7] valid, yet its conclusion seems to leave unexplained exactly what we most want explained in cases of compulsion: why *do* we do these things when we know better? The argument, even if valid and convincing, doesn't help us to understand this puzzling conflict. We need a way to explain those *actions* –not reflexes, not seizures, not twitches– that the person performs *given* the conflict of the intentional action with the person's other intentional states. The most obvious ways to explain those actions are that the person's intentional states are bypassed or that some intentional states are themselves irresistible, so the person acts this way despite also wanting not to act that way. Therefore, in the absence of any better explanation of irrational actions as genuine actions, we continue to talk of compulsion as something like literal irresistibility because there is no alternative way to talk about compulsions that makes sense of them as genuinely irrational actions.

This puzzle goes beyond compulsion. In the case of compulsion, we respond to this dilemma either by rejecting the claim that compelled actions are genuinely actions –addicts should be pitied, not punished, since the action is not genuinely their own intentional action–, or by insisting on finding or positing the reasons for the action (she claims she wants to quit, but *really* she wants… or *really* she's using because she's self-medicating some underlying…). Neither way out of the dilemma is satisfying as an explanation of irrational actions, though, since neither explains the action as an action while also adequately acknowledging the conflict. And this is true for many other types of irrational action, from procrastinating –did I procrastinate because (my reason) I was actually afraid to start working on this project, or did something in my psychology or neurobiology make me procrastinate, regardless of my intentional states?– to mental illnesses –did the OCD sufferer rewash because his hands don't feel clean yet or because something in her brain made him rewash regardless of his desires?–. The explanatory dilemma remains the same.

My goal here is to make the case for this explanatory difficulty and to suggest that it should be taken seriously, not to propose the solution. I don't have one, and there may be no general solution, even if particular cases seem explicable. But I'll conclude with the hint of a possible general solution.

One common assumption is that explaining an action by citing a reason imputes something like a decision or choice –perhaps implicit– to the actor. And it is understood as a constitutive claim about action

7 The argument has problems, at least in a simple form (*cf.* Sinnott-Armstrong 2013; Sripada 2014).

that actions are done for reasons in that they issue from such (implicit) choices (*cf.* Katsafanas 2013).

But perhaps our view of actions shouldn't be the constitutive claim about what actions *are* but should instead be an aspirational claim about what makes for good actions, or for actions constituting a good life. Perhaps actions *ideally* issue from explicit deliberation about their choices, but an alternative sense of "action" refers to behavior grounded in intentional states: something more than reflex, but something less than choice. Irrational actions may be but one example of such actions broadly understood, since they are clearly actions (*e.g.* not reflexes), but not reasons-based actions of the sort that an ideal agent should aspire to. Whether irrational actions can be satisfactorily explained as non-ideal actions but actions nonetheless is not obvious, but other strategies for explaining such actions are also far from obvious.

References

Arpaly, N., and Timothy Schroeder. *In Praise of Desire*. Oxford: Oxford University Press, 2013.

Berridge, K. C., and Robinson, T. E. "The Mind of an Addicted Brain: Neural Sensitization of Wanting Versus Liking." *Current Directions in Psychological Science* 4.3 (1995): 71-76.

Broome, J. "Wide or Narrow Scope?" *Mind* 116.462 (2007): 359-370.

Darke, S. "Pathways to Heroin Dependence: Time to Re-Appraise Self-Medication." *Addiction* 108.4 (2013): 659-667.

Davidson, D. "Actions, Reasons, and Causes." *Essays on Actions and Events*. Oxford: Oxford University Press, 2001. 3-19.

Davidson, D. "Incoherence and Irrationality." *Problems of Rationality*. Oxford: Oxford University Press, 2004a. 189-198.

Davidson, D. "Paradoxes of Irrationality." *Problems of Rationality*. Oxford: Oxford University Press, 2004b. 169-187.

Feinberg, Joel. "What is So Special About Mental Illness?" *Doing and Deserving*. Princeton: Princeton University Press, 1970. 272-292.

Horgan, T., and Timmons, M. "Morphological Rationalism and the Psychology of Moral Judgment." *Ethical Theory and Moral Practice* 10.3 (2007): 279-295.

Katsafanas, P. *Agency and the Foundations of Ethics: Nietzschean Constitutivism*. Oxford: Oxford University Press, 2013.

Khantzian, E. J. "The Self-Medication Hypothesis of Addictive Disorders: Focus on Heroin and Cocaine Dependence." *Am J Psychiatry* 142.11 (1985): 1259-1264.

Kolodny, N. "Why Be Rational?" *Mind* 114.455 (2005): 509-563.

Korsgaard, C. M. "The Normativity of Instrumental Reason." *Ethics and Practical Reason*. Eds. Garrett Cullity and Berys Gaut. Oxford: Oxford University Press, 1997. 215-254.

Pickard, H. "The Purpose in Chronic Addiction." *AJOB Neuroscience* 3.2 (2012): 40-49.

Pickard, H. "Psychopathology and the Ability to Do Otherwise." *Philosophy and Phenomenological Research* 86.2(2013): 135-163

Schroeder, T. "Irrational Action and Addiction." *What Is Addiction?* Eds. Don Ross, Harold Kincaid, David Spurrett and Peter Collins. Cambridge, MA: MIT Press, 2010. 391-407.

Sinnott-Armstrong, W. "Are Addicts Responsible?" *Addiction and Self-Control.* Ed. Neil Levy. New York: Oxford University Press, 2013.

Sripada, C. "The Second Hit in Addiction." *Moral Psychology, Volume 4: Free Will and Moral Responsibility.* Ed. Walter Sinnott-Armstrong. Cambridge, MA: MIT Press, 2014. 295-304.

Summers, J. S. "Post Hoc Ergo Propter Hoc: Some Benefits of Rationalization." *Philosophical Explorations* 21. Sup. 1 (2017): 21-36.

Watson, G. "Disordered Appetites: Addiction, Compulsion, and Dependence." *Agency and Answerability.* Oxford: Clarendon Press, 2004. 59-87.