

GIDEON YAFFE

Recent Work on Addiction and Responsible Agency

We tend to sympathize with addicts who behave illegally or immorally in service of their addictive cravings more readily than we do with those who act in exactly the same ways but who are not addicted. The addict who kills for money to buy crack seems less a moral monster than the unaddicted person who coldly plots the same murder for the same purpose. This distinction in our moral sentiments sometimes manifests itself in a distinction in legal and moral treatment: addicts are rarely thought blameless, but they are often taken to be less at fault than their unaddicted counterparts. But is the fact that a person's objectionable conduct springs from an addiction of genuine moral or legal weight? And, if it is, what is it about addiction that produces some form of diminished responsibility? In the last few years, a startling amount of literature relevant to these topics has appeared, produced by theorists in a wide variety of disciplines from jurisprudence, psychology and ethics to economics, political science and neurobiology. This essay critically examines some of the most prominent recent

This essay critically reviews issues and arguments raised in a number of recent books and articles on addiction and self control. Particular emphasis is placed on the following: George Ainslie, *Breakdown of Will* (Cambridge: Cambridge University Press, 2001), henceforth *BW*; Jon Elster, ed. *Addiction: Entries and Exits*, (New York: Russell Sage Foundation, 1999), henceforth *AEE*; Jon Elster and Ole-Jørgen Skog, eds. *Getting Hooked: Rationality and Addiction*, (Cambridge: Cambridge University Press, 1999), henceforth *GH*; Jon Elster, *Strong Feelings* (Cambridge: The MIT Press, 1999), henceforth *SF*; and two special issues of *Law and Philosophy* devoted to the topic ed. Michael Corrado: 18, no. 6 (1999), and 19, no. 1 (2000), henceforth *LP*.

Thanks to Michael Bratman, John Fischer, Janet Levin, Al Mele and the Editors of *Philosophy & Public Affairs* for invaluable comments and suggestions.

efforts to explain the impact, if any, of addiction on freedom and rationality, and, in turn, legal and moral responsibility.¹

There is something like consensus in the literature, and with good reason, that if addiction does diminish responsibility it is not for the reason that, say, epilepsy diminishes responsibility. The epileptic might do damage when in the fit of a seizure, but she is not responsible for that damage since her spasmodic movements are not motivated. She differs from a person who is thrown to the ground by the wind only in that the “wind” that blows her about springs from a condition within her own brain. But behavior stemming from addiction is not like this. The addict is motivated to get that to which she is addicted. As Gary Watson puts the point, “One who is defeated by appetite is more like a collaborationist than an unsuccessful freedom fighter.”² The first question is how, if at all, the motivational structures involved in addiction differ from those of the unaddicted; the second question is what difference, if any, this makes to responsibility for behavior stemming from addiction. While the bulk of the recent work on addiction is concerned with the first of these questions, the second will be considered here as well.

There is both a legitimate moral and legal basis for distinguishing among (1) those who wholeheartedly and unreservedly pursue illegal or immoral courses of action, (2) those who do wrong out of compulsion, that is, unfreely, and (3) those who do wrong as a result of transitory powerful impulses and thus manifest irrational weakness—crimes of passion, for instance. One approach to understanding the impact of addiction on responsibility places addicts in category (2). To adopt this approach is to say that addicts are subject to irresistible desires or are in some other way compelled to act as they do.³ This is to contrast

1. Recent discussions of policy issues about addiction are not discussed here, although there is very interesting work on the topic to be found. Cf. Helge Waal, “To Legalize or Not to Legalize: Is That the Question?” in *GH*, pp. 137–72; Douglas N. Husak, *Drugs and Rights* (Cambridge: Cambridge University Press, 1992); Husak, “Addiction and Criminal Liability” in *LP* 18 (1999): 655–84.

2. Gary Watson, “Disordered Appetites” in *AEE*, p. 7.

3. One naïve way to make good on the thought that addiction undermines freedom would be to argue for the claim that addicts are compelled to do what they do in something like the way in which a man falling from a bridge is compelled to hit the water. That is, we might think that addiction takes control of our bodies independently of our wills rather than inducing irresistible desires. In the popular imagination, this concep-

behavior stemming from addiction with behavior reflective of an agent's capacity to control what she does. We can imagine a variety of ways of pursuing this strategy differing with respect to their analyses of the freedom necessary for moral responsibility, on the one hand, and addictive behavior, on the other.⁴ But whatever the details of such accounts, adopting this approach amounts to claiming that addiction is a familiar excusing condition analogous to other conditions, such as insanity, that excuse from responsibility by removing the agent's capacity to engage in legally or morally responsible behavior. Those who take insanity to undermine freedom often argue that insanity removes its victim's capacity to act rationally, and further claim that such a capacity is required for the freedom necessary for moral or legal responsibility. We might adopt a similar position with respect to addiction. Still, to take this approach is to say that addiction undermines responsibility by eliminating freedom.

An alternative to the view that addiction eliminates freedom takes addiction to influence the agent either not to employ, or to misemploy, her capacities for rational conduct. This approach contrasts addictive

tion of the addict's behavior is encouraged by recent discoveries in neurobiology mapping the neurological effects of drug consumption. (For a useful survey of recent research of this sort see Eliot Gardner and James David, "The Neurobiology of Chemical Addiction" in *GH*, pp. 93–136.) But it is only when mind-body relations are understood very naïvely that such results are taken to indicate any such thing. After all, even deliberate action typical of free agency must have some kind of neurological basis. Besides, it runs directly contrary to the phenomenology of addiction to suggest that the cravings felt by addicts play no role in generating their behavior. For closely related remarks, and helpful discussion of the limitations of the disease model of addiction, see Stephen J. Morse, "Hooked on Hype: Addiction and Responsibility," *LP* 19 (2000): 3–49.

4. One of the best known ways of accounting for the diminished responsibility of addicts appears in Harry Frankfurt, "Freedom of the Will and the Concept of a Person" in *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988), pp. 11–25. Frankfurt's explanation depends on both a controversial conception of freedom and a controversial conception of addiction. Freedom of the sort that addiction undermines, according to Frankfurt, is enjoyed by an agent just in case the motivational efficacy of her first order desire depends upon her wanting that first order desire to be efficacious. Frankfurt takes the addicted agent to be such that her first order desire for that to which she is addicted will be effective regardless of whether or not she wants it to be. There is a substantial literature assessing both of these aspects of Frankfurt's position, the examination of which would take us too far afield. For helpful discussion and an overview, see Gary Watson, "Free Action and Free Will" in *Mind* 96 (1987): 145–72 (esp. pp. 147–53). See also Olav Gjelsvik, "Freedom of the Will and Addiction" in *AEE*, pp. 29–54.

behavior with the peculiarly repugnant behavior of a person who acts objectionably but whose conduct is the product of correctly functioning rational processes.⁵ Perhaps, that is, addictive behavior is as much under control as the behavior of the unaddicted, but the conduct of the addicted agent does not spring from the exercise, in service of something objectionable, of rational capacities, in the way that the most morally objectionable conduct does. To adopt this approach is to see addiction as irrational behavior—weak-willed behavior—performed by agents in possession of the capacity to act as they ought.⁶ Action expressive of weakness, it seems, is diminished in responsibility, rather than excused entirely from responsibility, and so this approach, in placing the addict in category (3), provides grounds, for instance, for lesser sentences for addicts, or for lesser moral censure than would be appropriately applied to an unaddicted agent who acts similarly.

According to a third, deflationary, approach, there is little reason to think that addiction diminishes responsibility at all; addicts, that is, fall under category (1). I begin, in Section I, with consideration of the grounds for advocating such a position provided by rational choice theory of the sort practiced by most economists.⁷ Under views of this

5. A certain sort of ethical rationalist will deny that there is any immoral behavior that is fully rational. But even ethical rationalists of this sort think that immoral behavior could be the product of processes that are “rational” in some sense of the term. Such rationalists, for instance, can distinguish between instrumentally or procedurally rational and irrational immoral conduct. So, the approach under discussion for accounting for the responsibility-undermining force of addiction is open to ethical rationalists.

6. Self-deception may play an important role in many addictions. The alcoholic may drink, for instance, believing that he must since it would be rude not to toast his host when, in fact, the host couldn't care less and he is really drinking to satisfy his craving. This essay doesn't discuss issues of cognitive irrationality but focuses, instead, on the way in which one can choose, or be motivated to choose, irrationally even while having rational beliefs. For an important recent discussion of self-deception see Alfred Mele, *Self-Deception Unmasked* (Princeton, NJ: Princeton University Press, 2000).

7. Another way of pursuing this deflationary approach starts with the thought that addicts are usually responsible for the fact that they are addicted, and so the fact that addictive behavior is irrational does not ameliorate the addict's responsibility. Such views encounter a variety of obstacles; perhaps the most important is this: People are very often excused from responsibility for behavior springing from conditions acquired voluntarily. The responsibility of a parent who takes objectionable steps to prevent separation from a child is diminished. But a parent's attachment to a child is no less voluntarily acquired than many addictions. For further discussion, see Section III.

sort, addiction influences action in something like the way in which poverty influences action. For those unlucky enough to be in poverty, it can be rational to commit crimes. But we don't ordinarily think that the poverty-stricken are thereby excused from responsibility, even if we feel empathy for their predicament. Perhaps an agent can find herself in circumstances such that, because she is rational, she ends up engaging regularly in destructively high levels of drug consumption. Section I focuses on the work of economist Gary Becker who argues that addictive behavior is a result of rational efforts to satisfy preferences for certain special goods given temporal constraints. It is argued that while illuminating in certain respects, such views are able to accomplish their aim of characterizing the addict as fully rational, and thus fully responsible, only by maintaining an implausibly thin conception of that which can be rationally assessed. Thus, rational choice models of addictive behavior tell us something important about addiction, but they don't tell us as much about the responsibility of addicts as we would like.

The failures of a pure rational choice view of addiction encourage approaches that exploit some of the tools of rational choice theory but, at the same time, deny that addicts are rational in the sense meant by rational choice theorists. Section II looks at the view of George Ainslie. Like Becker, Ainslie thinks that addiction is the outcome of rational efforts to satisfy preferences in the face of temporal constraints. However, Ainslie thinks that the way in which addicts respond to such constraints is irrational and indicative of weakness of will. Thus, Ainslie places addicts in category (3). Section II argues that Ainslie's account, as developed in his most recent work, coherently identifies a form of irrationality to which addicts are subject only by failing to make room for an adequate account of the relationship between the will and our capacities for rationality. What this implies, as we'll see, is that while Ainslie is right that a fully satisfactory account of addictive behavior must depart from the rational choice theorist's conception of the agent, the needed departure might be more radical than Ainslie envisions.

Section III examines the views of George Loewenstein, Jon Elster and Gary Watson, all of whom depart from the standard rational choice model's picture of the addict differently from the way Ainslie departs from it. Loewenstein, Elster and Watson question the rational

choice theorists's assumption that there can be no break at any particular time between what the agent judges to be best and what she is most motivated to pursue. All three of these theorists analyze addiction as a condition of susceptibility to certain special "visceral," or as Watson puts it, "appetitive" motives for action. There remains problematic ambiguity about the implications for responsibility of such views. It is unclear, that is, whether such views imply that addicts belong in category (2) or in category (3). Section III examines recent efforts to remove this ambiguity. While progress has been made on this front, there remains a great deal of work to do.

I. ADDICTION AND TRADITIONAL RATIONAL CHOICE THEORY

According to the influential economic model of addiction proposed by Becker, addicts rationally act so as to maximize their preferences at each and every moment.⁸ However, they differ from the unaddicted by virtue of the fact that (1) they engage in regular heavy consumption of a particular substance, where the unaddicted do not consume the substance at all, or only at much lower levels, and (2) the overall welfare of the low level consumers (the unaddicted) is significantly higher than the overall welfare of the high level consumers (the addicted). How could a rational agent end up in such a predicament? Becker demonstrates that this occurs when a substance that is pleasant to use induces tolerance and reinforces its own consumption, and the agent weighs the goods of the present much more heavily than those of the future.⁹

8. Becker's overall approach is expressed in his *The Economic Approach to Human Behavior* (Chicago: Chicago University Press, 1976). The approach is applied to addiction in Gary Becker and Kevin Murphy, "A Theory of Rational Addiction," *Journal of Political Economy* 96 (1998): 675–700. Also relevant is Gary Becker, Michael Grossman and Kevin Murphy, "Rational Addiction and the Effect of Price on Consumption" in *Choice Over Time* (henceforth *CT*), George Loewenstein and Jon Elster eds. (New York: Russell Sage Foundation, 1992).

9. Becker's derivation of this implication is summarized very nicely in Ole-Jørgen Skog, "Rationality, Irrationality and Addiction—Reflections on Becker and Murphy's Theory of Addiction" in *GH*, pp. 173–207. Skog's simplified presentation of Becker's position is an important contribution to the philosophical literature on addiction and rationality, since Becker's own presentation of his view relies on mathematical reasoning that few philosophers are able to follow. The presentation of Becker's position offered in the main text differs from Skog's only in style.

Becker analyzes tolerance and reinforcement as follows:

Tolerance: A substance induces tolerance if and only if the more the agent has consumed in the past the less utility she receives from a given level of present consumption.

Reinforcement: A substance reinforces its own consumption if and only if the more the agent has consumed in the past the more utility she will receive from an increase in consumption.

So, for instance, crack cocaine induces tolerance since the more hits one has taken, the more one needs to take to receive the same high. It also reinforces its own consumption, since by increasing consumption—say, by taking two hits today after taking only one yesterday—one increases one's pleasure *and* avoids the pain of withdrawal. The second hit makes more of a difference to the agent than it would have had she not consumed in the past since past consumption places the agent in a position of suffering withdrawal should she consume at the same or a lower level.

Notice that it is quite possible for a regular consumer of a substance that induces tolerance (putting aside, for a moment, reinforcement of consumption) to have a lower level of overall welfare than an agent who abstains entirely. Imagine, for instance, that the welfare level of an agent over the course of a day can be measured on a scale from -10 to 10 , with 10 being the best. And consider the ten day career of two agents one of whom uses and one of whom abstains from use of a substance that induces tolerance. Let's say that the Abstainer enjoys a level 6 day each day, and so has an overall welfare of $6 \times 10 = 60$ over the ten days. The User's overall welfare is more complicated to calculate. Since the substance is pleasurable to use, we can imagine that someone who has never used before can increase his welfare by $+4$ by using; so someone who has never used is choosing, the first time, between having a level 6 day in which he abstains and a level 10 day in which he uses. Since the substance induces tolerance, let's say that someone can increase his welfare by one point less through use for each day in the past in which he has used. If he has used only once in the past, he gets a boost of $+3$ through use; if twice, then a boost of $+2$ and so on. So, we have the following equation which I will call "The Welfare Through Use Equation":

[The total welfare enjoyed by someone who uses on a particular day] = [default welfare level] + [boost from use corrected for tolerance level]

In this equation, the “default welfare level” is the amount of welfare that someone who never used would experience on that day. So, on the first day, the User has a level 10 day (6 + 4). On the second day he has a level 9 day (6 + 3), on the third a level 8 day (6 + 2), and so on. After the fourth day, then, his welfare is reduced through use rather than increased, but if he continues to use, nonetheless, his overall welfare over the ten day period will be the sum of the integers from 10 to 1. This yields a total ten-day welfare level of 55: 5 points less than the person who abstained over the same ten-day period.

If the substance merely induced tolerance no rational agent would continue to use over a ten day period even if she made her decision about what to do on a given day without any regard for her welfare on future days. On the fifth day, the agent’s tolerance is such that using isn’t worth the cost. But, if we add in consideration of the fact that the substance also reinforces its own consumption, then on any given day someone who has used in the past might do better on that day by using. If the substance reinforces its own consumption, then were the agent not to use on a particular day, her expected welfare for that day would be reduced proportionally to the amount that she had used in the past. So, if we say, for instance, that her expected welfare were she to abstain is equal to the welfare of the person who never abstains minus the number of days she has used we derive the following equation, which I will call “The Welfare Through Abstention Equation”:

[The total welfare enjoyed by someone who abstains on a particular day] = [default welfare level] – [degree to which use is reinforced]

For someone choosing between use and abstention, use promises the degree of welfare specified by the Welfare Through Use Equation, and abstention promises the degree of welfare specified by the Welfare Through Abstention Equation. So, consider the position of someone who has never used before, and who chooses to use on each of ten days. Here the first number of the ordered pair is the level of welfare she can expect should she use and the second is the level of welfare she can expect should she abstain:

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
(10, 6)	(9, 5)	(8, 4)	(7, 3)	(6, 2)	(5, 1)	(4, 0)	(3, -1)	(2, -2)	(1, -3)

Since the first number of each ordered pair is higher than the second, an agent who chooses what to do on a given day only by considering her expected welfare on that day will use every day of the ten and will end up with a net overall welfare score lower than that of the person who abstains on all ten days.

How is a rational agent to avoid this unattractive result? Clearly, it is important to consider more than just one's welfare on a particular day in deciding what to do. Rational agents also take the future into consideration. So, in our example, a rational agent who weighs the future with exactly the same strength as the present will abstain on each of the ten days, for she will see that through using she enjoys only a temporary increase in welfare and pays the price later.¹⁰

Becker has demonstrated that whether or not a rational agent is an addict in his sense—that is, whether or not one finds oneself in a lifestyle of high usage of a substance that induces tolerance and reinforces its own consumption and where a high usage lifestyle promises a lower level of welfare than a low usage lifestyle—depends upon the degree to which the agent discounts the future. We have seen how this happens in the extreme cases—a rational agent who weighs only the present in her deliberations ends up an addict and a rational agent who weighs the future no more weakly than the present does not. Becker has shown, however, that someone who does discount the future, but does so only weakly (weighing tomorrow's hang-over at, say, a welfare level of 2.1 when, in fact, it will be experienced at a level 2) will not fall into a cycle of heavy use, while someone who takes the future into account but discounts it very heavily (allowing tomorrow's level 2 hang-over to weigh into his present deliberations as, say, 3.4) will end up in a cycle of heavy usage despite the fact that both agents, at every step of the way, rationally do what they most prefer at the time.

10. Things will be more complicated if tolerance or reinforcement have a tendency to subside. Depending on how quickly one can bounce back to default welfare levels in the absence of use, it might be rational to use for a few days and then stop, thereby reaping the benefits of use without eroding the expected utility of abstention to the point at which the use itself was not worthwhile.

Can the degree to which an agent discounts the future be rationally assessed? If so, then Becker could be taken to have shown (although he would not endorse this characterization of his view) that some forms of discounting of the future are rational—those that lead the agent to adopt a lifestyle of low consumption of tolerance-inducing, reinforcing substances—and others are not. The result would be that even under his model of addiction there might be a distinction in rationality between addicts and nonaddicts. Elster, who endorses an instrumental conception of rationality, holds that such a conception commits one to the claim that there are no standards of rationality by which tendencies to discount the future can be rightly assessed. He writes,

A time preference is just another preference. Some like chocolate ice cream, whereas others have a taste for vanilla: this is just a brute fact, and it would be absurd to say that one preference is more rational than another. Similarly, it is just a brute fact that some like the present, whereas others have a taste for the future. . . . If some individuals have the bad luck to be born with genes, or be exposed to external influences, that make them discount the future heavily, behavior with long-term destructive consequences may, for them, be their best option. We cannot expect them to take steps to reduce their rate of time discounting, because to want to be motivated by long-term concerns ipso facto *is* to be motivated by long-term concerns.¹¹

11. Elster, *SF*, p. 146. As discussed in Section III, in connection with Elster's own positive view of addiction, the opinion Elster offers is consistent with the view, which he also holds, that a tendency to discount the future can be irrational by virtue of its causes. But Elster does hold that a tendency to steeply discount the future is not irrational merely by virtue of its steepness.

The last sentence of the quotation seems to be offering the following, unsound argument: "(P1) Someone's preference for the present is irrational only if that person could be motivated to correct it. (P2) Someone who is motivated to have a preference for the future over the present already has that preference. (Conclusion) A preference for the present is never irrational; one could never both have such a preference and be motivated to correct it." The argument is unsound since P2 is certainly false, and P1 may be false. P2 is false, since one can be motivated to acquire a preference one lacks by many things other than the preference one is aiming to acquire; for instance, one can have a second-order preference for having a particular preference without already having the preferred preference. A person might wish, that is, that she cared more about the future

It is not clear that Elster is right about this, however. Becker makes an assumption, standard among economists, that is, in essence, a limitation on the rationality of particular tendencies to discount the future. He assumes that tendencies to discount the future must be consistent in the sense that one's preferences between pairs of outcomes will not flip-flop over time: If the agent prefers future good G_1 to future good G_2 at t , then she will maintain the same preferential ranking at all other times in which the two goods both remain future; she won't, that is, change her mind at some point and come to prefer G_2 to G_1 simply because of the way in which she tends to discount the future. This assumption makes it much easier to assess the rationality of a particular plan of conduct. From the point of view of a purely instrumental conception of rationality, if a plan of conduct is useful for getting G_1 but undermines the possibility of getting G_2 , then it is a rational plan of conduct in so far as the agent prefers G_1 to G_2 . If before the completion of the plan the agent will come to prefer G_2 to G_1 , even if only temporarily, then it becomes difficult to assess the rationality of the plan. In such cases, further principles of rationality need to be brought in to tell us how to weight the preferences of the agent at different times. The preferences of which temporal stages in the life of the agent are to be taken seriously as indicators of what the agent really prefers, and which are to be considered merely as indicative of passing fancies? (More on preference reversal in Section II.)

In order to avoid flip-flopping preferences as a result of temporal discounting Becker (following the standard practice in economics and rational choice theory) assumes that rational agents discount the future exponentially. This is to assume that rational tendencies to discount the future function something like the way in which one would rationally assign value to future payments now given a set degree of risk of failure to be paid. \$100 now is worth more to me than \$100 a year from now, even putting aside considerations of interest, if I think there is only a 90 percent chance that I will actually receive the \$100 a

without thereby caring about the future in much the same way as she might prefer that she preferred spinach to ice cream rather than the reverse without thereby preferring spinach to ice cream at all. P_1 is probably false as well. At least, a claim like P_1 about rational belief would clearly be false: one can have an irrational belief while lacking any kind of motivation to correct it. Why should the irrationality of a preference require motivation to correct it when the irrationality of a belief does not?

year from now. If I'm rational, I will value the \$100 a year from now as worth \$90 today. Similarly, \$100 two years from now is worth \$81 today if we assume that the assurance of receiving a good that is a year away is always 90 percent. The chance that I will not receive \$100 two years from now is $90\% \times 90\% = 81\%$. In general, if the payment is n years away, I should value it today as $(.90)^n$ of its face value, assuming that one year of time places acquisition of the good at risk by 10 percent.¹²

So, the assumption that rational agents discount the future exponentially can arise from the plausible view that temporal discounting is rational as a way of hedging against risk of nonpayment of future goods. This suggests that an agent discounts the future rationally if and only if the degree to which she discounts it matches the degree to which the futurity of a good places its acquisition at risk. But if this is so, then it is not clear that Becker has shown that any fully rational agent has ever actually ended up an addict in his sense. He has shown that this is theoretically possible, but whether or not it actually ever happens depends on whether or not the actual risk of failure to attain future goods prompts rational discounting of the future to the degree to which an agent would need to discount to end up an addict. To put the point more simply, it is clearly not rational to be entirely myopic, since there is surely less than a 100 percent chance that future goods will fail to be obtained. But, similarly, it is not rational, given epistemic limitations, not to discount the future at all, since there is some risk that future goods won't be obtained as expected. But whether or not the appropriate degree of discounting—the degree to which the future ought, rationally, to be discounted—will lead one into high levels of consumption of the sort that Becker takes to be definitive of addiction

12. Rational choice theorists usually justify the assumption that rational agents discount the future exponentially on the grounds that exponential discounters can turn those who discount nonexponentially into money pumps. This is a consequence of the fact that nonexponential discounters may suffer flip-flops in preference during which time they will be willing to buy goods at rates higher than those at which they are willing to sell the same goods at different times. The inconsistency in one's preferences over time, that is, can make one into an economic victim of those with temporally consistent preferences. However, to avoid being a money pump, one needs only to have temporally consistent preferential rankings. An agent who discounts the future linearly, for instance, will, like the exponential discounter, enjoy such consistency. So, the fact that rational agents are not money pumps does not provide a justification for the claim that rational agents discount exponentially rather than discounting in any other temporally consistent manner.

is yet to be determined. However, without such a determination it is unclear what the implications are of Becker's model to the questions of the addict's responsibility. The rational choice model of addiction encourages the thought that addicts are no less responsible for their behavior than any other rational agent who finds herself in a predicament in which rationality requires objectionable behavior. But if, as has just been argued, it is unclear that those who discount the future so steeply as to put themselves into a cycle of heavy usage are genuinely rational, it is unclear what Becker's model implies about the responsibility of addicts.

Various other objections to Becker's model can be found in the recent literature on addiction. Ole-Jørgen Skog, for instance, points out that Becker has only shown that a rational agent who heavily discounts the future can find himself in a lifestyle of consumption *higher than* an alternative, and better, level of consumption at which he would have rested had he discounted the future less heavily. But, Skog observes, it follows that agents who actually consume very little, by ordinary standards, will count as addicted under Becker's model if there is a lower level of consumption that they could have reached had they discounted the future less heavily.¹³ In addition, Olav Gjelsvik, Ainslie and Elster all argue that Becker's model fails to account for defining features of addiction. Gjelsvik argues that under Becker's model there is no reason to think that an addict who has managed to quit is more likely to relapse than any other rational agent is to consume in the first place. But this is clearly false: addicts have a much greater chance of relapsing than nonusers have of starting to use.¹⁴

Ainslie notes that addiction is often characterized by deep ambivalence manifested in efforts to quit that sometimes impose great costs on the agent. (In fact, the effort to characterize the nature of this ambivalence is one of the primary motivations behind Ainslie's own the-

13. Skog, "Rationality, Irrationality and Addiction—Reflections on Becker and Murphy's Theory of Addiction," pp. 185–86.

14. Olav Gjelsvik, "Addiction, Weakness of the Will and Relapse" in *GH*, pp. 48–49. Becker may have room to respond to Gjelsvik's criticism. After all, Becker points out that there is no reason to think that the degree to which an agent discounts the future should remain constant. (Gary Becker, Michael Grossman and Kevin Murphy, "Rational Addiction and the Effect of Price on Consumption," p. 329; quoted in Elster and Skog's introduction to *GH*, p. 24.) It is quite possible that an addict may quit when she comes to discount the future less steeply and will relapse when she returns to her usual manner of discounting.

ory of addiction, discussed below.) The behavior of an alcoholic who takes the drug Antabuse, and thus guarantees herself stomach-wrenching sickness when she takes her next drink, is not easily accounted for under Becker's model.¹⁵ Is such a person really discounting the stomach pains when she takes a drink? She will suffer the pains, after all, precisely because she chose to take Antabuse.¹⁶

Elster notes that certain behavioral compulsions that are thought of as addictions both colloquially and for purposes of treatment (most notably gambling addictions) are not adequately described by any model, of which Becker's is just one example, that takes either tolerance or the suffering of withdrawal symptoms as a result of abstinence (the assumed mechanism behind reinforcement in Becker's model) as a defining feature of addiction. Elster steps through various ways in which tolerance and withdrawal might be interpreted for gambling addiction—the most plausible interpretation of tolerance, for instance, is illustrated by the increase in size of the stake needed for a gambler to get the same level of excitement from the bet—and argues that none of the various ways of interpreting these concepts serves to identify adequately the distinctive mechanisms that are driving the behavior of the gambling addict.¹⁷

In addition to encountering powerful criticisms, however, in recent years Becker's approach to modeling addiction has been extended in various ways.¹⁸ Karl Ove Moene takes an approach similar to Becker's

15. George Ainslie, "A Research-Based Theory of Addictive Motivation," *LP* 19 (2000): 83; idem, *BW*, p. 18.

16. In addition, since rational choice theorists assume that there can be no distinction between what one judges to be best and what one is most motivated to pursue, it is very difficult for the rational choice theorist to account for weakness of will at a particular point in time. Given that assumption, how can an agent authentically judge one thing to be best and yet do another? For a related point, see Ainslie, *BW*, pp. 24–26. See also Gjelsvik, "Addiction, Weakness of Will and Relapse," pp. 49–52.

17. See Jon Elster, "Gambling and Addiction" in *GH*, pp. 208–34 (esp. pp. 215–17); and idem, *SF*, pp. 65–66. The point is, perhaps, even clearer in the case of certain eating disorders. The same degree of food deprivation has the same effect in decreasing a person's weight, even if she has been depriving herself in the past. Thus, at least one way of understanding tolerance cannot be naturally applied to anorexics and bulimics. There may be other possibilities. For instance, perhaps the more the anorexic has deprived herself in the past the more weightless she requires to feel the same level of relief.

18. One way in which Becker's model has been extended is by showing that a tendency to steeply discount future goods is not the only mechanism that can lead a rational drug user into a lifestyle of destructively high consumption. Richard Herrnstein and

in order to account for the way social dynamics can influence drug usage within populations.¹⁹ Moene argues that groups of rational agents, each pursuing the best means to satisfy their preferences, can end up in a society in which a higher percentage of people use than prefer use to abstention. The overall welfare of the group is likely to be significantly lower if the group has a high rate of consumption, and so the social dynamics of drug use might be an example of the “tragedy of the commons”: in each individual’s hurry to satisfy her preferences, the preferences of the group are damaged.

Moene generates this result from the following four, intuitively plausible, assumptions: people prefer to use when others are using, prefer to abstain when others are abstaining, prefer others to use when they are using, and prefer others to abstain when they are abstaining. Given these assumptions, Moene shows that there are two stable levels of consumption, one high, one low, where a level of consumption within a population is “stable” just in case it will perpetuate itself over time: If n percent of the population consumes at a particular time, and if n is a stable rate of consumption, then in the next time period the same percentage of people will consume. Whether a particular population reaches the low stable point or the high stable point depends on what happens when a consumer encounters a nonconsumer. Do both consume, or do both abstain? If consumption is the sufficiently frequent outcome of such encounters, for whatever reasons, then the society will soon find itself in a stable and excessively high level of consumption; if, on the other hand, the social pressures tend in the other direc-

Drazen Prelec, “A Theory of Addiction” in *CT*, pp. 331–60 show for instance, that destructively high levels of consumption can be reached by an otherwise rational drug user who ignores the fact that her behavior will lead to addiction. Such an agent need not discount future goods that she correctly anticipates. Instead, she ends up in a pattern of high consumption by failing to anticipate the effects of tolerance or reinforcement. There will be cases in which the kind of addiction-producing mechanism that Herrnstein and Prelec identify involves self-deception and so the resulting situation cannot be characterized as “rational addiction.” Also, Athanasios Orphanides and David Zervos have shown that a lifestyle of destructively high consumption can be reached by a rational drug user who doesn’t entirely ignore the possibility that she will become addicted, but underestimates the chances that the substance she consumes will cause tolerance and reinforce its own consumption through threat of withdrawal. (Athanasios Orphanides and David Zervos, “Rational Addiction with Learning and Regret,” *Journal of Political Economy* 103 [1995]: 739–58.)

19. Karl Ove Moene, “Addiction and Social Interaction” in *GH*, pp. 30–46.

tion, a low level of consumption will be reached. Moene's result is to group consumption what Becker's is to individual consumption. The crucial factor for Becker that determines whether a high or low level of consumption is reached is the degree to which the agent discounts the future; the crucial factor for Moene is the degree to which social pressures—whatever factors account for the choices made in cases in which those who prefer to consume encounter those who do not—tend towards use rather than abstention.²⁰

Since a theory like Moene's is primarily aimed at modeling the mechanisms governing group drug-consumption behavior, such theories do not have any clear implications regarding the responsibility of individual addicts. However, models like Moene's can help us to answer questions that are all but intractable without employing such a model, questions with immediate policy implications. For instance: How do the costs of obtaining a drug influence the level of drug consumption that we find within a society?²¹ If it can be shown that a very high cost for obtaining a drug will lead to a low, stable level of consumption, then that might speak in favor of tough penalties and enforcement policies for offenders, or for heavy taxation on, for instance,

20. The similarities in structure between Moene's view and Becker's make Moene's theory subject to some of the same criticisms that have been launched against Becker's view. In particular, Skog's point that "high levels of consumption" and "low levels of consumption" are too gross measures by which to distinguish addicts from nonaddicts applies equally to Moene's model of social consumption. A society could have what is, by ordinary standards, a low level of consumption of a substance whose use is subject to the social constraints Moene imagines even though there is a yet lower stable level of consumption that could be reached if social pressures weighed differently than they actually do. Should we say that this would be a society of drug abusers? To do so would be to distort our ordinary concept of drug abuse, or addiction. In addition, the sort of criticism developed above in the main text (under which an agent's tendencies to discount the future are rationally assessible) can be extended to Moene's theory. Perhaps there are rational ways for encounters between those who prefer use and those who prefer abstention to be settled. Whether or not this is so is a difficult problem in bargaining theory, but it is not clear that there are no rational standards to be brought to bear in the adjudication of disputes between those with conflicting preferences. However, whatever problems the theory might encounter when interpreted as a general model of an addictive society of rational agents are irrelevant to the main purpose of Moene's theory, which is to model the way in which social factors can result in less than optimal equilibria of usage, the question of the rationality or irrationality of the members of the society being orthogonal to this question.

21. See Moene, "Addiction and Social Interaction," pp. 38–40 for a discussion of the implication of drug costs given his model.

alcohol and tobacco. Or, alternatively, a model like Moene's might have the opposite implication suggesting that costs would have to be increased to impossibly high or morally objectionably high levels in order to produce a low, stable level of consumption. What this suggests is that the approach to understanding addiction developed by Becker is a powerful one, despite its limitations.

Both the power and the limitation of the approach to addiction taken by classical rational choice theorists like Becker derives from the power and the limitation of classical rational choice theory itself. Rational choice theory, especially when harnessed to model agential weaknesses such as addiction, is neither purely descriptive nor purely normative. It neither aims to provide an analysis of our "ordinary" concepts nor does it function as advice to would-be rational agents. Rational choice models, then, are difficult to assess philosophically. Whether or not they serve as adequate analyses depends on what we take them to be attempting to analyze, and it is not always clear what their target is. Still, the rational choice theorist's model of human deliberation, and the behavior in which it issues, does reflect something important about what might be happening when the addict acts, even if it does not provide answers to all of our descriptive and normative questions about addiction. In particular, rational choice models like Becker's do accurately describe the mechanics of traps into which rational agents may fall simply by exercising their capacities for rationality. Whether or not the conditions that would trap a rational agent into non-optimal levels of consumption are actually those that drive the behavior of any actual, fully rational addict, is another matter. Without an answer to this further question, however, it is unclear what normative lesson to draw from the rational choice model of addiction. Whether or not addiction diminishes responsibility, in the context of rational choice theory, will depend on whether or not addictive behavior is rational behavior and, as we've seen, Becker's theory, in any event, does not provide as clear an answer to that question as one would hope for.

II. AINSLIE AND HYPERBOLIC DISCOUNTING

Rather than try to account for the addict's behavior by employing the tools of traditional rational choice theory, we might, instead, model

addiction by denying that one or more of the assumptions underlying the rational choice approach are true of addicts. We might try, that is, to explain how a stable level of unhealthy consumption could be motivated, even in the face of knowledge that it is unhealthy, without assuming full rationality on the part of the addict. This is Ainslie's approach.²² As we've seen, rational choice theorists assume that the discounting of future goods by a rational agent is exponential. However, Ainslie claims that the distinctive feature of addiction is that the addict discounts future goods hyperbolically.

Mathematically, hyperbolic discounting can be understood as follows. The equation by which we calculate the value now of a future good under exponential discounting is of the following form: $Y = G \times C^D$, where Y equals present evaluation of a future good of actual worth G , C equals the rate at which the agent discounts the future (i.e., 0.9, if the agent takes the acquiring of future goods to be only 90 percent assured), and D equals the delay from now until the time the future good will be acquired. On the other hand, the equation by which we calculate the value now of a future good under hyperbolic discounting is of the following form: $Y = G/(1 + D)$. With both exponential and hyperbolic discounting, as the delay D approaches zero, as the time of acquisition draws near, the value assigned to the good, Y , approaches its actual value, G . But there is an important difference between the two kinds of discounting: Consider the change in assigned value of a future good over the course of the final unit of time before the good is acquired. Say an agent is offered \$100 to be paid one day from now. Both the exponential and hyperbolic discounter will value that \$100 as worth \$100 when tomorrow arrives. But if you are an exponential discounter who weighs future goods at 90 percent of their face value, you will value \$100 a day from now as worth $\$100 \times 0.9^1 = \90 today. If you are a hyperbolic discounter, however, you will

22. Ainslie has expressed his theory of motivation in a variety of places. Cf. "Derivation of 'Rational' Economic Behavior from Hyperbolic Discount Curves" in *American Economic Review* 81 (1991): 334–40; *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person* (Cambridge: Cambridge University Press, 1992); "The Dangers of Willpower: A Picoeconomic Understanding of Addiction and Dissociation" in *GH*, pp. 65–92; "The Intuitive Explanation of Passionate Mistakes, and Why it's Not Adequate" in *AEE*, pp. 209–38; "A Research-Based Theory of Addictive Motivation"; *BW*.

value that same \$100 as worth $\$100/(1 + 1) = \50 today. So, in the final day before the \$100 is received, the exponential discounter's evaluation only increases by \$10, or 10 percent of the good's actual value, while the hyperbolic discounter's evaluation increases by \$50, or 50 percent of the actual value of the good. Hyperbolic discounters care very little about a good until the final moments before it is available, during which time they experience a drastic shift in their evaluation of it. Exponential discounters do not experience much more of a change in their evaluation over the final day before the good becomes available than they do over the next to last day, or the second to last.

This difference between exponential and hyperbolic discounters suggests that hyperbolic discounting more closely approximates the phenomenology of craving. Someone who is subject to strong cravings for sugar might look forward to tonight's dessert more at lunch than at breakfast as her desire for the dessert increases over the course of her day; but after the dinner entree is completed, and before the dessert arrives, she might find her desire for dessert increasing much more rapidly than it did in a comparable amount of time earlier in the day. In addition, recall that part of the motivation for thinking that a rational agent's tendencies to discount the future were best modeled exponentially came from the fact that an exponential discounter's preferences are temporally consistent: if he ranks one outcome over another at one time, he will not change his ranking merely because time has passed. However, hyperbolic discounters will experience flip-flopping preferences between two future goods when the attainment of one is farther in the future than the other, provided that the time gap between the two goods is such that the craving for the first will kick in before the craving for the second.

Figure 1 illustrates the change over time in evaluation of two goods by an exponential discounter and a hyperbolic discounter. At the point at which the hyperbolic discounter's evaluation curves cross, she ranks the two goods to be of equivalent value. Prior to that time she ranks the level 10 good below the level 20 good, and after that time, and before she attains the level 10 good, she ranks it higher than she ranks the level 20 good. So, she experiences a preference shift. On the other hand, the exponential discounter ranks the two goods similarly throughout the time interval, although each becomes more attractive in her eyes as the time of its acquisition draws closer. The problem

faced by the hyperbolic discounter is that in the case of drug use, the two goods are not jointly attainable. If we think of the level 10 good as the pleasure received from consuming the drug, and the level 20 good as the good that the addict can have if she manages to avoid using at time 10, then it appears that a person who discounts the future hyperbolically will end up taking the lesser of two goods despite the fact that

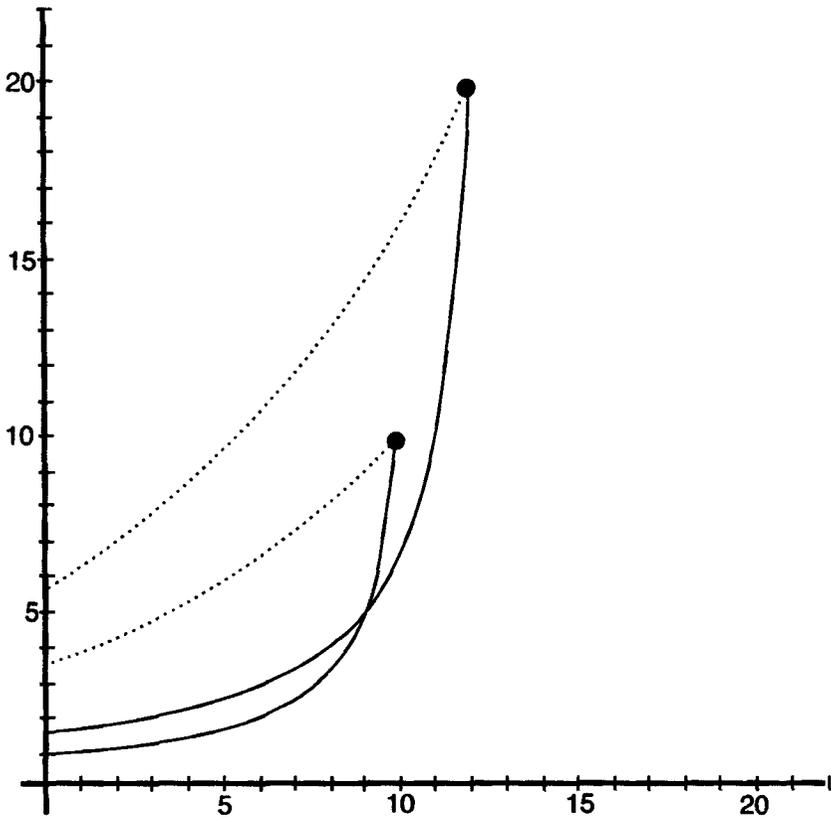


FIGURE 1

The x-axis is time. The y-axis is value. The dot at (10, 10) indicates a level 10 good attainable at time 10. The dot at (12, 20) indicates a level 20 good attainable at time 12. The solid lines indicate the evaluation of a hyperbolic discounter of the two goods. The dotted lines indicate the evaluation of an exponential discounter of the two goods.

at each and every moment she acts so as to satisfy her present preference at the moment at which it can be satisfied.

Hyperbolic discounting is surely irrational, and for a number of reasons: unlike the exponential discounter, there is ambiguity about what the hyperbolic discounter really prefers—depending on which of her temporal time slices we ask, we will get different answers; and unlike the exponential discounter, the hyperbolic discounter does not consistently assess the risk that the futurity of a good places on the chances of its being acquired.²³ Therefore, under Ainslie's model of addiction, the high levels of consumption in which addicts engage are a product of irrationality. However, under Ainslie's model, there is no time at which the addict acts contrary to her present preference. She finds herself in a cycle of high usage solely because of her tendency to hyperbolically discount the future.

Ainslie's model is superior to Becker's in a variety of respects. Most importantly, Ainslie is able to account for the ambivalence that many addicts experience. Under Becker's model the agent is never in disagreement with herself; there is no time at which she is divided, and her different time-slices are also in agreement—at least, in their rankings of various outcomes. Ainslie's addict never experiences conflict *at a particular time*; there is no time at which she can be rightly said to want to abstain more than consume and, at the same time, to want to consume more than abstain.²⁴ In this she is like Becker's addict. But, Ainslie's addict experiences cross-temporal conflict: there is a time at which she wants to abstain more than consume (when she is not in the grip of craving); and there is another time (while in the grip of craving) in which she wants to consume more than she wants to abstain.

Since under Becker's model, different time slices of oneself—the addict during the day at work, the addict that night busy in efforts to score—are not in preferential disagreement, they can work together. At work that day, the addict might make plans to ease her efforts to

23. In addition, unlike the exponential discounter, the hyperbolic discounter can be turned into a money pump by buying goods from her before she is in the grip of a craving and selling them back to her at inflated rates when the craving strikes.

24. This possibility is not as analytically puzzling as it might appear. The possibility can be accounted for in a number of ways, most notably for our purposes by specifying distinct kinds of "wants." See Section III.

score that evening; she might, for instance, skip lunch so that she won't have to stop at the bank machine before visiting her dealer; or, anticipating the aftereffects of use, she might tell her boss to expect her late the next day. However, the person-stages of Ainslie's addict find themselves at war. When in the grip of craving, the addict takes steps that preclude the satisfaction of her later preference. Later she will wish she had stayed sober so that she could attend her child's recital, but now, while her discount curve climbs steeply and crosses her evaluation of sober attendance of the recital, she will actively take steps that prevent her later self from satisfying its preference.

This effect of preference shift suggests that an addict's predicament might be understood along the lines of preferential conflict between individuals or nations—conflict, that is, in which satisfaction of one party's preference precludes satisfaction of the other's. In some cases, such conflicts are simply settled by power: the strongest individual gets her way. However, such conflicts are also settled by the exploitation of opportunity: a weaker party might get her way because she gets to act first; such is the case, for instance, in most draft lotteries for professional sports teams: the team with the worst record, the weakest team, gets first pick of new prospects. In someone who habitually gives in to temptation, her tempted person-stage exploits the fact of her temporal priority to act for her own satisfaction before the competing sobriety-preferring self has an opportunity to act. The bind in which addicts find themselves is that the tempted self seems always to have temporal priority, and, therefore, would appear always to have the opportunity to satisfy her craving at the expense of the agent's long term interests. Addiction, then, can appear inevitable for agents who discount hyperbolically.

Addiction, however, is obviously not inevitable. Many people who are quite vividly aware of the immediate pleasures promised by drugs never engage in unhealthy levels of consumption, even though they do, in fact, discount hyperbolically. Ainslie accounts for this, however, by arguing that the appearance of inevitability derives from the assumption that the earlier tempted-self gains a decisive strategic advantage over the later-self. This is not so, however, since the right model is not one of preferential conflict between agents just once, but, instead, of repeated conflict. The temptations will subside and the untempted agent will find herself regretting her early choices; *but* she will also

anticipate later temptation, and can take steps to preclude her later tempted self from taking advantage of her time of control.

There are a variety of mechanisms by which an untempted self can do her best to prevent her later self from giving in to temptation. She can lock the liquor cabinet, for instance. Thomas Schelling gives the example of a cocaine addiction clinic in which patients write a letter admitting their addiction that will be sent to family, friends or business associates should they relapse.²⁵ But could there be a mechanism through which a *tempted*-self could act so as to defeat the satisfaction of her own preference for giving in to temptation? Ainslie thinks he has identified a mechanism whereby a tempted-self, by recognizing her involvement in an iterated preferential conflict with a self who will enjoy motivational control prior to such control being regained by the tempted-self, will be led, rationally, not to act on the present temptation. He calls the exercise of this mechanism "exercise of will," and associates it with the adoption of what he calls "personal rules."²⁶

The mechanism is supposed to function as follows.²⁷ Imagine an agent who has the opportunity to smoke crack every evening at 8 P.M. When she wakes in the morning on Monday, she evaluates that evening's prospective high at a very low level, and values having a sober morning on Tuesday more highly. Since she discounts the future, she assigns neither option the value that it would actually have to her at the time it would be enjoyed, but she does rank a sober morning on Tuesday over a high this evening. During the course of her day, assuming she is a hyperbolic discounter, this ranking remains constant, until, say, 7 P.M., when craving sets in, her discount curves cross, and she comes to rank the high at 8 P.M. over the sober morning on Tuesday. So far it would appear that she will use at 8 P.M. if she is endeavoring to satisfy her present preferences, as Ainslie, following traditional rational choice theory, assumes. However, the key to recognizing how she can avoid indulging her craving, Ainslie thinks, comes from examination of the preference that she has at 8 P.M. on Monday for indulgence of the same craving on each successive day, compared to morning-after sobriety. Since craving for a Tuesday evening high has not yet set in on

25. Thomas Schelling, "Self-Command: A New Discipline," in *CT*, p. 167.

26. See, particularly, Ainslie, *BW*, pp. 78–88.

27. For another discussion of Ainslie's view of will power see, Ole-Jørgen Skog, "Hyperbolic Discounting, Willpower, and Addiction" in *AEE*, pp. 151–68.

Monday evening, she still prefers Tuesday evening abstention to Tuesday evening use, and the same can be said for every day after. Further, if she adds together the temporally discounted expected utility of not using on each successive day and compares it to using on each successive day, she finds that she is better off in the long run not using at all—that, after all, is one of the hallmarks of addiction: consumption at an *unhealthy* level compared to abstention. At 8 P.M. on Monday, then, three things are true of the agent: (1) she prefers the action of using right now to the action of abstaining right now, (2) she prefers abstention in the future, to use in the future, and (3) the preference in (2) is stronger than the preference in (1): she prefers never using again to using now. As yet, however, the relative strength of these two preferences doesn't motivate the agent to abstain on Monday at 8 P.M. since the preference for present use is perfectly compatible with a preference for future abstention. The best way to satisfy her Monday 8 P.M. preferences, that is, is by using on Monday at 8 P.M. and never using again. Action of this sort, however, will result in a pattern of use since on Tuesday at 8 P.M. her best course of action will appear to be using on Tuesday and never using again. How can the agent make it the case that abstention on Monday at 8 P.M. is favored by her preference for future abstention over present use? Ainslie suggests that this is done through the adoption of "personal rules." To adopt a personal rule is to conceive of your present choice as evidence of what you will choose in similar circumstances in the future. If you conceive of your choice in this way, according to Ainslie, then it will be the case that present use is actually incompatible with future abstention: if you use now, you will also use in the future. And, since you prefer future abstention to present use, you will thereby choose abstention now. Ainslie is recommending placing all of one's temporally discounted evaluations of future abstention on the opposite side of the scale from present use. Although each is of very small value, there will be many of them, and so they can outweigh the very high value assigned to present use. This is only possible if present use precludes the possibility of future abstention. Ainslie claims that for those who adopt personal rules, this is true, and thus the adoption of personal rules is a means of manipulating one's present preferences for future abstention in order to avoid the shortsighted preference for present use. Those who manage to act in accordance with personal rules, and thereby overcome temptation,

are rational agents, according to Ainslie. They manage to motivate themselves in such a way as to overcome the weaknesses induced by hyperbolic discounting.²⁸

Ainslie gives a prominent role in the mechanism of “willpower” to the belief that a choice at a particular time is an indicator of what one will choose at future times. It is important to Ainslie’s position, however, that the belief is not irrational. If it were irrational, then “willpower” would be a way of inducing one kind of irrationality in oneself so as to overcome the irrational weakness induced by hyperbolic discounting. But if we were looking for a mechanism whereby an agent could overcome temptations even while slipping into irrationality, we might as well just say that the best way for an agent to overcome her powerful temptation to use is simply to abstain, period. To be sure, this would be irrational (she would be acting contrary to her strongest preference) but if we allow irrationality into the picture, there is no reason to point to complicated rule-based irrational mechanisms for overcoming temptation; simpler irrational mechanisms would do just as well.²⁹

Michael Bratman has argued that there is no reason to think that an agent is generally under rational pressure to think that her choice now is an indicator of what she will choose in the future.³⁰ There are, to be sure, cases in which a choice on Monday to violate a rule is very good evidence that the agent will not follow the rule in the future, evidence that no rational agent could ignore. We can imagine a world, for in-

28. Howard Rachlin, *The Science of Self-Control* (Cambridge: Harvard University Press, 2000), thinks of successful exercises of self-control as the initiation of patterns of behavior that supplant other, less healthy, patterns. Although Rachlin places no strong emphasis on the preference shifts experienced by hyperbolic discounters, he does take hyperbolic discounting seriously, and there are strong affinities between his views and Ainslie’s. In addition, Rachlin’s book is of particular interest for its thorough examination of recent empirical work in behavioral psychology regarding the effectiveness of various techniques for overcoming patterns of unhealthy behavior.

29. Ainslie holds that the belief that one’s present choice is decisive evidence about what one will choose in the future is self-fulfilling: if one has it, then it is more likely to be true than if one lacks it. (See *BW*, p. 88.) However, even if the belief does sometimes cause the conditions that make it true, this fact is only relevant if the causal chain passes in the right way through the agent’s capacities for rationality. This point is elaborated below in the main text.

30. Michael Bratman, “Planning and Temptation” in *Faces of Intention: Selected Essays on Intention and Agency* (Cambridge: Cambridge University Press, 1999), pp. 35–57.

stance, in which there is decisive evidence that there are only two kinds of agents: those who always choose to use and those who always choose to abstain. If an agent rationally believes that there are only these two kinds of people, then she can expect that should she violate a rule against use now, she will also do so in the future. But the conditions specified in this example do not necessarily hold. The only kinds of circumstances relevant to Ainslie's project that would also place an agent under rational pressure to expect her future-self to act as her present-self is acting, are those circumstances that are logically required by a presumption of rationality on the part of the agent. For instance, the presumption of rationality implies that the agent will act according to her best reasons, so if choosing to violate the rule against use on Monday at 8 P.M. would give a future-self a decisive reason to use Tuesday at 8 P.M., then the agent would be under rational pressure on Monday to abandon an expectation of abstention on the part of the Tuesday-self should she choose to use on Monday.

Bratman notes that past abstention doesn't generally give an agent reason to abstain now, so it is no condition of rationality that a choice of abstention encourage an expectation of future abstention. What follows is that an agent has no reason to expect a choice of abstention on Monday to contribute to her gaining the long-term anticipated rewards of abstention, and so she has no reason to abstain on Monday. If Bratman is right, then Ainslie has not identified a mechanism that will help a rational agent to overcome temptation. However, Alfred Mele has argued that there is a large class of cases in which past abstention does provide strong reason to abstain, and, he claims, cases of addiction are often of this sort.³¹ In the kinds of cases Mele has in mind, to use in the face of past abstention would be to waste the effort spent on past abstention.³² If, for instance, in order to be released on your own

31. Alfred Mele, "Addiction and Self-Control," *Behavior and Philosophy*, 24 (1996): 99–117 (henceforth *BP*).

32. Both Mele and Bratman in his description of Mele's examples (see Bratman, "Planning and Temptation," p. 49 n. 21) talk about past abstention as providing an agent with a reason to abstain by encouraging her that she will refrain in the future and thereby reap the benefits of continued abstention. Notice, however, that the encouragement provided by past abstention is not crucial to such examples. All that matters is that one cannot gain the goods promised by past abstention if one doesn't abstain now, and so the fact that one abstained in the past gives one further reason to abstain now. Anticipating this gives one's earlier self a reason to abstain that is rooted in the expectation

recognizance you need to be clean for thirty consecutive days, then days of past abstinence are like money in the bank that would be squandered entirely by using today.³³ In such cases, a rational agent can expect abstinence now to give reason to her future self to abstain; this in turn grounds an expectation of future abstinence, which, in turn, allows one to take the goods of continued abstinence into one's rational deliberations about what to do now.³⁴

However, even in cases in which Ainslie's mechanism of willpower

that one's future self will appreciate the reason-giving force of one's abstinence now and, therefore, will abstain.

33. Contestants on "Who Wants to be a Millionaire?" face situations of this sort. With each right answer the possible reward increases, but with a single wrong answer the contestant leaves with some lower reward. So, a contestant who has answered enough questions correctly will have to choose between answering a question or not answering it where a correct answer earns him \$1,000,000; not answering will earn him \$500,000; and answering incorrectly will earn him only \$32,000. In choosing whether or not to answer the question, the contestant is given an incentive, the possibility of leaving with \$1,000,000, to risk wasting his past correct answers, which are worth \$468,000.

34. Mele has also argued that the kinds of cases in which Ainslie's "personal rules" do provide a rational agent with the tools for resisting present temptation—those in which to violate the rule would be to waste past efforts that led to successfully following it—are closer to the predicament of the addict than the kinds of cases in which personal rules do no good (*BP*, 107). Mele is probably right about this in the case of nicotine, but it is less clear in the case of other more dramatically and immediately damaging drugs. The primary problem with smoking is the long term negative effects on one's health. Further, the nicotine addict knows that a week of abstinence will do little for his long term health if he returns to smoking today. That is, the value of past abstinence is lost at the point of relapse. The smoker who quit and starts again goes back to "square one." (This may be even clearer in cases of compulsive overeating.) But compare the case of nicotine to the case of crack. While it's true that some of the goods of past abstinence are lost with relapse—abstinence from crack does have an incremental positive effect on long term health that can be ruined by relapse—these goods are minor compared to other goods obtained through abstinence the acquisition of which do not depend on past abstinence. By abstaining today, for instance, the crack addict avoids the degrading things that she does for another hit once she has run out of money to pay for it. These goods are gained just by abstaining now, and do not depend on past abstinence. In cases of this sort, Ainslie's model of the mechanism of willpower cannot help a rational agent to overcome temptation, and for the reasons that Bratman suggests. While the fact of past abstinence might give the agent some reason to abstain now, it is a very weak reason indeed and not one that will support an earlier expectation of later abstinence. If anything will prevent the crack addict from giving in to temptation, it must be reflection on the horrible things that she will do for more once in the grip of the drug. But if she discounts these evils hyperbolically, as on Ainslie's model, there is no reason to think a rational agent capable of giving them the weight in her present deliberations needed to topple the attractions of use.

does help a rational agent to overcome temptations produced through hyperbolic discounting, there is reason to think that what Ainslie has identified is not actually “willpower” in the sense in which we usually think of it. In fact, the best reasons for thinking that Ainslie is not talking about the will at all come from some of the very effects of adoption of personal rules that he articulates in his recent work. As Ainslie summarizes his examination of the effects of using “willpower” in his sense, “Nothing fails like success.” (*BW*, pp. 141ff.) That is, Ainslie identifies a variety of fascinating psychological mechanisms through which the adoption of personal rules results in alienation, disassociation and even the development of compulsive disorders structurally very similar to addictions.³⁵

Ainslie distinguishes compulsion and addiction as follows: An addict suffers preference shifts due to hyperbolic discounting at intervals of hours or days, while a compulsive suffers the same preference shifts at intervals of weeks or months.³⁶ The adoption of personal rules serves the interests of longer-term goods over short-term goods, and so personal rules could be harnessed to serve compulsive cravings over addictive cravings, even though the agent’s longest-term welfare would be best served if she were to overcome her compulsive cravings as well. So, for instance, you might find yourself, each evening after dinner, preferring to leave the dishes until the next day, even though before dinner you prefer to do them that evening, and regret the next morning having left them to the cockroaches. To fight against this lazy craving, you might adopt a rule, “Do the dishes right after dinner,” and thereby provide yourself with the greater rewards of a clean, cockroach-free kitchen. But there are also advantages to flexibility: enjoying time with friends whom you’ve had over for dinner, rather than busily doing dishes, might be worth the added costs of dealing with the mess in the morning, even if it is not worth it to wait on nights in which you

35. Ainslie discussed the negative side-effects of willpower in *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person*, cf. pp. 205–13. His examination of these negative side-effects has been substantially extended in his more recent work. See, especially, “The Dangers of Willpower: A Picoeconomic Understanding of Addiction and Dissociation” and *BW*, pp. 143–97

36. *BW*, pp. 48–51. In a fascinating section of *BW* (pp. 54–61), Ainslie argues that pains are addictions in which the temporal gap between cravings is almost instantaneous. Thus, Ainslie thinks of pain, addiction and compulsion as being on a continuum.

don't have company. By overcoming the dangers brought on by hyperbolically discounting the goods of a clean kitchen, you come to "crave" the act of following the rule, even when you will regret, the next day, having not enjoyed your friends' company. You don't, that is, necessarily confront the rule "Do the dishes right after dinner" with the goods gained through following the rule "Do the dishes right after dinner except when you have company," and so you end up giving in to the short term preference for avoiding laziness over the longer term preference for enjoying your friends. The tool for solving your craving for laziness, the adoption of the rule for doing the dishes, becomes an affliction; you become a compulsive dish washer.³⁷

Ainslie thinks that results of this nature show only that "willpower" is a relatively blunt tool for solving the problems that hyperbolic discounting engenders. However, it is possible that what these results really show is that the motivational mechanisms that Ainslie dubs "the will" are really not to be equated with the will at all. I explain.

There is a long tradition of distinguishing the will from desire. To have a will is to be capable of directing one's conduct contrary to all, or at least many, of one's desires. As intuitive as this distinction is, it is far from obvious what it really amounts to. Since desires can be brought into motivational combat with other desires, showing that an agent is motivated contrary to a particular prominent desire is not sufficient to show that the moving force behind her conduct is to be equated with the will: she might just be acting on another desire. A more important distinction between desire and the will is this: Where desires place us under *equivalent* rational pressure either to act as they direct *or* to give them up, acts of will place us under *greater* rational pressure to act as they direct. Someone who abandons her desire to go to Paris is no less rational than someone who wants to go and books a ticket with Air France. However, someone who *intends* to go to Paris, or *chooses* to go to Paris, owes us more of an explanation if she abandons her intention or her choice than if she takes steps to do as it directs. Similarly, if I tell you that I intend to spend the weekend in Paris, you can count on me doing so with greater surety, if I am ratio-

37. What Ainslie is offering here is closely analogous to a well-known criticism of Kantian ethics, namely, that the rule-based conception of the best life that the Kantian advocates results in an inappropriate subordination of one's own personal projects and interests.

nal, than you can if I merely say that I want to. What this suggests is that exercises of will are a product of our rational capacities in a way mere desires are not. Both direct us towards conduct, but since exercises of will spring in some way from rationality itself, they place us under rational pressure to do as they direct that desires do not.³⁸ There is great truth, that is, to the Aristotelian-Scholastic equation between will and “rational appetite.”

Exercises of will in Ainslie’s sense, however, do not obey this asymmetry with other forms of motivation. Since, ultimately, exercises of will are merely preferences for particular patterns of conduct brought to weigh in against preferences for single and immediate outcomes, they have no greater rational status than the preferences they are to combat. An agent who abandoned one preference or the other would be no less rational than one who acted to satisfy it instead. In fact, Ainslie’s observations to the effect that employment of his mechanisms of willpower often results in alienation and compulsive rule-following bears this point out. Any account of the will ought to respect the fact that the irrationality of desire and the connections between the will and rationality are such that when the will topples desire it has a rational justification available. It hasn’t just won a battle through force, it

38. There are a wide variety of different ways to account for the connection between the will and rationality. At the extreme, we might think, as Kant did, that to will is to direct conduct in accordance with categorical principles of action dictated by the very nature of practical reason. (Cf. Christine Korsgaard, *The Sources of Normativity* [Cambridge: Cambridge University Press, 1996], esp. Lecture 3.) At the other extreme is a view that accounts for the difference between will and desire by taking the will to be a special species of desire and then arguing that the species-defining characteristic suggests that acts of will are connected with our capacities for rationality in a way in which other desires are not. One might, for instance, associate the will with the strongest desire and then argue that there are rational grounds to act in accordance with the strongest desire that do not apply to desires across the board. Or, more promisingly, one might associate the will with the desire to act in accordance with reasons. (Cf. J. David Velleman, “What Happens When Someone Acts?” in *Perspectives on Moral Responsibility*, John Fischer and Mark Ravizza eds. [Ithaca: Cornell University Press, 1993], pp. 188–210.) Between these two extremes are a variety of other positions. One might, for instance, take acts of will to be mental states distinct from desire that play certain special roles in practical reason, and thus are governed by principles of rationality that do not govern desires, without thereby associating the will with practical reason itself. (Cf. Michael Bratman, “Toxin, Temptation and the Stability of Intention” in *Faces of Intention: Selected Essays on Intention and Agency*, pp. 58–90.) Clearly, a full discussion of these issues cannot be undertaken here.

also has right on its side. For all that, of course, one might wish that desire had won; there might be a strong sense of loss, but it cannot be a sense of loss for which good reasons can be given, for desire lacks the rational basis that exercises of will enjoy. However, on Ainslie's model of the will, the compulsion and alienation that the agent endures as a result of mastering temptation is no more supported by reasons than her conduct would have been had she given in to temptation. She has really just traded one unhealthy cycle for another. Since the compulsion into which she has fallen and the addiction she has left behind both have their roots in the same irrational tendency (the tendency to discount the future hyperbolically) Ainslie does not have room within his theory to make the kind of distinction in rationality between them that is required to make a convincing case for the claim that the adoption of personal rules is really to be equated with exercise of will. The person who exercises willpower, in Ainslie's sense, has not genuinely confronted passion with reason—instead she has just confronted one passion with another and so she has not, in fact, exercised her will, truly speaking, at all.³⁹

In denying that the addict's affliction involves a subversion of her rationality, the traditional rational choice theorists, like Becker, fail to account for many of the most important features of addicts that distinguish them from the unaddicted. In conceiving addictive motivation to spring from irrational tendencies that plague all forms of motivation, Ainslie fails to provide an adequate account of the way in which the will can be a genuine source of rational motivation. When we look to a model of addictive behavior for guidance regarding the sense and degree in which the responsibility of addicts is diminished, the failing of Ainslie's theory appears serious. From the point of view of assessing responsibility, the attraction of a view that presents addictive behavior as a product of irrationality is its promise to help to distinguish addictive behavior from the worst kind of deliberate immoral or illegal be-

39. R. Jay Wallace, "Addiction as Defect of Will," *LP* 18: 621–54, suggests that no account of addictive motivation will be adequate that fails to give a special motivational role to the will as a motivating capacity different from desire. In arguing for this claim, he writes that by exercising the will, ". . . persons can bring about a kind of rational action that is not merely due to fortuitous coincidence of rational judgment and given desire, but that is a manifestation of the very capacities that make them, distinctively, agents." (pp. 637–38).

havior. However, Ainslie's view does not provide us with the needed contrast, for the central motivational mechanism behind addiction—hyperbolic discounting—is endemic to addicts and nonaddicts alike. Those who avoid giving in to addictive temptations do so through the exercise of the very same form of irrational mechanism that plague addicts and consequently there is no meaningful sense in which they are more rational when they act wrongly than addicts are. As psychologically rich as Ainslie's model of the addicted agent is, it remains too impoverished to confront the pressing normative questions about the addict's responsibility.

III. VISCERAL FACTORS AND THE DISTINCTION BETWEEN COMPULSION AND WEAKNESS

While Ainslie denies that one of the central features of traditional rational choice theory's picture of the rational agent (the tendency to discount future goods exponentially) is true of addicts, he accepts another assumption of the traditional rational choice model: an equation, at any given time, between an agent's evaluative rankings and her motivating preferences.⁴⁰ Rational choice theorists, that is, tend to assume that what a person judges to be best is what she prefers most and vice versa.⁴¹ This might be true of fully rational agents, but it is not clear that it is true of addicts, and it could be that to understand addiction we need to understand how evaluation and motivation can pull apart.⁴² Although they do not put it in quite this way, George

40. There is room in Ainslie's theory to pull these two things apart, but there is not room to do so while giving any meaningful motivational role to evaluations. An agent might judge a future good to be worth a merely exponentially discounted value, while "feeling" attracted to it to a hyperbolically discounted degree. However, Ainslie is committed to the claim that it is only the hyperbolically discounted feeling that actually influences present behavior. The exponentially discounted judgment does not compete in the marketplace of motivation.

41. Sometimes the assumption is thought to be essential to a naturalistic conception of human motivation. It is sometimes thought, that is, that a causal theory of the motivational role of evaluations requires an equation between one's preferences and one's judgments. Alfred Mele, *Irrationality: An Essay on Akrasia, Self-Deception and Self-Control* (Oxford: Oxford University Press, 1987), pp. 31–49 argues for the compatibility of a causal theory of agency and the view that evaluative judgments have a different motivational role from one's desire-based preferences.

42. As an historical note, John Locke felt that motivation and evaluation had to be

Loewenstein's contribution of the "visceral factors" view of addiction, and Jon Elster's development of it, help to show how this might be possible.

Loewenstein has coined the term "visceral factors" to describe motivational influences that fall into one of three categories: drives (such as hunger and sexual desire), emotions (such as anger and jealousy), and bodily sensations (such as pains and itches).⁴³ The visceral factors in motivation are contrasted with "higher level" motivating factors, where these are understood to be motives that are "cognitively mediated." Elster has specified in some detail what such "cognitive mediation" involves and has recognized that some visceral factors, especially emotions, also involve cognitive mediation in the weak sense that they are sensitive to the agent's beliefs.⁴⁴ Since the visceral factors and the cognitive factors compete to guide every agent's behavior, Loewenstein and Elster are going against the usual assumption of rational choice theory eliding the evaluational and the motivational. Our evaluative judgments, which are cognitive factors in motivation, can influence behavior independently from the influence of the visceral, or purely appetitive, motives. Where the addicted differ from the unaddicted is in their susceptibility to powerful visceral factors. Loewenstein and Elster understand addiction, that is, to be an acquired susceptibility to visceral motives directing the agent towards that to which she is addicted.

Loewenstein has also done interesting empirical work indicating that, in fact, people are very bad at predicting the likelihood that they will act as directed by a visceral motive when they are not experiencing it. People who are not sexually aroused, for instance, underestimate the motivational efficacy of future sexual arousal.⁴⁵ What this

distinguished in order to account for cases of weakness of will. See John Locke, *An Essay Concerning Human Understanding*, Peter Nidditch ed. (Oxford: Clarendon Press, 1975), pp. 252–64 (book 2, chap. 21, sec. 35–47). For discussion, see Gideon Yaffe, *Liberty Worth the Name: Locke on Free Agency* (Princeton, Princeton University Press, 2000), pp. 32–61.

43. Loewenstein has presented his "visceral factors" view in a number of different places. Cf. "Out of Control: Visceral Influences on Behavior," *Organizational Behavior and Human Decision Processes* 65 (1996): 272–92; "A Visceral Account of Addiction" in *GH*, pp. 235–64; "Will-Power: A Decision Theorist's Perspective," *LP* 19 (1999): 51–76.

44. Cf. Elster, *SF*, pp. 31–35.

45. Cf. George Loewenstein, Daniel Nagin and Raymond Paternoster, "The Effect of Sexual Arousal on Predictions of Sexual Forcefulness," *Journal of Research in Crime and Delinquency* 3–4 (1997): 443–73.

suggests is that the tendency to discount future goods in one's present motivation may be a product of the influence of visceral factors. Before the craving for the drug sets in at 7 P.M. tonight, the agent expects herself to be more motivated by the desire to have a sober morning the next day only because she underestimates the degree to which she will be moved by her visceral craving when it sets in. Hyperbolic temporal discounting, then, under Loewenstein and Elster's view, is a symptom of addiction rather than a defining feature of it.⁴⁶

This view provides a natural way of distinguishing addictions from compulsions that differs importantly from the way in which Ainslie drew the distinction. Since Ainslie makes no principled distinction between motives—all of the relevant motivationally effective mental states are preferences—he distinguishes addiction and compulsion through appeal to temporal factors: both are periodic cravings, which differ only with respect to the length of time between cravings. According to the visceral factors position, however, one distinctive feature of addictions is their appetitive nature and in this they might differ importantly from compulsions. The driving motive of a compulsive handwasher, for instance, might be to rid herself of guilt, while the driving motivation of an addicted smoker is an appetite much like her appetite for water or air.⁴⁷

Loewenstein's and Elster's positions do not differ substantially from the view adopted by Gary Watson in his important 1975 paper "Free Agency."⁴⁸ Although he has repudiated some aspects of the position since,⁴⁹ Watson there argued that there are two distinct types of motivation corresponding to the two "parts of the soul" described by Plato: the appetitive and the rational. Watson went on to claim that in addiction, appetite wins the battle with evaluation in guiding the agent's behavior, and that this fact is the crucial feature of addiction by virtue of which it undermines responsible agency. This position encounters a formidable problem. Under the account, there appears to

46. Elster, then, would agree that a tendency to discount the future can count as irrational if it is a product of a nonrational, visceral motive, even though he thinks that tendencies to discount the future cannot count as irrational merely because of their "steepness," or some other feature of the temporal discounting function.

47. Things are complicated by the fact that compulsions will often "hijack" appetites. So, for instance, the compulsive overeater may experience the compulsive desire as an appetite. Still, in such cases the deepest motivation for the behavior is not appetitive.

48. Gary Watson, "Free Agency," *Journal of Philosophy* 72 (1975): 205–20.

49. Idem, "Free Action and Free Will," *Mind* 96 (1987): 149–50.

be no meaningful motivational distinction between unfree choice and weak-willed choice. Both are rightly described as the victory of the appetitive over the evaluative the visceral over the cognitively mediated, to use Loewenstein's terms.

The problem has immediate implications with regard to responsibility for behavior stemming from addiction. Compulsion genuinely excuses from responsibility; e.g., someone with Tourette's syndrome who blurts out obscenities at a fancy dinner party deserves no rebuke. But weakness only diminishes responsibility, without eliminating it entirely. Someone who gives in to irrational impulses is certainly less to blame than someone who coldly and rationally plots objectionable conduct, but she is, nonetheless, to some degree blameworthy. The distinction between first and second degree murder, for instance, is not without a rationale. It is then imperative for a theory of addiction to tell us whether addicts are to be classified as compulsive or weak, and the theory advocated by Loewenstein, Elster and by Watson in the mid-1970s does not seem, at first glance at least, to do the trick.⁵⁰

One might try to overcome this problem by claiming that the difference between the weak-willed and the unfree is that the weak-willed remain capable of guiding their behavior in accordance with their evaluative judgments, while the unfree do not. As Watson put the suggestion in his 1977 follow-up paper, "[W]e are inclined to contrast weakness and compulsion like so: in the case of compulsive acts, it is not so much that the will is too weak as that the contrary motivation is too strong; whereas, in weakness of will properly so-called, it is not that the contrary motivation is too strong, but that the will is too weak."⁵¹ However, Watson argues persuasively against this way of contrasting the unfree and the weak-willed, when the term "strength", as applied to the will or to desire, is understood to be identifying a degree of causal force (*PR*, pp. 326-29). The problem is clearest when the distinction between strong and weak desires is made in the following way: strong desires win out over evaluative judgments and weak de-

50. This problem is brought out very nicely in Kadri Vihvelin, "Stop Me Before I Kill Again," *Philosophical Studies* 75 (1994): 115-48. See esp. pp. 124-30.

51. Gary Watson, "Skepticism About Weakness of Will," *Philosophical Review* 86 (1977): 327, henceforth *PR*. In this remark, Watson uses the term "the will" to refer to the agent's "practical judgment" (*ibid.*), or her evaluation of what is, all things considered, best for her to do.

sires do not. Under this analysis, to deem a desire to be “strong” is a *post hoc* way of noting that the agent acted in accordance with it instead of in accordance with her evaluative judgment. But under this account of “strength” there still remains no meaningful distinction between those whose freedom is diminished by their desires and those who give in to them only through weakness; in both cases, the appetite, or the visceral factor, was “too strong” to resist.

There are, to be sure, subtler ways of drawing the distinction between strong and weak desires. In his recent work, Watson has considered, for instance, efforts to do so by appealing to the agent’s susceptibility to countervailing considerations.⁵² For instance, a desire is strong enough to count as compulsive, on such an account, if the agent would still act as it dictates even if given good reason not to. Watson has pointed out that instances in which agents turn away from a particular course of conduct when given a reason to are only instances of weakness rather than compulsion if it can be shown that the agent turns away from the course of conduct specified by her desire because she recognizes the reasons to do so and responds to them appropriately. For instance, the fact that a heroin addict would not take steps to satisfy her craving for heroin should she have to swim to do so doesn’t show that she chooses heroin weakly rather than compulsively, since she might be a hydrophobe. The response to reasons to act counter to the dictate of a desire must, itself, not be compulsively motivated.⁵³ But if we were able to specify what it was to recognize countervailing reasons and respond to them appropriately, rather than in the way that the hydrophobe does, we would already have a good test for determining whether or not the agent was

52. As Watson points out, proposals of this sort can be found in the literature. Cf. John Fischer, *The Metaphysics of Free Will* (Oxford: Blackwell, 1994), p. 94.

53. Alfred Mele, “Irresistible Desires” in *Nous* 24 (1990): 455–72, makes a similar point in discussion of Wright Neely’s closely related account of irresistible desire. Neely says that a desire is irresistible if and only if an agent who recognized a good and sufficient reason not to act on it would still act on it (“Freedom and Desire,” *Philosophical Review* 83 [1974]: 32–54). But Mele points out that an agent’s desire would then count as irresistible if the recognition of a good and sufficient reason not to act on it gave him a fatal heart attack (see Mele, “Irresistible Desires,” p. 456). The point is very similar to Watson’s: any counterfactual test must specify that the reason–action relation is normal in the counterfactual circumstance. But if we could specify what normality of this sort consists in we wouldn’t need the counterfactual test in the first place.

responding weakly or compulsively when she acts contrary to what she judges to be best and would not require an appeal to responsiveness to countervailing reasons. That is, a test for determining whether or not a person acted compulsively that appeals to *noncompulsive* responsiveness to countervailing reasons is viciously circular, but without such a qualification the test seems not to draw the compulsive-weak lines in the way it should.⁵⁴

However, in response to concerns of this sort Elster and, in a similar way, John Fischer and Mark Ravizza have offered subtler ways of identifying compulsion-inducing desires by appeal to sensitivity to countervailing reasons. Both Elster, on the one hand, and Fischer and Ravizza, on the other, suggest that agents must exhibit a willingness to turn away from desire when given reason to do so *in a consistent and coherent pattern*, if her acting in accordance with that desire is a manifestation of weakness rather than compulsion.⁵⁵ The problem, that is, with the hydrophobic heroin addict is that she would not choose heroin if it required her swimming, but would choose it if it required her to suffer other evils that are, by all rational measures, just as bad or worse. The remaining problem, however, is to specify which patterns of response to incentives are rational and which are not. It would be

54. Gary Watson, "Disordered Appetites," pp. 7–9. Here is another way to see the problem that Watson is raising: An analysis of the weak-compulsive distinction through appeal to susceptibility to countervailing reasons would have to overcome the fact that addicts are often susceptible to some countervailing reasons, even if they are not as susceptible as the rest of us; a sufficiently severe threat might keep the addict on course even if a weaker threat, sufficient to keep the unaddicted from using, wouldn't do the trick. But what is the difference between a weak countervailing reason and a strong one? The worry is that the distinction between strong and weak countervailing reasons is just the distinction between the compulsive and the weak reappearing in a different place.

55. Elster puts this test entirely in monetary terms: If the agent would act contrary to her desire when offered a certain amount of money, or any greater amount, then she merely acts weakly. However, Elster proposes this as only a sufficient condition. (See Elster, *SF*, pp. 140–41.) Alternative sufficient conditions could be devised for agents who don't care much about money, or who have reasons not to think better of more of it. Whatever the reasons are that would draw an agent away from her desire, she must respond to such reasons in a coherent pattern. Someone might, for instance, choose contrary to her desire if offered \$1,000 to do so, but not if offered \$10,000 while still showing herself to be responsive to reasons; say she knows that after accepting money above \$9,999 she would be audited by the IRS. In accordance with examples of this sort, a more general recipe for the construction of sufficient conditions is supplied in John Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: Cambridge University Press, 1998), pp. 65–91.

circular to appeal, at this point, to responsiveness to countervailing reasons. But neither Elster nor Fischer and Ravizza provide an alternative account.⁵⁶ The worry is that any adequate account will involve a circular appeal to an unanalyzed distinction between compulsive and merely weak choice making. There is surely more work to be done here, but in the absence of a noncircular specification of a distinction between rational and irrational patterns of choice making we are still in need of an account of “strength” of desire that will serve to distinguish between the compulsive and the merely weak.⁵⁷

In his 1977 paper, and also in his most recent work, Watson suggests that there is truth to the claim that the weak have weak wills and the compulsive have strong motives if “strength” of motivation is understood normatively; to say that one suffers under desires so strong as to induce compulsion is simply to say that those desires are not ones we expect morally upright agents to deny. The reverse is true in the case of weakness.⁵⁸ Watson’s proposal involves a further, and yet more radical departure from the approach of rational choice theory than anything advocated even by Loewenstein and Elster. All the theorists whose views are under discussion here, with the exception of Watson, assume that the question “How does the behavior of an addict differ

56. In correspondence, Fischer has suggested that this problem could be overcome through appeal to a criterion of “mechanism individuation.” For Fischer and Ravizza, that is, the relevant question is whether an agent who acted from the very same mechanism on which she actually acted would be moved by countervailing considerations. But, Fischer is suggesting, cases like the hydrophobic heroin addict might involve a switch in motivational mechanism. The mechanism that leads her to actually take heroin, that is, is not the same as the mechanism that leads her to choose not to when she would have to endure the water to do so. Fischer is well aware that there are serious challenges that an adequate criterion of mechanism individuation would have to meet. For our purposes here it is necessary to note only one: whatever criterion one produces it must not distinguish between the actual mechanism on which the heroin addict acts and the hydrophobic mechanism that would lead her not to choose to take heroin by appealing to the fact that the latter, and not the former, leads to compulsively made choices. To specify the criterion of individuation in this way would be to argue in the same viciously circular manner identified by Watson.

57. An important recent discussion of strength of motivation is Mele, “Motivational Strength,” in *Nous* 32 (1998): 23–36. See also Mele, “Strength of Motivation and Being in Control: Learning from Libet,” in *American Philosophical Quarterly* 34 (1997): 319–33.

58. See Watson, “Skepticism About Weakness of Will,” p. 331, and idem “Disordered Appetites,” p. 11. Although Watson does make this proposal, he adds that he is “sure it is unsatisfactory as it stands.” (Ibid., p. 11.)

from the behavior of the unaddicted?" takes priority over the question "Does addiction diminish responsibility?" They think that the first question can be answered without answering the second, although not vice versa. However, by claiming that the compulsive–weak distinction is a normative one, Watson is denying this. The theory of addiction (the model of the behavior of the addict) cannot tell us, all by itself, whether the agent is to be excused entirely from responsibility (as she would be if she is acting unfreely) or if she is to be assigned merely diminished responsibility (as she would be if she is acting merely irrationally). In order to draw the compulsive–weak distinction, we must ask and answer a normative question; we must ask whether we think it objectionable that she acted contrary to her evaluative judgment in this instance. Is that fact indicative of a failing on her part? If it is then she is weak; if it is not, then she is compulsive.⁵⁹

It may be correct to draw the compulsive–weak distinction normatively. However, notice that doing so doesn't yet help us to understand our initial dilemma: we started by asking what it was about addiction that diminished the responsibility of the afflicted agent for conduct stemming from her affliction. The answer supplied by a "visceral factors" account is this: addicts are acting from motives that are not products of rational capacities; their conduct is controlled by the appetitive, rather than the evaluative, part of the soul. But when we are asked what the implications are for responsibility—does this imply that addicts are weak or does it imply they are compulsive?—we are told that the compulsive–weak distinction is drawn through determining which way of regarding the addict involves a correct assessment of responsibility. The account of addiction, then, is not providing justification for treating addicts as compulsive rather than weak, or weak rather than compulsive; instead, the account remains silent on the question, leaving it to be dictated by which responsibility assessment is appropriate. In the absence, then, of another, nonnormative way of

59. Watson's way of drawing the weak–compulsive contrast fits nicely with the view of freedom of will recently offered in Gideon Yaffe, "Free Will and Agency at Its Best" in *Philosophical Perspectives*, 14: *Action and Freedom* (Oxford: Oxford University Press, 2000), pp. 203–29. There I suggest that freedom of will is a "thick" evaluative concept. That is, no analysis of the concept in purely descriptive terms will be satisfactory, but, instead, those to whom the concept applies must be thought of as exemplifying something that is intrinsically evaluative: agency at its best.

drawing the distinction between strong and weak appetites and strong and weak rational motives, the visceral factor conception of motivation does not help us to determine what to say about the addict's responsibility.

It is important to see the status of this critical point. What is being identified is not a point of incoherence in the claim that the compulsive-weak distinction is a normative one. Rather, the truth of the claim would suggest that the question of the degree to which addicts are or are not to be held responsible for their behavior cannot be answered simply by determining what is happening when the addict acts, and how the addict's exercises of agency differ from those of the unaddicted. But if this isn't the way to approach the question, what is? It is natural to look into the metaphysics of agency in order to find the bridge to normative concepts. We look, for instance, for a metaphysical test for determining whether or not an agent could have done otherwise, thinking that we are thereby identifying a necessary condition of moral responsibility. But if Watson is right, metaphysics alone are not sufficient for helping us to determine whether or not the addict acts compulsively or weakly, and so we must start where we originally thought we would end: with an answer to the normative question of the addict's degree of responsibility. But how is it possible to start there? What does starting there amount to?

In his most recent work, Watson has examined the legal notion of duress as a means of investigating the distinction between compulsion and weakness.⁶⁰ Watson points out that in the law defendants are rarely able to successfully employ a duress defense if they found themselves in the duress-producing circumstances as a result of voluntary conduct on their parts. A robber who injures a store owner when the owner pulls a gun during a robbery cannot defend himself on the grounds that the owner was threatening him with lethal force, and

60. Gary Watson, "Excusing Addiction," *LP* 18 (1999): 605ff. In developing his view there, Watson draws heavily on Dan Kahan and Martha Nussbaum, "Two Conceptions of Emotion in the Criminal Law," *Columbia Law Review* 96 (1996): 269-374. Patricia Green-span, "Behavior Control and Freedom of Action" in *Moral Responsibility*, John Fischer ed. (Ithaca: Cornell University Press, 1986), pp. 191-204, argues that those who have motivations like the addict are unfree because of duress. Kadri Vihvelin, "Stop Me Before I Kill Again," pp. 120-24, argues that whether or not duress of this sort undermines freedom, it does not undermine moral responsibility.

therefore he was under duress. This, despite the fact that an innocent bystander who injures the owner when the owner threatens him with the gun, mistaking him for one of the robbers, can launch a duress defense. These practices are usually justified on the grounds that the robber, and not the innocent bystander, got himself into the duress-producing circumstance voluntarily.⁶¹ Since the link between voluntary drug-use and later addiction is just as strong as the link between voluntarily entering into a robbery and finding oneself threatened with a gun, it follows that addicts face formidable obstacles in mounting a defense based on duress. Watson points out, however, that there are many conditions in which people find themselves that, like addiction, are entered into as a result of voluntary conduct and that do form the basis of a defense of duress. For instance, someone who lies to police to protect her child from arrest can use a duress defense against a charge of obstruction of justice. In cases of this nature, the “dependency” on the child might be no less voluntarily acquired than most drug dependencies. Watson argues that the difference between cases of this nature and most cases of addiction is that we often take dependencies on children (and parents, spouses, and the like) to be of great worth and importance, and so we excuse those who act from such dependencies and contrary to the letter of the law on the grounds that they were under duress. What this implies is that the legal concept of duress is normatively loaded. The law is recognizing that we cannot draw the distinction between those circumstances that do and do not produce duress without appeal to our evaluative assessments of the circumstances that are putatively duress-producing.

The point can be extended to the distinction between weakness and compulsion. Whether behavior stemming from addiction is weak or compulsive is not merely a matter of the psychological and metaphysical facts about the causal etiology of such behavior. Also relevant is our evaluative assessment of the dependency to the drug.⁶² If Watson is right about this, then he is taking one step towards providing a sub-

61. Notice that the robber, in this example, didn't literally choose to be threatened with a gun, and, conversely, the bystander very well may have voluntarily walked into the store at the wrong moment. So, what link between the duress-producing circumstances and the agent's voluntary conduct is needed to invalidate a duress defense is a complex matter.

62. Watson, “Excusing Addiction,” p. 616; “Disordered Appetites,” p. 18.

stantive test for determining whether addicts are suffering under compulsion or are guilty of weakness.⁶³ Watson's point helps to explain why, for instance, conduct motivated by appetites for things that we view as necessary for a flourishing life is often thought to be compelled, while similar conduct springing from other sorts of appetites is not. Why does dependency on air, for instance, provide the basis for a defense of duress—people have the right to do many things in order to prevent themselves from being suffocated that they wouldn't ordinarily have the right to do—while dependency on nicotine does not? Watson's answer is that in the former case we think of the dependency as intertwined with a healthy life, while in the latter case we do not. Unfortunately, however, Watson's point will only take us so far in understanding the impact of addiction on responsibility. After all, there are a variety of different evaluative assessments of a dependency that we might make. The famous mathematician Paul Erdős took amphetamines in huge doses and believed himself to be incapable of creative mathematics without their assistance. In fact, mathematical progress was the only thing that really mattered to Erdős and so, arguably anyway, his amphetamine dependency was intertwined with a flourishing life for him in much the way that a dependency on a spouse or a child is for the rest of us.⁶⁴ But it is far from clear that Erdős's responsibility for seemingly objectionable conduct would be any less than that rightly attributed to a person who acted in the same way in order to prevent amphetamine deprivation but for whom the dependency was not part of a flourishing life. So, while Watson is certainly right that an evaluative assessment of dependency is one of the factors involved in drawing the compulsion–weakness distinction, it is far from clear what role such assessments play.

Nonetheless, Watson's work points the way towards a promising avenue for further work on addiction: we can certainly go farther in

63. R. Jay Wallace, "Addiction as Defect of Will," pp. 627–28, suggests that this sense in which the concept of compulsion is sensitive to normative assessments is of little significance. Wallace is certainly right that evaluative judgments of the sort that Watson appeals to are not sufficient for distinguishing compulsions from mere weaknesses. However, at issue for Watson at least, although not for Wallace, is not sufficiency, but necessity.

64. The definitive biography of Erdős is Paul Hoffman, *The Man Who Loved Only Numbers* (New York: Hyperion, 1998).

identifying the range of normative assessments that enter into a judgment of the applicability of the concepts of compulsion and weakness, and the role that such assessments play in applying those concepts. Such work will help us to go farther in determining the degree to which addicts manifest one of these faults as opposed to the other. However, such work will not put us in position to supply a simple answer to the question of the degree, if any, to which the responsibility of an addict is diminished; we won't find ourselves with a decision procedure for answering that question. What such work will provide is a more accurate picture of the way in which our assessments of the responsibility of addicts are linked to other normative evaluations of agents and the circumstances under which they labor. It is possible that the best picture of the addict's responsibility is more like an Impressionist painting than a blueprint.

IV. CONCLUSION

Recent contributions to the philosophical literature on addiction can be classified by the degree to which they depart from the model of human motivation provided by traditional rational choice theory. At one extreme is work being done by theorists who do not depart at all from that model. Such work does not provide a satisfactory account of the impact of addiction on responsibility, since it can only be used to provide such an account by making questionable assumptions about what agent-rationality amounts to. Still, the application of rational choice theory to the case of addiction serves as a fascinating guide to the traps in which a rational agent can find herself. Sometimes our rationality can be our affliction; in exercising it, we can end up worse off than we would have been had we acted irrationally. Of course, rationality cannot provide the ladder for climbing out of this condition since we cannot, by definition, have decisive reason to act irrationally.⁶⁵

Departures from the traditional rational choice model, in moving away from the hydraulic conception of the causal force of preferences, come up against deep questions about the nature of rational motiva-

65. This kind of line of thought is explored in an interesting and entertaining way in Thomas Schelling, "Rationally Coping with Lapses from Rationality" in *GH*, pp. 265–84.

tion and how it differs from other forms. While Ainslie would not agree with this point, his work indicates precisely how much of a departure is really necessary in order to account for the kind of struggle that addiction involves. Insofar as exercises of will serve as a corrective to the irrationality of preferences over time that Ainslie identifies, it must be that the will serves as a source of motivation differing intrinsically from the preference-based motivation that both Ainslie and the rational choice theorists take to be the only sort.

How much more of a departure from the rational choice model the recognition of this point really requires remains, to some degree, an open question. Elster, Loewenstein, and prominent thinkers in the free will literature, such as Fischer and Ravizza, aim to provide normatively neutral tests for determining the difference between rationally entwined forms of motivation and the sort to which addicts, arguably, are subject. These efforts may succeed. But they may not, or even if they do, they may end up converging on the normatively loaded conception of addictive motivation to which Watson is drawn. In this area, anyway, important and foundational conceptual work remains to be done.