Teleology and Degrees of Freedom[1]

By Scott Sehon (Brunswick, Maine)

I. Introduction

There is a debate in philosophy of mind about the nature of reason explanations of action, and this volume is testament to a resurgence of interest in non-causal accounts. In *Teleological Realism: Mind, Agency, and Explanation,*[2] I have proposed a non-causal account according to which common-sense reason explanations of action are irreducibly *teleological* in form. I claim that we explain behavior by citing the state of affairs towards which the agent was directing her behavior, i. e., by citing the *purpose* or *goal* of the behavior. I will not be defending that account of action explanation here, but will be assuming it and applying it to the free will debate. I will argue that the teleological account of action explanation leads to a view of free will with some interesting and attractive features.

Philosophical accounts of free will typically propose some sort of criterion for determining which behaviors count as free, e. g., that the agent could have done otherwise, or that the behavior was in accord with the agent's second-order volitions. The account is then usually used to answer the question of whether we have free will, and especially whether we can have free will even if determinism is true. To test these accounts, philosophers make arguments of various sorts, but the primary method is with cases. Here's the recipe for an objection to an account of free will: come up with a case where the theory says that the agent is free, but our intuition says that the agent is unfree, or vice versa.

I won't question the method of cases in principle; it's what we have and it makes some sense. On the picture of analytic philosophy as conceptual analysis, it makes perfect sense. Philosophers put forward a precisely spelled out version of our concept of *free will*; but if such an account is at odds with firm intuitions in particular cases, then it can't be what we mean. We know that the situation is not really so simple as that. Our ordinary concept of free will might be messy and not perfectly consistent; accordingly, any particular theory might get a few cases ›wrong‹ as judged by our intuitions. Instead of supposing that we are doing straightforward conceptual analysis, it makes more sense to assume that we are aiming for something like Rawlsian *reflective equilibrium*, in which we consider

1 Portions of this paper have been presented to audiences at Auburn University, the University of New Mexico, and to classes at Bowdoin College. In each case, I am grateful for the many questions and comments I received. For comments on an earlier draft, thanks to Sarah Conly, Kristen Ghodsee, Larry Simon, and Matthew Stuart.

2 Scott Sehon, *Teleological Realism. Mind, Agency, and Explanation*, Cambridge MA 2005.

both high level general principles and considered judgments about particular cases. By going back and forth between these levels, we hope to come up with a theory that both seems plausible in the abstract but also accounts for our strongly held intuitions about cases.[3]

I will approach the free will debate in something like this way. In section 1, I begin at the bottom level by discussing some cases and intuitive judgments, particularly cases involving weakness of will and addiction. Our intuitions and practice suggest a view of free will that seems to be largely overlooked in the literature, namely, that freedom comes in degrees. In many cases, we are neither wholly free nor wholly unfree, but in some grey region in between. In section 2, I turn to higher level principles. I begin with an overview of the teleological account of action explanation, and I will argue that this account suggests a certain positive view of free will. Roughly, the view is that behaviors are free to the extent that the agent is rationalizable when performing them. In section 3, I return to diagnosing particular cases and intuitions with this positive view in hand. I argue that the view aligns well with the intuitive idea that freedom comes in degrees, and, in particular, that the view harmonizes with our intuitions and practice regarding cases of weakness of will and addiction. Moreover, I will suggest that the teleological account of free will goes further than merely agreeing with clear intuitions; in addition, in the grey areas, the view helps us to think the cases through further, both agreeing with our strong intuitions and clarifying our muddier ones. Finally, in section 4, I consider some objections. Here too, I engage in the reflective equilibrium process, for at first glance it might appear that my view of freedom gets certain clear cases quite wrong. I will fend off this misunderstanding, thereby both defending and clarifying the view.

II. Cases and intuitions

The free will debate is informed by a large array of stock examples, and a prominent part is played by cases of weakness of will, compulsion, and addiction. Typically, though not always, philosophers assume that when we behave akratically we are still free, but that when compelled or addicted we are not free.[4] Either way, philosophers usually assume that in these test cases the answer is all or nothing: the agent is either free or not free. I suggest instead, that with cases like this, we should see freedom as impaired in varying degrees.

---

3 For Rawls's version in the context of political philosophy, see: John Rawls, *A Theory of Justice*, Harvard 1971.

4 See Jeanette Kennett/Michael Smith, »Philosophy and Commonsense: The Case of Weakness of Will,« in: Michaelis Michael/John O'Leary Hawthorne (eds.), *Philosophy in Mind. The Place of Philosophy in the Study of Mind*, Boston 1994, 141–157.

In the first instance, my claim that freedom comes in degrees is an autobiographical point about my own intuitions. If I get up from my chair to get a glass of water, I'm perfectly comfortable saying that this was simply a free action; if you push me out of a building, then my falling towards the ground is simply an unfree happening. But weakness of will and addiction cases seem much fuzzier. Consider the smoker who desperately wants to quit but lights up anyway; I have difficulty categorizing this behavior either as a simply free action (like getting up to get a glass of water) or as an unfree happening (like falling from a building after being pushed). Or consider acting against one's best judgment by eating more than one should. In some cases I might attribute full responsibility for such behavior, but if the temptation is very strong (feeling extremely hungry and having one's favorite kind of cake offered), then again I am less clear. There is some medical evidence suggesting that certain people have either a defect in the production of the hormone leptin, or have a genetic mutation that makes them insensitive to the leptin that is produced by their bodies; such people apparently feel ravenous hunger nearly all the time and become morbidly obese.[5] If such a person devours a bagful of cookies, knowing full well that they have already had sufficient calories for the day, is this a free action? Again, I have a tough time giving a clear »yes« or »no.« It seems to me that such a person is neither completely free nor completely unfree.

I have some evidence that my reactions about these cases are not completely idiosyncratic. I polled undergraduates in a philosophy of mind class about a number of cases, and I gave them five options, from option A, *completely free*, to option E, *completely unfree*. Here are the results from two of the examples mentioned above:

	A	B	C	D	E
To what extent is the following behavior free: lighting up a cigarette against your best judgment, against what you say you want to do, and where you feel like you can't help it.	5	2	6	5	0
To what extent is the following behavior free: eating 20 cookies when, because of a db mutation in your leptin receptor gene, you feel ravenously hungry, though you know you have had a sufficient number of calories for the day.	4	2	5	6	1

5 See Joseph Proietto, »The Weight Debate: Why staying lean is not a matter of ethics,« in: *The Medical Journal of Australia* 171 (1999), 611–613.

In each case, some students were willing to say that the behavior was completely free, and very few were willing to say it was completely unfree – which already seems to go against most philosophical treatments which assume that cases of addiction are unfree. But my main point is that in both of these hard cases, nearly 3/4 of the students gave an answer in the middle, somewhere between simply free or unfree.

There is other evidence that we think of freedom as coming in degrees, namely, how we think of the agency of young children. I hold my 10 year old daughter largely responsible for most things she does, good and bad; in other words, I assume that she is a free and responsible agent, although even here there will be times when I think of her freedom and corresponding responsibility as somewhat attenuated. When she was a newborn infant, of course I did not take her to be a free and responsible agent. Even basic motions of her limbs did not at first seem under her control, and it would take an awfully callous parent to blame a newborn for crying or soiling her diaper. As they proceed from newborn to normal adult, human beings become free and responsible agents. It strikes me as almost silly to maintain that there was some magic moment at which that happens.

There might be another way of accounting for these intuitions besides claiming that freedom comes in degrees. One could claim that freedom is still essentially all or nothing, but that there is a vague middle ground. The vagueness in the middle could either be purely epistemological, just a matter of our uncertainty, or it could be that there are cases that are indeterminate in principle. Some ordinary terms arguably work like this, at least in some contexts. To borrow a stock example from philosophy of law,[6] if a town ordinance specifies that vehicles are prohibited in the park, it may be hard to know whether bicycles and skateboards are allowed. Arguably, the ordinance, and the term »vehicle,« is simply vague and indeterminate. One might say the same about the word »free.«

On the basis of intuitions alone, it might be difficult to distinguish between freedom coming in genuine degrees as opposed to being all-or-nothing with vagueness in the middle. But other judgments we make point to freedom coming in degrees. For starters, we should note that legal contexts like the ordinance are somewhat artificial precisely because they tend to force a yes or no answer: bicyclists and skateboarders need to know whether they are subject to being ticketed if they go in the park. When it comes to judgments of freedom and responsibility, we can and do allow for shades of grey. We hold someone responsible if their action was free; if their behavior was completely unfree, we do not hold them responsible at all. But when we have a hard case, we do not always force a yes or no answer; rather, we mitigate the level of responsibility, and thus

6 See Herbert Lionel Adolphus Hart, »Positivism and the Separation of Law and Morals,« in: *Harvard Law Review* 71 (1958), 594–629.

blame or praise we attribute. If a recovering alcoholic loses her resolve not to drink and goes on a binge drinking episode, we will typically regard her as responsible for that action. However, if we find out that, through no fault of her own, she was placed in circumstances in which everyone around her was drinking and pressuring her to join, then we are less inclined to hold her as fully responsible.

I would not claim that any of the above constitutes a knockdown argument that freedom comes in degrees. At most, the view is suggested by our intuitions and our practices, and this should at least motivate an account of free will that makes sense of this possibility. This perhaps does not bode well for typical libertarian accounts according to which free will depends on the existence of physically possible alternatives, for the existence of alternatives would seem to be an all or nothing matter. Nor is it great for simple causal compatibilist models, according to which whether a behavior is free depends on the nature of the causal chain that led to it, for there too, depending on the details view, the causal chain will either be in the right category or not. In the next section I will lay out some pertinent aspects of the teleological account of action explanation, and I will show that this account suggests a view of free will that more nicely accords with intuitions about degrees of freedom.

III. Teleological realism and the positive account

Let us now temporarily back away from the free will debate, and look instead at action explanation. We typically explain human behavior by citing the agent's *reason* for acting. For example:

(1) Steve went to the coffee shop because he wanted to work without interruptions from students.
(2) Gina ordered decaf so that she wouldn't be kept awake that night.
(3) Hayden shucked corn because his father told him to.
(4) Josephine went upstairs to get her shoes.

Within philosophy of action, there are those following Donald Davidson[7] who construe these as causal explanations. On this view, the explanations imply that a relevant mental state of the agent caused the behavior. By contrast, on my preferred view, ordinary explanations of action are best seen as *teleological* explanations rather than causal explanations. Teleological explanations explain by citing a state of affairs or goal towards which the behavior is directed; such explanations need not cite any antecedent mental state of the agent at all, and even when they

7 Donald Davidson, »Actions, Reasons, and Causes,« in: *The Journal of Philosophy* 60 (1963), 685–700.

do, the point of citing the mental state is to identify the state of affairs towards which the behavior was directed and perhaps to say something about why that state of affairs would be of value to the agent.

In the last three of the examples above, no mental state is explicitly mentioned, and this is hardly uncommon in ordinary action explanation. On the causal construal, the explanations must implicitly refer to some mental state of the agent and cite it as the cause of the action; in some cases which mental state to posit would be more obvious than in others. By contrast, on the teleological construal, the explanations are not aiming to answer the question of the antecedent cause of the behavior; rather, the explanations are answering a different question: what was the behavior aimed at accomplishing? Towards what state of affairs was the behavior directed? Putting these four explanations into more explicitly teleological form would yield the following:

(1a) Steve went to the coffee shop in order to be able to work without interruptions from students.
(2a) Gina ordered decaf so that she wouldn't be kept awake that night.
(3a) Hayden shucked corn in order to fulfill his father's request that he do so.
(4a) Josephine went upstairs in order to get her shoes.

The causalist will presumably not dispute that (1)–(4) can be read as (1a)–(4a) (especially in the case of (2), since I did not change it at all). But the causalist will claim that even such explanations, which are teleological in form, are nonetheless *reducible* to causal explanations. I have argued elsewhere[8] that the causalists are wrong about this, and that attempts to reduce teleological explanations fail, and thus I will not here defend the irreducibility of teleological explanation. However, because it is relevant to the free will debate, I do need to address the question of how we make and justify teleological explanations: what considerations do we draw on when determining the truth or falsity of candidate teleological explanations?

IV. How we make teleological explanations

First, just a bit about explanation in the natural sciences. When we explain events in the natural world, of course we aim for a theory that is consistent with the observational data. But as has been oft noted, observational data alone scarcely suffice to constrain our theories. All manner of crazy theories can be made consistent with the specific observations we have. My favorite is Bertrand Russell's

8 See Scott Sehon, »Connectionism and the Causal Theory of Action Explanation,« in: *Philosophical Psychology* 11 (1998), 511–531; Scott Sehon, *Teleological Realism. Mind, Agency, and Explanation*, Cambridge MA 2005, chapters 6 and 7.

example: a theory according to which the entire universe simply popped into existence five minutes ago, with everything in place just as we naively think things were at that time.[9] We rule out such craziness, and all manner of less obviously crazy theories, by appealing to theoretical norms that go beyond consistency with the data. In particular, I take it that both in common sense and natural science we assume something like the following general principle, labeled »(S)« for »simplicity«:

(S) Given two theories, it is unreasonable to believe one that leaves significantly more unexplained mysteries.

The five-minute theory, along with other faulty theories, fails precisely because it leaves so many coincidences or mysteries utterly unexplained.

When we are interpreting *agents*, we likewise aim to be consistent with observational data, and we construct theories in accord with (S). But, I claim, we also do something different. Loosely following Davidson's views of interpretation (but not his endorsement of the causal theory of action), I suggest that we arrive at teleological explanations as part of an overall attempt to construct a theory of an agent, and part of our aim is to produce a theory according to which the agent is as rational as possible. In general terms, I suggest that our theorizing about agents is constrained by something like the following principle.

(R) Given two theories of an agent, it is unreasonable to believe one according to which the agent is significantly less rational.

Rationality can be assessed in various different ways, of course, but two aspects are particularly relevant here. First, we assume that,

(R_1) Agents act in ways that are appropriate for achieving their goals, given the agent's circumstances, epistemic situation, and intentional states.

But not just any state of affairs can count as an intelligible goal for an agent. We assume that,

(R_2) Agents have goals that are of value, given the agent's circumstances, epistemic situation, and intentional states.

So, roughly put, there are two axes on which a candidate explanation is judged: the degree to which it makes the behavior appropriate for achieving the goal, and the degree to which the goal is of value.

---

9 Bertrand Russell, *The Analysis of Mind*, London 1921.

In simple and straightforward cases, application of these principles is almost entirely automatic. Recall Josephine who went upstairs, and suppose that the circumstances were these: another ten year old friend had shown up at the door and asks Josephine wants to come outside to play; Josephine starts to run excitedly outside, but her father says, »wait! you need shoes! they're upstairs«. And off she goes up the stairs. And here we can see that it is quite obvious that going upstairs would be appropriate to achieving the goal of getting her shoes, and we can easily understand the value that this would have for Josephine. We might also believe various counterfactual conditionals that point in the same direction: If Josephine had believed that her shoes were in the kitchen, she would have gone there instead of going upstairs; had she believed that she already had shoes on, which perhaps she did prior to her father's admonition, she would have simply gone on outside; etc. The general point: our theory of Josephine is constructed so as to make the most rational sense we can out of her behavior in the actual and in nearby counterfactual circumstances.

Of course, further information might lead us to reject the initial hypothesis. Josephine might return downstairs entirely shoeless but carrying a favorite toy, and when her father reminds her that he told her to get her shoes, she might say that she didn't hear him. So we revise our original explanation and conclude that she went to get the toy rather than her shoes. But this is still in accord with the rationalizing principles – it's merely that we have gained further knowledge of her intentional states, and we are giving a rationalizing theory about a broader set of data. Naturally, we must also make our theory of Josephine conform with the simplicity principle, (S). Apart from our allegiance to (S), nothing would stop us from concluding that Josephine *did* get her shoes upstairs, but that they magically transformed into the toy on the way downstairs.

However, if we were *only* concerned with observational consistency and simplicity, and if we were not concerned with making rational sense of the behavior, all sorts of explanations would be possible and fully consistent with the data and with (S): e. g., that Josephine went up the stairs to get to France, or that she went upstairs hoping thereby to become Pope. If we are willing to attribute crazy enough beliefs and desires, any interpretation becomes possible. This is, I take it, exactly why Davidson thinks that Quine's radical interpreter can get nowhere at all without assuming a principle of charity. We rule out such interpretations precisely because they fail in our common sense psychological aim of making *sense* of the person with whom we are dealing.

Naturally, these rationalizing principles do not constrain our theorizing about the behavior of inanimate things like rocks or planets. Or, to put it the other way around, on any theory according to which a rock was an agent, the rock would either come out as quite irrational, or would have too impoverished a set of goals to count as a genuine agent. If we attribute to the rock one and only one desire, the desire to follow the laws of physics, then of course the rock comes out as

always acting in ways appropriate to its one goal. But it is not clear why this one goal would be of value to a rock or anything else. Moreover, I claim that being an agent requires a complex set of goals – a life. We cannot successfully attribute anything of the sort to the rock. So we conclude that the rock is not an agent at all.

Similarly, not everything that a human body does counts as goal-directed behavior. If someone tells an embarrassing story about me and I blush, my blushing is not goal-directed. Having my face turn red typically serves no goal of value to me. And even if turning red did serve some purpose I valued, I would equally have blushed in circumstances that were otherwise identical but in which I did not value having my face turn red. That is to say, it is not enough that a piece of behavior happens to serve an interest of mine in the actual circumstances; if there are nearby counterfactual situations in which the behavior did not serve any such goal, and if I would have done the same thing in those situations, then we would conclude that the behavior was not goal directed. So, in determining whether a given theory makes an agent come out as rational, we look both at the actual situation and at what the agent would have done in slightly different circumstances. And if, for a given piece of behavior, we cannot tell a rationalizing story, then we conclude that the behavior was not a goal-directed action.

Of course in doing all of this, we will take note of the agent's mental states, but we are not in the game of determining which of the physical antecedent events served as the cause of the agent's behavior. We could do that, of course, and we would presumably find various physiological states playing key causal roles. But, in our common sense psychological project of making sense of the agent, we are seeking a theory of the agent's behaviors and thoughts that is consistent with the data and is best in accord with principle (R). We are not, I claim, even assuming that the mental states of the agent can be identified neatly with the sort of physical states that are eligible to be causes of the behaviors in question. We are up to something different.

V. Rationalizability and freedom

Some behaviors (e. g., Josephine getting her shoes) are eminently rationalizable, and we have little trouble determining the goal towards which they were directed. Others (my blushing, someone falling to the ground after being pushed) are scarcely rationalizable at all, and we put them into the category of behaviors that are not goal-directed. Nonetheless, it is not a case of binary opposition, in which behaviors are either rationalizable or they are not. Rationalizability comes in degrees; even the best among us are prone to less than perfectly rational behavior. On the teleological account, we aim for a theory that makes as much rational sense of the agent as possible, just as the simplicity principle requires that we find the simplest theory of the physical world that we can manage; in neither case do

the principles guarantee that we will be able to find either a perfectly rational or perfectly simple story. Since teleological explicability thus comes in degrees, the degree to which a behavior counts as goal-directed or on purpose comes in degrees.

Now we can turn to the free will debate. I will start by suggesting that free actions, if there are any, are those for which we are legitimately held responsible. Thus if an action is not free, then I am not responsible for it, and if I am responsible for it, then it is a free action. On some views, freely performed actions are a small subset of the behaviors for which we are responsible; i. e., on such views, there are many routine behaviors for which we are fully responsible, but which are nonetheless not metaphysically *free*.[10] While I cannot do justice to them here, I'll just say that such views make me lose my grip on the meaning and significance of the notion of free will. Freedom *matters* to us because we genuinely hold ourselves and others responsible for our actions. We indeed diminish that responsibility on some occasions (weakness of will, addiction, brainwashing, or the like), and we do so on the grounds that the agent did not seem fully free.

For which behaviors are we responsible? At least to a first approximation, it is those things we do on *purpose*, for *reasons*. And those are the goal-directed actions, i. e., the behaviors that are teleologically explicable. In other words, I am suggesting that there is a tightly knit circle of concepts: behaviors for which we are responsible, freely performed actions, and goal-directed or teleologically explicable actions. As we saw above, we make teleological explanations by determining which goal state, if any, makes the most rational sense of the behavior in question. Thus a behavior is free if it is rationalizable in accord with the principles that govern teleological explanation.

So, for example, when Josephine went upstairs to get her shoes, her behavior was well suited to her goals, and her goals were of easily comprehensible value; so, other things being equal, she counts as quite free. On the other hand, when I blushed upon the retelling of the embarrassing story, the blushing was not rationalizable. However, rationalizability comes in degrees. Thus, whether a behavior is teleologically explicable comes in degrees. Accordingly, on this account, whether or not a behavior is free comes in degrees as well. So, more specifically, a behavior is free to the extent that the agent is rationalizable at the time.

On this view, free will is, I take it, fully compatible with determinism. Whether and to what degree an agent is free is a matter of whether her behavior is teleologically explicable, and, I have claimed, teleological explanation does not reduce to causal explanation. If this is right, then the causal history of a behavior and its status as an action would seem to be independent issues. Whether a physical

---

10 See John Martin Fischer/Mark Ravizza, *Responsibility and Control*, Cambridge 1998; Laura Waddell Ekstrom, *Free Will. A Philosophical Study*, Boulder 2000; Nomy Arpaly, *Merit, Meaning, and Human Bondage. An Essay on Free Will*, Princeton NJ 2006.

behavior is determined by antecedent causes is one question; whether it is teleologically explicable is a different question. We answer the teleological question by coming up with the best theory of the agent that we can manage, where the theory must be consistent with the data but where it also must make coherent sense of the agent. This project is, on its face, different from the project of identifying causes and determining their nature. So if I am right about the irreducibility claim, then there is no obvious way that the incompatibilist can argue that determinism somehow shows that, really, all behaviors fail to be in the free action category. Teleological realism makes determinism irrelevant to agency and freedom.

VI. Diagnosis of the cases

Now that we have the outlines of the positive account in hand, let's return to diagnosing particular cases, especially those where I've suggested that freedom seems to come in degrees. But I'll start with a case where the agent's freedom is not typically thought to be impaired, but which nonetheless nicely illustrates the view I am defending. Martin Luther was famously brought before the Diet of Worms in 1521, and was asked to recant certain propositions he had written, propositions that had been condemned by Pope Leo X.[11] After contemplating this request, Luther is reported to have refused: »To go against conscience is neither right nor safe. I cannot, and I will not recant. Here I stand. I can do no other«.

If we take Luther's words seriously, then we would conclude that he could not have done otherwise than he did. On certain incompatibilist views of freedom, this would show that Luther was not free, which strikes most of us as counterintuitive, a point emphasized by Daniel Dennett.[12] Far from being a case of weakness of will, Luther was acting with full conviction and in accord with his best judgment. If he was being compelled, he was being compelled by what he at least thought was right reason. On the teleological account, his behavior seems eminently rationalizable, and hence it counts as a goal-directed action, a freely performed behavior for which he is responsible.

We don't know for sure whether Luther actually said, »I can do no other,« and, in any event, we might take that claim to have been hyperbole. But the general point is familiar enough. Suppose that I am walking into the voting booth, having firmly made up my mind to vote for the Democrat; I believe that she is eminently qualified, her positions on the issues are similar to mine, and, besides, I think

11 See »Diet of Worms,« *Encyclopædia Britannica* 2008. *Encyclopædia Britannica Online*. 2 May 2008, http://search.eb.com/eb/article-9077510.

12 See Daniel Dennett, *Elbow Room. The Varieities of Free Will Worth Wanting*, Cambridge MA 1984.

that the Republican candidate is dangerously wrong on most of the issues, and would be a disaster in office. All of my reasons point towards voting for the Democrat, so that's what I do. There is a strong sense in which I can do no other. Voting for the Republican would make no rational sense, given my set of beliefs and desires. Of course, I could imagine my arm twitching involuntarily and thereby marking a vote for the Republican, and I can even (dimly) imagine suddenly reevaluating my beliefs on the spot. But, keeping my beliefs and desires the same as they were at that moment, it is hard to see myself *intentionally* voting for the Republican. (Nor would I *want* it to be the case that I could have done otherwise than I did; i. e., it is hard for me to see the value of freedom as the libertarian defines it.) On the other hand, I take full responsibility for my vote and see it as a paradigm case of a free action, precisely because it is so clearly something I did on purpose – a goal-directed behavior. So, again, this sort of case fits neatly with teleological account of free will, but it is much more problematic on a view that identifies freedom with some sort of ability to do otherwise.

By contrast with the case of Luther, the characteristic feature of weakness of will is that we act against our own best judgment, though this can be complicated by the fact that our judgment itself might change at the moment of temptation. Indeed, I think it is useful to distinguish between what we might call our *hot* rationality versus our *cool* rationality. Our hot rationality is what seems rational to us in the heat of the moment, at the moment where some urge (whether for sweets, alcohol, nicotine, sex, etc.) is felt very strongly and when a fairly simple momentary action could put us directly on the path of satisfying it. Our cool rationality is what we would want ourselves to choose, when judged from the vantage point of circumstances in which the specific urge is not felt so strongly or where, even if it is felt reasonably strongly, we don't have the path of immediate gratification available to us.

So at one extreme, we can imagine the drug addict who, literally shaking with desire for another dose, succumbs to the temptation, despite having told herself many times in cool moments that she must quit. At the other end, we can imagine a fairly healthy person who is slightly overweight, and would like to lose 10 pounds or so, and thus has decided that he will forego desserts. But at a special dinner, he is presented with an exquisitely prepared fruit tart. Like the drug addict, he might give in to the temptation, but the case is quite different. It is a special occasion, the stakes are not that high, and the dessert is not that unhealthy; that is to say, even if he were to judge the situation from the vantage point of his cool rational self, it might be a close call. So while it still might be a case of weak willed action, it has little of the tinge of desperation had by the drug case. In between, there are all manner of other sorts of cases. Many of us are subject to cravings of various sorts and degrees. In the heat of the moment, all such desires can impair our judgment, as compared with what we would think in cooler moments.

So are cases of weakness of will rationalizable? Yes, to varying degrees. In each of these cases, when the agent does act and succumbs to whatever temptation is in play, the action is done in order to satisfy the urge in question. Satisfying an urge of one sort or another will typically be an intelligible goal. Thus the behavior will be teleologically explicable, and thus count as within the realm of free actions. However, depending on the nature of the action, while the agent's behavior will be intelligible, it may be far from perfectly reasonable. We may understand that the agent acted in order to satisfy the urge, but it will also be true that there was another action available to the agent that would have been more rational overall, an action serving a goal of greater value, even given the agent's own intentional states and epistemic circumstances. Moreover, if the agent's beliefs and desires themselves are far from rational, then this too will diminish the agent's overall rationalizability.

In minor cases of weakness of will, like the dessert case, resisting the temptation would be more in line with the agent's desires and values, but it might be a close call. The action of taking the offered fruit tart is easily rationalizable, even if there was an alternative that would have been somewhat more in line with what the agent valued. On the other hand, the morbidly obese person with the defective leptin receptors knows that it is very much in her best interests to leave the bag of cookies alone, and she knows she will regret it later, but she acts instead to satisfy the ravenous hunger she feels at the moment. Moreover, given the hunger she feels, she would have eaten the cookies almost irrespective of how much she had eaten that day already and how bad the cookies were for her. So, although satisfying hunger makes some rational sense, the fact that she ate the cookies despite her intentions, along with the fact that she would have eaten the cookies almost no matter what, shows that her behavior was markedly insensitive to what was of most value to her at the time. Thus, while we can tell a rationalizing story when she in fact eats the cookies, the story is less good than if she had done something else. On the suggested view of freedom, this means that her weak-willed action, while still an action, is less rationalizable and considerably less free than paradigm actions.

Cases of extreme addiction or compulsion will be even harder to rationalize. Consider a drug addict who is about to inject something that may induce a euphoric state, but which has negative long-term consequences, and even in the relatively short term will likely render the addict unable to go to work that morning, and will probably immediately cost him his job. The addict knows that all of his interests, except that of getting high at that moment, speak strongly against taking the drug, but, feeling that he can't help it, does so anyway. We can build up the case more by stipulating that the agent would have taken the drug in the face of almost any positive reason not to do so. We can see that he took the drug to satisfy the immediate urge, or for the euphoric state he hoped to induce, so it is minimally rationalizable, but both in the actual circumstances and in nearby

counterfactual circumstances, the agent would be ultimately far better served by other actions. Accordingly, on the present account of freedom, we conclude that the addicted agent is only minimally free.

What if we are dealing with a willing drug addict? For example a cigarette smoker with extremely strong nicotine cravings, but who positively affirms her smoking habit and has no desire to give it up? On standard sorts of incompatibilist views, the unwilling drug addict and the willing addict are in exactly the same position, and it all depends on whether or not it was physically possible for them to do otherwise. Given the incompatibilist position, if determinism is true then we are all unfree, including the addicts. If determinism is not true, then whether this particular action was free depends on the physical possibility of doing otherwise. *Maybe* we can conclude that it is physically impossible for either addict to act otherwise, and conclude this because in cases of severe addiction it *feels* like one has no choice. But it is a real question whether this feeling is indicative of a genuine physical impossibility. In any event, the incompatibilist will probably put the willing and unwilling addicts into the same category, and that would probably be the unfree category. Some compatibilists, on the other hand, might distinguish between the two cases. Harry Frankfurt[13] says that the unwilling addict is not free, because his actions are not in accord with his second order desire to stop taking the drug. The willing addict, however, is different, for his second order desire is to continue using the drug, and Frankfurt asserts that he does act of his own free will.

On the teleological account of freedom, there is no need to shoehorn the addicts into the »free« box or the »unfree« box. The unwilling addict, as suggested just above, is only minimally free, and is less free to the extent that her drug use goes against her interests and violates her own judgment about what is best for her to do. What of the willing addict? Obviously, she is not acting contrary to her own judgment of what she should do, and thus her values and beliefs are at least, to that extent, more internally coherent and thus more rationalizable than those of the unwilling addict. So she is more free than the comparable unwilling addict, and this seems right: we will generally hold someone more responsible if she affirms and relishes her drug use than if she despises it and struggles against it. But the drug user does not become fully free simply by having the conviction that she likes using drugs. Where she falls along the spectrum of freedom and unfreedom will depend on other details. For example, some coffee drinkers report feeling quite addicted to coffee in the morning and will go to some trouble to obtain coffee if they find themselves out of it at home, or if they are staying in the home of a non-coffee drinker. Nonetheless, they might fully endorse their habit.

---

13 See Harry Frankfurt, »Freedom of the Will and the Concept of a Person,« in: *Journal of Philosophy* 68 (1971), 5–20.

At the other extreme, there could be a heroin addict whose life revolves around getting her next dose, who has lost her friends, family, and job, and whose life is on the brink of collapse; but she might nonetheless endorse her heroin use. While the behavior of the willing heroin addict is somewhat more rationalizable than that of the unwilling heroin addict in analogous circumstances, the willing heroin addict is still very different from the willing caffeine addict. Her heroin use is at odds with all sorts of other things that she *ought* to value in life, even if she is currently blind to seeing those values. While we can understand the attraction of the euphoric high said to go with use of the drug, we conclude, with reason, that the heroin addict's life is far from ideal, and we suspect that the addict herself would see this too if she had some appropriate distance from her use of the drug. Moreover, we might suspect that in nearby counterfactual circumstances in which she no longer affirms the value of the drug use, she will likely still behave the same way. All of this paints a picture of an agent whose actions and life are hard to fully rationalize, and thus we would take it that a severely, even if willingly, addicted heroin user will not be very far towards the free end of the free will spectrum. Little of this applies to our willing coffee addict, for needing a couple of cups of coffee in the morning typically has very few detrimental effects in one's life generally. In addition, were it to be shown to the typical coffee drinker that her coffee drinking was having seriously deleterious effects, she would likely cut back or stop. So the willing coffee addict can, so far as her coffee drinking is concerned, be firmly on the free side of the spectrum.

The examples I've discussed indicate how weakness of will can fall in a spectrum from fairly mild cases to extreme addiction, and that the teleological account of free will classifies them accordingly: in mild cases of weakness of will, we are still substantially rationalizable and thus substantially free and responsible, but in more extreme cases of addiction, agents are much less free. It is worth noting, however, that even a heroin addict is markedly different from, say, someone who collapses in a heap upon being struck by lightning, for the latter person would have collapsed regardless of what she valued on the occasion and was not even satisfying some sort of urge in the process. Even if she did have the urge to collapse right before doing so, in face of the lightning strike, she would have collapsed even in those counterfactual circumstances in which she did not have that urge; whereas if the drug addict's urge disappeared, he would not take the drug anyway.

Weakness of will holds an odd place in the philosophical literature. Some authors seem forced by their philosophical positions to deny, or all but deny, that weakness of will is even possible.[14] In the free will literature, some compatibilists are forced to say that any case of weakness of will is therefore unfree. On Frank-

14 George Frederick Schueler, Reasons and Purposes. Human Rationality and the Teleological Explanation of Action, Oxford 2003.

furt[15] type views, for example, whether our will is free depends on whether we act in accord with our second order volition – i. e., whether the desire we wish to act on is the one we in fact act on. By definition, we exhibit weakness of will when we act contrary to our own best judgment, so Frankfurt should count all such actions as simply unfree. Similarly, on Gary Watson's view, we are free when our desires and our values are in harmony, which is precisely what is not happening in cases of weakness of will; thus, again, cases of weakness of will are automatically unfree.[16] Surely this is a counterintuitive result, especially in minor instances like occasional overeating or watching mindless television rather than washing the dishes. On the other hand, I find it equally counterintuitive to claim that severe cases of weakness of will are simply and straightforwardly free actions. The common sense thing to say is that weak willed actions fall somewhere in the middle, and that it depends on the strength of the temptation and the level of irrationality involved. My view accomplishes this.

VII. Objections

I count agents as less and less free the further they go from being rationalizable. To some minds, this might make my view look like a non-starter. By appearing to suggest that non-rationalizable agents are not free, I seem to be putting reason and cool-headedness into a position of ridiculously high privilege. We might put this objection in one of two ways. First, it might seem that on my view the only completely free individuals are those who approximate Mr. Spock, of the 60s television series *Star Trek*, namely individuals who are coldly logical and fight their emotional impulses. Second, it seems clear that we can freely choose actions that are wrong, unjustified, or even stupid, and it might appear that my view automatically makes such actions less than free. I'll deal with each of these thoughts separately.

a. Objection: Makes Spock-like creatures the only truly free agents

A Spock-like person, who calmly and logically considers each action, may indeed be rationalizable, and thus count as free. But it is a mistake to think that, in general, one's behavior is more rationalizable to the extent that one shuns emotional reactions. There can be times when the most reasonable thing a person can do is to act against her own best logical judgment. Consider Huck Finn in the climactic moment of *The Adventures of Huckleberry Finn.* Huck has just written a note back to Miss Watson giving the location of her runaway slave, Jim. After writing

15 See Frankfurt, *Freedom of the Will and the Concept of a Person* (fn. 13).

16 George Watson »Free Agency,« *The Journal of Philosophy* 72 (1975), 205–220.

the note he says, »I felt good and all washed clean of sin for the first time I had ever felt so in my life.« But as he reflects on the time he has spent with Jim, he reconsiders:

> It was a close place. I took [the note] up, and held it in my hand. I was a trembling, because I'd got to decide, forever, betwixt two things, and I knowed it. I studied it a minute, sort of holding my breath, and then says to myself:
> »All right, then, I'll *go* to hell« – and tore it up.[17]

Huck acts against his own best judgment, believing that he will even suffer eternal damnation for his action. But in fact it was the right thing to do, and we have no trouble in seeing the value in his action, even if Huck himself is convinced otherwise, and thinks that he has just been incredibly weak and has committed a grave sin.[18] Huck has been taught and is operating within a system of values and beliefs according to which black people are mere property; Huck appears incapable of consciously thinking his way out of that system; he can't bring himself to consciously affirm that it is some of the values he has been taught that are gravely mistaken. But on another level, he knows this, and when he acts he goes with his emotionally laden instincts rather than his conscious judgment. And this choice, besides being the admirable and right one, ends up being at least as free as the alternative of turning Jim in. Huck would not have been more rationalizable as an agent had he taken a Spock-like approach, ignored his emotions and thought only about what seemed most logical.

In some ways, Huck seems like a special case, since the issues are so weighty, and Twain is so adept at portraying the inner torment of the boy whose emotional perceptions fly in the face of societally imposed values. But the point applies to more trivial contexts as well. One might impulsively decide to stop working on a philosophy paper, grab the kids and head to the beach; or one might spontaneously put the housecleaning on hold and go off to a bar with a friend. Such actions are quite rationalizable. Indeed, even if these precipitate actions are contrary to the agent's carefully considered plans for the day, the actions might nonetheless be at least as valuable as the more deliberate alternative. It might even be important to the value of the actions that the agent *felt* as if she was playing hookey. Being rational, in the broad sense I have in mind, does not necessarily mean always carefully planning and weighing options logically. It means instead pursuing courses of action that are of value from the agent's perspective, and actions with this feature are not necessarily coextensive with actions that are carefully planned. Perhaps if we were Godlike in our ability to deliberate and plan, things would be different, for if we were always perfectly rational thinkers and

17 Mark Twain, *Adventures of Huckleberry Finn*, 1884, chapter XXXI.

18 Susan Wolf makes a similar point about this example (see *Freedom Within Reason*, New York 1990).

planners, then it would presumably be irrational ever to spontaneously discard our plans and act on whim. But since we are not perfectly rational when deliberately making plans, it can happen that our emotionally informed impulses or whims are, on occasion, more trustworthy and more reasonable overall than our more consciously employed rational faculties.

Antonio Damasio describes a famous case that also illustrates this point.[19] Phineas Gage, a railroad worker in the mid 19th century, suffered a terrible accident in which a large metal bar was driven completely through the front part of his skull. Astonishingly, Gage survived, and his speech, memory, and intelligence seemed intact. However, Gage was far from himself. His ability to interact socially, and his ability to make intelligent decisions, were both greatly impaired. Damasio also describes a contemporary case of a man, whom he calls Elliot, with somewhat similar symptoms. Elliot had had a large brain tumor removed, and there had been damage to frontal lobe tissue. After the surgery, Elliot's intellectual abilities, as measured by a large variety of tests, were still intact. Nonetheless, Elliot could no longer hold down a job, and was constantly making decisions with detrimental consequences. Damasio observes that Elliot seemed virtually devoid of emotional reactions, either when telling his own sad history, or when being shown images of natural disasters and the like. Damasio says, »We might summarize Elliot's predicament as *to know but not to feel*«[20]. Damasio concludes that Elliot's defect in decision making and social behavior was connected to this emotional deficit:

> I was certain that in Elliot the defect was accompanied by a reduction in emotional reactivity and feeling […] I began to think that the cold-bloodedness of Elliot's reasoning prevented him from assigning different values to different options, and made his decision-making landscape hopelessly flat.[21]

In ordinary life, especially in social interactions, we operate with the help of emotional cues. We don't even consider certain courses of action because they seem literally laughable; we reject other possible plans because we recoil upon considering the effect the plan would have on someone else; we might perceive that a friend is in no mood for idle chit-chat, even if we cannot pinpoint or verbalize the behavioral cues from which we draw this conclusion.[22] For our behavior to

19 Antonio Damasio, *Descartes' Error. Emotion, Reason, and the Human Brain*, New York 1994.

20 Damasio, *Descartes' Error* (fn. 19), 45.

21 Damasio, *Descartes' Error* (fn. 19), 51.

22 For insightful discussion of examples of this sort, see Nomy Arpaly, *Unprincipled Virtue. An Inquiry into Moral Agency*, Oxford 2003.

be fully rationalizable, we need what Elliot and Phineas Gage apparently lacked, namely, the right sort of emotional reactions. So, to be rationalizable is not to be narrowly logical in a Spock-like way. Paying attention to emotions and instinctive reactions is very much a part of being a rational agent, and is crucial to being able to regularly pursue goals that are of value.

b. Objection: choosing badly becomes impossible

Surely sometimes we can, with full freedom, *choose* to take a less than rational course of action. But on my proposed view, it seems that the very fact that I choose to do something dumb (or immoral) is, *ipso facto,* a reason for exonerating me to some extent. Or, to put it slightly differently, the more it is true that I should have known better, the more my action *thereby* becomes excusable. This would be highly problematic. If my view denies that we can freely choose to do something less than rational, then the view must be mistaken.

To address this objection, it will help to consider specific examples of someone choosing an action that has less value than some alternative. Consider a case where someone chooses an action that seems fairly clearly immoral. Suppose that Jake, the owner of a small factory, has two choices: he can either pollute the local river with effluent from his plant, or he can install anti-pollution devices. If he installs the devices, his profit margin will decrease somewhat, but will still be substantial (and we can stipulate that all of the extra income will simply go to him, not to workers or shareholders). If Jake fails to install the devices, the resulting pollution will have various bad consequences, including greatly damaging the livelihood of poor fishing families who live downstream. Jake decides not to install the devices, knowing that this is the morally wrong thing to do. Doing the morally wrong thing can be considered a failure of rationality, at least in the broad sense I have in mind, for the agent is choosing a course of action that is of less value than the morally correct course. (There are tricky issues involved here; it makes a difference, for example, if moral values override other values. I need not decide such issues. The objection to my view depends, in this instance, on it being the case that Jake deliberately decided to choose an action that is of less value than others available to him, and I won't question that description of this case.) In terms of my view, the situation can be set up as follows. Before Jake's decision, we assumed that one of two explanations would ultimately be appropriate:

(1) Jake failed to install the anti-pollution devices in order to maximize his profits.
(2) Jake installed the anti-pollution devices in order to protect the environment in general and the downstream families in particular.

Once Jake has made his decision and fails to install the devices, the second explanation is, of course, ruled out. But comparing (1) and (2) becomes relevant when we rate Jake's overall rationality. If Jake *would have been* more rational if he were such that (2) was true, then this means that the truth of (1) makes him less rational than he otherwise might have been.

To illustrate, compare this with a case of weakness of will. Suppose that there is a different agent, and one of two explanations might become true of him tonight:

(3) Scott spent the evening watching television and eating potato chips for the pleasure of salt, fat, and mindless entertainment.
(4) Scott spent the evening writing in order to finish his free will paper

If in fact Scott spent the evening watching television, then (4) cannot be the correct explanation of his behavior. But in rating the rationality of Scott's action, the alternative of (4) becomes relevant: Scott would have been more rational had he taken that course. But note that, if (4) were true, Scott would be more rational, in part, because his action would match his own best judgment about what he should do. As seen above, it is not always irrational to act contrary to one's best judgment, but such actions automatically involve *some* tension; acting against one's best judgment at least shows some failure of coherence between behavior and values. That is part of why akratic actions are less rationalizable.

But the factory owner, Jake, is not acting against his own best judgment. We are supposing that Jake made the calm, cool judgment that he didn't really care about long-term environmental consequences or about the fishing families living downstream; he just wanted to maximize his own already substantial profits. To those who know Jake, this comes as no surprise, for he has always been rather self-centered and disdainful of the interests of others around him. If Jake had suddenly chosen to pay for the anti-pollution devices, then, at least in one respect, he would have been *less* rational, for he would have been acting contrary to his own beliefs, desires, and judgments. In other words, there are different ways in which an action can be of value. In terms of objective, overall value, it would be better if Jake put the anti-pollution devices on. But, as Jake sees things, it is better for him to make more money. Recall that we are trying to tell as rational a story as we can about the agent, given the agent's circumstances and intentional states. This last proviso means that self-consistency and coherence make for a more rationalizable agent.

This is not to say that one's intentional states can rationalize any behavior or that Jake is perfectly rational so long as he is acting in accord with what *he* values. Jake's rationality is still diminished by his mistaken judgment that he should maximize his own profits at the expense of the fishing families trying to eke out a living. But the mistake itself is comprehensible, for people are often and somewhat understandably blinded by their own self-interest. Jake's behavior makes sense given his existing values, beliefs, and desires; and his intentional states form

a reasonably consistent and coherent whole. Still, the very fact that we believe him to be making a *mistake* in his value judgments does mean that there is a sense in which we believe that he could be yet more rational, and in this same sense he fails to be perfectly rationalizable.

But before we conclude that Jake is less than a free agent, we should consider a couple of things. First, if we demand a status of perfect or near perfect rationality before we count an agent as free, then who among us will achieve that status? While the view is that freedom comes in degrees, and this suggests a spectrum from zero to a perfect 100 %, we needn't be mindlessly rigid and require 100 % rationalizability before we simply count an agent as free. There is an analogy here to degrees of confidence in our beliefs. We are more confident of some of our beliefs than of others, even if we would be hard pressed to attach an actual percentage degree confidence to individual beliefs. If we are clear headed and even remotely impressed with skeptical arguments, we should admit that there are few, if any, of our beliefs concerning which we should claim absolute 100 % confidence; but we still straightforwardly and correctly say that we believe certain propositions even if we are not completely and utterly certain of their truth. Similarly with freedom: we can straightforwardly and correctly say that an agent is free even if the agent is not perfectly and completely rationalizable.

The second point is that it may be misguided or incoherent to talk of a simple linear scale of degrees of rationality or rationalizability. There are various differing facets of rationality: the consistency and coherence of our beliefs, of our values, the correctness of our beliefs and values, and the degree to which our behaviors are in accord with what we value. And there are arguably distinct kinds of value: short-term prudential, long-term prudential, concern about friends and loved ones, moral values, etc. If there is a way of translating all of these factors into a linear scale of rationality, I will certainly not be attempting it. But there can still be a clear sense in which some actions and agents are more rationalizable than others, even if the standards are defeasible and open-textured, and even if there are epistemological problems in ascertaining correct answers in certain cases. So, even though Jake may not be perfectly rational, he is still free, and the very fact that he chooses the wrong action does not exonerate him from responsibility. Of course, the point generalizes: when we make poor choices, then this does show that we fail to be perfectly rational, but it does not thereby sink us into degrees of unfreedom that will begin to exonerate us from moral responsibility.

VIII. Conclusion

Descartes's claims in the *Meditations* notwithstanding, philosophical argument rarely proceeds by ironclad argument from supposedly indubitable premises to now-indisputable conclusions. In this paper, I have suggested that a plausible view

of action explanation, a view I've defended elsewhere, leads to an account of free will that is compatibilist and that accords nicely with the independently plausible claims that freedom comes in degrees and that cases of addiction and weakness of will fall along a spectrum. The theory is, I claim, an attractive package, particularly when compared to incompatibilist accounts that make our free will at any give moment hostage to strong and all but unverifiable assumptions about physical possibilities. Of course, incompatibilists have specific arguments for their position, and these need to be answered; but that will have to wait for another occasion.