ON THE RATIONALIST SOLUTION TO GREGORY KAVKA'S TOXIN PUZZLE

BY

KEN LEVY

Abstract: Gregory Kavka's 'Toxin Puzzle' suggests that I cannot intend to perform a counter-preferential action A even if I have a strong self-interested reason to form this intention. The 'Rationalist Solution,' however, suggests that I *can* form this intention. For even though it is counter-preferential, A-ing is actually rational given that the intention behind it is rational. Two arguments are offered for this proposition that the rationality of the intention to A transfers to A-ing itself: the 'Self-Promise Argument' and David Gauthier's 'Rational Self-Interest Argument.' But both arguments – and therefore the Rationalist Solution – fail. The Self-Promise Argument fails because my intention to A does not constitute a promise to myself that I am obligated to honor. And Gauthier's Rational Self-Interest Argument fails to rule out the possibility of rational irrationality.

1. Introduction

Suppose that there is a certain toxin that makes people suffer pain for only one day. Suppose also that a trustworthy billionaire offers me $1m if I will merely form the intention by midnight tonight of drinking the toxin tomorrow afternoon. So I do not actually have to drink the toxin tomorrow afternoon to win the money. Again, I need merely to intend tonight to drink the toxin tomorrow afternoon. If the very advanced brain-scanner confirms tonight that I have formed the intention of drinking the toxin tomorrow afternoon, then – whether or not I actually end up drinking the toxin tomorrow afternoon – the billionaire will give me the money tomorrow morning.

It seems at first that the money is 'in the bank.' After all, I do not even have to drink the toxin. All I need to do is *intend* to drink the toxin. And

Pacific Philosophical Quarterly 90 (2009) 267–289

intending is easier than doing. Indeed, it seems that I can form virtually *any* intention for a sizeable enough reward. But Gregory Kavka (1983; 1984, pp. 156–57) argues that this intuition is false, that the intention cannot be formed and therefore that the money cannot be won. 'Kavka's Argument' goes like this:

(1) Whether or not I have won the $1m by tomorrow morning does not at all depend on what I do tomorrow afternoon. By tomorrow afternoon, I will already have either won the money or not.
(2) ∴ I will have no good reason to drink the toxin tomorrow afternoon.
(3) I will also have a (substantial) reason not to drink the toxin tomorrow afternoon – namely, pain!
(4) I will recognize this tonight. I will recognize tonight – as I recognize now – that come tomorrow afternoon, I will have no good reason to, and a (substantial) reason not to, drink the toxin.
(5) I cannot intend to do what I (rightly) believe to be fundamentally irrational – i.e. something that I have no good reason to do and a (substantial) reason not to do.[1]
(6) ∴ I will not be able to form the intention tonight of drinking the toxin tomorrow.
(7) ∴ I will not be able to win the money.[2]

(6) and (7) are counter-intuitive. They both conflict with our intuition that I *can* form the intention – especially for $1m.[3,4] For this reason, Kavka refers to this hypothetical situation as a *puzzle* – the 'Toxin Puzzle' (henceforth, 'TP').

In the end, there are three and only three strategies by which the Toxin Puzzle can be solved – that is, by which it can be shown that something is wrong with Kavka's Argument and therefore that, contrary to (6) and (7), I *can* intend tonight to drink the toxin tomorrow and thereby win the money. It must be shown either:

(8) that I can form the intention to drink the toxin;
(9) that I can drink the toxin; or
(10) that it is rational – i.e. that I have a sufficiently good reason, all things considered – to drink the toxin.

The first of these three strategies (i.e. (8)) would be a direct route to ~(6) – i.e. the conclusion that I can intend tonight to drink the toxin tomorrow. The second and third strategies are indirect routes to ~(6) insofar as both require additional assumptions. The second strategy (i.e. (9)) proceeds to ~(6) through the assumption:

(11) if I can drink the toxin, then I can intend to drink the toxin.

And the third strategy (i.e. (10)) proceeds to ~(6) through:

(12) if it is rational for me to drink the toxin, then I can intend to drink the toxin.[5]

In this paper, I will investigate the third strategy, which I will refer to more transparently as the 'Rationalist Solution' to TP. The Rationalist Solution is arguably the most popular approach to TP, at least in the philosophical literature.[6,7] I will ultimately argue, however, that the Rationalist Solution fails.

2. *The rationalist solution*

In this section, I will explain what the Rationalist Solution involves in more detail. The key objective of the Rationalist Solution is to demonstrate (10) – i.e. that drinking the toxin is rational. For this proposition in conjunction with (12) entails ~(6) – i.e. that I can intend to drink the toxin. And this result would help to show that Kavka's Argument fails.

2.1. WHY (10) CANNOT BE WEAKENED

It might initially be objected that the Rationalist Solution need not show anything as strong as (10); that in order to defeat Kavka's Argument, the Rationalist Solution need not show that I have a sufficiently good reason, all things considered, to drink the toxin. Instead, it needs to show only that I have *a* reason to drink the toxin. In response to this objection, however, merely having *a* reason to drink the toxin will not be sufficient to establish that I can drink the toxin, which is the ultimate goal of the Rationalist Solution. It will not succeed because this reason for drinking the toxin may be overridden by other, better reasons for refraining from drinking the toxin. These better reasons, then, would help to show not merely that drinking the toxin is irrational but also – and possibly therefore – that drinking the toxin is impossible.[8] So in order for the Rationalist Solution to prevail, it must be shown – as (10) represents – not merely that I have a reason for drinking the toxin but that this reason is, all things considered, stronger than the set of reasons for refraining from drinking the toxin.[9]

2.2. WHY MY INTENTION TO DRINK THE TOXIN IS RATIONAL

The Rationalist Solution assumes that my intention to drink the toxin is rational. And the reason that it is rational is because of the reasonably expected *outcome* of my having or forming the intention.[10]

I ultimately face only two options tonight – what Kavka (1978, p. 295) refers to as a 'cruel dilemma' and what McClennen (1990, p. 231) refers to as an intra-personal 'coordination problem.' I may either (a) intend to

maximize my utility tomorrow afternoon or (b) intend not to maximize my utility tomorrow afternoon. And the way in which a rational utility-maximizer such as myself should make this choice is by seeing which of these two options will likely yield the overall maximum utility. If I adopt (a), then I will not intend to drink the toxin tomorrow afternoon and will therefore not win the $1m. But I will also avert a day of toxin-induced pain. If I adopt (b), then I will intend to drink the toxin tomorrow afternoon and therefore will win the $1m. But by intending to drink the toxin tomorrow afternoon, I am in effect intending not to act as a utility-maximizer tomorrow afternoon. And, let us suppose, this intention significantly increases the probability that I will actually drink the toxin and suffer a day of pain.[11]

Given these different (reasonably likely) consequences of my choice, (b) is the (more) rational choice for me to make.[12] The net gain of (b) – i.e. $1m plus (very possibly) one day of toxin-induced pain – is greater than the net gain of (a) – i.e. no money plus no toxin-induced pain. The $1m more than compensates for the (reasonably likely) day of pain. So in this particular situation, the (more) rational thing to 'do' is to be 'rationally irrational' – i.e. to intend tonight *not* to act as an ideally rational utility-maximizer tomorrow afternoon.[13] I should instead resolve tonight to minimize it tomorrow afternoon – at least with respect to this particular choice of drinking or not drinking the toxin. This is one of the few situations in which it can be shown, ironically enough, that a utility-maximizer should intend to be self-destructive.[14,15]

The intention to drink the toxin is rational – i.e. utility-maximizing – *despite the fact* that there are *no* good reasons directly supporting the intended action (drinking the toxin) *and* a (substantial) reason *against* this action. What makes the intention rational *does not* derive from the action itself or from any reasons supporting the action.[16] What makes the intention rational has nothing to do with the action's nature or expected consequences. Rather, the rationality of the intention derives solely from one of its 'autonomous effects'[17] – namely, the fact that I will earn a substantial profit merely from forming it.[18]

One might argue that the proposition that it is rational to intend to drink the toxin is all that is needed to solve TP. For if it is rational to intend to drink the toxin, then it must be possible to intend to drink the toxin. But the possibility of forming an intention does not necessarily follow from its being rational. And it is TP itself that may help to demonstrate this point. So it would beg the question against Kavka's Argument to assume otherwise.

2.3. FORWARD- AND BACKWARD-LOOKING REASONS

The Rationalist Solution next makes a distinction between 'forward-looking' or 'outcome-oriented' reasons and 'backward-looking' reasons.

On the one hand, an outcome-oriented reason for performing some action A is a consideration in favor of A-ing based solely on some expected good consequence(s) of A, where *good* is defined subjectively – i.e. as whatever the agent would regard as positive or favorable (e.g., pleasure, profit, or promotion). On the other hand, a backward-looking reason for A-ing is a consideration in favor of A-ing *not* based on some expected consequence of A but rather based on something that *precedes* A. Backward-looking reasons include such things as my having made a promise and my having behaved badly. The former would help to justify my now keeping the promise, and the latter would justify my now being blamed or punished.[19]

Naturally, as with every philosophical distinction, the distinction between forward- and backward-looking breaks down in certain cases. It blurs, for example, with respect to rule-utilitarian type reasons. For a rule-utilitarian reason tends to be both forward- and backward-looking. It is backward-looking insofar as it suggests that I should A even though A may very well yield an overall balance of negative consequences simply because that it is what the prior-established rule says that I should do. And it is outcome-oriented insofar as it suggests that I should follow the rule simply because rule-following will produce the best consequences in the long run. (I will discuss rule-utilitarianism further in section 3.)

2.4. THE RATIONAL ACTION PRINCIPLE (RAP)

Given this distinction between forward- and backward-looking reasons, the Rationalist Solution then holds that I *can* intend tonight to drink the toxin tomorrow even with full knowledge that I will have absolutely no outcome-oriented reason to, and a (substantial) outcome-oriented reason not to, drink the toxin. For the inference from (1) to (2) in Kavka's Argument is invalid. While it is certainly true that I have no *outcome-oriented* reason for drinking the toxin, it does not follow that I have *no* good reason whatsoever to drink the toxin. On the contrary, I *do* still have a good reason to drink the toxin – a *backward-looking* reason. My backward-looking reason for drinking the toxin is the fact that (a) I intended the night before to drink the toxin, (b) this intention was itself rational (see section 2.2), and (c) the circumstances surrounding the choice to drink or to refrain from drinking the toxin that I anticipated when I formed the intention are identical to the circumstances that actually obtain.

This point implicitly depends on the following more general principle – call it the 'Rational Action Principle' ('RAP'): if it is rational for me to intend to A at future time t when I anticipate circumstances C at t, then, whatever my preferences at t, it is rational for me to A at t if C

obtain at t.[20] Normally, a given intention of mine is rational because the intended action is rational. The rationality of the intended action is the ground or source of the rationality of the intention.[21] RAP, however, suggests that, in this peculiar case, the reverse is true: my action is rational because my intention was rational.[22] The rationality of the intention is the ground of the rationality of the intended action. The fact that my prior intention to drink is rational itself constitutes a good *reason* to drink.[23] In this way, my 'intention-based' reasons for drinking the toxin may 'rationally override' my 'preference-based' reasons for refraining from drinking the toxin.[24]

It is fairly easy to see why philosophers would adopt RAP in this context. Gauthier, for example, shares our intuition that I can form the intention to drink. But Gauthier also subscribes to step (5) in Kavka's Argument, which holds that it is impossible to intend to perform an action that I know or believe to be fundamentally irrational. So the only way to reconcile these two propositions in this context is by demonstrating that I may convince myself that drinking the toxin is rational. And, arguably, the easiest way to do that is by predicating the rationality of the action on the rationality of the intention.[25]

2.5. THE ANTI-RAP OBJECTION

One might propose the following objection against RAP – call it the 'Anti-RAP Objection.' The Anti-RAP Objection suggests that the rationality of the intention does not transfer to the action, as is usually the case. What makes my intention rational – namely, the fact that forming it will help me to win $1m – does not also make my actually drinking the toxin rational. Drinking the toxin will not help me to win any money at all, and it will cause me severe pain. Therefore the mere fact that my intending to drink the toxin is rational is not a good enough reason for me actually to drink the toxin. In this rare situation, my reason for forming the intention does not also amount to a reason for actually performing the intended action.[26] We simply cannot 'bootstrap' the rationality of a self-destructive act from the rationality of a self-benefiting intention.[27] Even though the intention to drink the toxin is rational, actually drinking the toxin is still irrational.[28]

3. Two defenses of RAP: the self-promise argument and the rational self-interest argument

The RAP supporter must hold that, contrary to the Anti-RAP Objection, the rationality of my intention to drink the toxin *does* transfer to my

actually drinking the toxin and thereby makes *it* rational too. This position may be defended in two different ways. The first defense is the 'Self-Promise Argument.' The second defense is David Gauthier's 'Rational Self-Interest Argument.' In section 3.1, I will discuss and reject the Self-Promise Argument. In section 3.2, I will discuss the Rational Self-Interest Argument. In section 3.3, I will argue that the Rational Self-Interest Argument ultimately fails to show that my drinking the toxin is indeed rational.

3.1. THE SELF-PROMISE ARGUMENT

According to the Self-Promise Argument, when I formed a rational intention to drink the toxin, I made a *promise* to myself. I promised myself that when the time came, I would act in accord with this intention, an intention that I realized might not carry as much weight with me at this later time as it did when I formed it. The mere fact that I have made a promise to myself imposes an obligation upon me to honor it. And this obligation is only strengthened by the fact that, in this particular situation, the promise was made for my benefit and actually helped to fulfill my best interests – namely, by making me $1m richer.[29]

To say that I have an obligation to drink the toxin is to say that it would be *wrong* for me to refrain from drinking the toxin. But what is the *nature* of this obligation and corresponding wrong? Despite first appearances, and despite what the proponent of the Rationalist Solution may *wish*, the obligation and wrong are *not* moral. When I formed the rational intention to drink the toxin, I did not impose a *moral* obligation on myself to drink the toxin. It would not be *morally* wrong for me now to refrain from drinking the toxin.

There are two reasons. The first reason is that the intention to drink the toxin is non-moral. It does not concern the welfare or rights of anybody or anything else. On the contrary, it was formed for the strictly *selfish* purpose of winning lots of money. It might be objected that my reasons for making a promise to myself cannot determine whether or not the promise imposes a moral obligation on me because the sheer fact of having made a promise – whether to myself or *to others* – imposes moral obligations on me no matter what reasons I have, moral or non-moral, for making this promise. In response to this objection, it is certainly true that I am presumptively morally obligated to fulfill promises to others; this is analytic. But because a promise *to myself* is *not* a promise *to others*, it does not necessarily involve any self-imposed *moral* obligations. We must therefore look to the intentions behind my self-promise to determine what the nature of these obligations are. If, for example, the intention behind my self-promise is to become a better person, then my obligation *is* moral and it *would be* morally wrong for me to break this promise. But if the intention

behind my self-promise is to become *richer*, as in the toxin situation, then the intention behind my self-promise – and therefore the obligation it imposes upon me – is *not* moral but rather purely selfish/self-interested/utilitarian.

The second reason why it would not be *morally* wrong for me to refrain from drinking the toxin is that it is tenuous at best to maintain either (a) that I have a moral obligation – and not just a self-interested imperative – to look out for my own (future) interests or (b) that I have a moral obligation to honor whatever I previously intended myself to do. Regarding (a), the reason that I break promises to myself much more easily and frequently than I break promises to others is precisely that the former are usually designed to serve my own self-interest rather than any moral imperative. As a result, whatever guilt I feel when I break these self-promises has a fundamentally different affect than the kind of guilt I feel when I break my promises to others. And a difference in affect is some evidence of the different kinds of moral content behind the two kinds of guilt. Regarding (b), I may have moral obligations to others around me. But it borders on the nonsensical to maintain that I have a moral obligation to honor a promise not merely to myself but to my past self – especially when, as in the toxin situation, my past self had motivations for making the promise *that have since been satisfied*. So I am not hurting myself, no less anybody else, by abandoning this promise.[30]

Instead, the most that the proponent of the Rationalist Solution can say is that I have a *utilitarian* or *practical*, not moral, obligation to follow my previously formed intention to drink the toxin. My intention to drink was formed for the very practical reason of maximizing my overall self-interest. This intention is still a *backward-looking* reason because what justifies my acting on it is *not* my *present* self-interest per se but rather the fact that it *was designed* to fulfill my overall (or net or long-term) self-interest. Still, in the end, this backward-looking reason is outweighed by my forward-looking reason of minimizing pain. For the object of the latter is to maximize my overall self-interest as well. So this object cancels out. And all that is left to determine which reason is stronger is a comparison of which reason does a better job of achieving this goal. Clearly, the forward-looking reason of minimizing pain wins *this* contest because the one and only consequential difference between them is entirely in the forward-looking reason's favor. While acting on this forward-looking reason would spare me a day of pain, adopting the backward-looking reason of paying homage to my previously formed rational intention would do precisely the opposite. (Incidentally, this point would be true not only on an act-utilitarian account but also a rule-utilitarian account as long as the rule was defined narrowly enough to cover only toxin-type situations – i.e. situations in which I have a rational intention to perform a somewhat self-destructive action.)

3.2. THE RATIONAL SELF-INTEREST ARGUMENT

The Rational Self-Interest Argument picks up where the Self-Promise Argument leaves off. According to the Rational Self-Interest Argument, the last paragraph above is incorrect. I actually do *better* overall by acting on my intention to drink the toxin than I do by refraining from drinking the toxin. And, for this reason, my drinking the toxin is indeed rational.

For the sake of convenience, call the plan or 'course of action' in which I intend to drink the toxin and subsequently drink the toxin the 'Course' and the course of action in which I do *not* intend to drink the toxin and so do *not* drink the toxin the 'Alternative Course.' Given this distinction, the Rational Self-Interest Argument proceeds as follows:

(13) I will benefit more overall from adopting the Course than I would from adopting the Alternative Course. While the Course will cost me one day of pain and the Alternative Course would not, it will also make me $1m richer than would the Alternative Course. And $1m plus one day of pain is overall more beneficial to me than no day of pain and no money. (See section 2.2.)

(14) ∴ It is in my rational self-interest to adopt the Course. [(13)]

(15) I am able to adopt the Course only if I intend to drink the toxin.

(16) I am able to intend to drink the toxin only if I believe that I will drink the toxin.[31]

(17) ∴ It is in my rational self-interest to believe that I will drink the toxin. I cannot adopt the Course without this belief. [(14), (15), (16)]

(18) I can believe that I will perform counter-preferential actions only if I believe that these actions are in my rational self-interest.[32]

(19) ∴ I can believe that I will drink the toxin only if I believe that drinking the toxin is in my rational self-interest. [(18)]

(20) ∴ My belief that drinking the toxin is in my rational self-interest is an essential part of the Course. I can adopt the Course only if I believe that drinking the toxin is in my rational self-interest. (Conversely, I can believe that drinking the toxin is in my self-interest only if I adopt the Course. For the only plausible reason to believe that drinking the toxin is in my self-interest is that it is an essential part of the course of action that will most benefit me.) [(17), (19)]

(21) ∴ My belief that it is in my rational self-interest to drink the toxin should persist *past* the point of my winning the $1m *all the way* up to tomorrow afternoon, when I am presented with the choice of drinking or not drinking the toxin. [(13), (20)]

(22) Assume I am a rational agent and that I have acquired my beliefs in this context through rational reflection.

(23) ∴ It is in my rational self-interest to drink the toxin.[33] [(21), (22)]

Two objections may be raised against the Rational Self-Interest Argument – specifically against the inference of (21) from (13) and (20).[34] First, one might challenge this inference by arguing that it wrongly assumes that there are only two options, the Course and the Alternative Course, when in fact there is a third – call it the 'Third Course.' In the Third Course, while I intend tonight to drink the toxin tomorrow, I change my mind tomorrow (after I have already won the $1m) and decide to refrain from drinking the toxin. *This* course of action seems to be more rational than the Course (not to mention the Alternative Course). For while I equally win $1m in both situations, the Third Course does not cost me a day of pain.

Gauthier offers two responses to this first challenge. First, if the Third Course occurs to me before I have actually formed the intention of drinking the toxin and therefore before I have won the $1m, then I may not be able to form the intention of drinking the toxin in the first place. So, as the Rational Self-Interest Argument itself argues, it is in my rational self-interest to reject the Third Course and go instead with the Course. Second, while it is certainly possible for me to form the intention to drink the toxin and then change my mind when the time comes to choose between drinking and refraining, this sequence of events does not qualify as a *course of action*.[35] A course of action is a plan that I end up executing, an intention that ultimately results in the intended action. And I cannot coherently plan to go back on my present intention. For then it is not really my intention in the first place. That is, there is not the third option of (a) convincing myself that I will drink the toxin and (b) secretly thinking that I will change my mind when actually presented with the toxin. For (a) and (b) are incompatible. If (b) is the case, then (a) is not. If I am secretly thinking that I will change my mind, then I am really *not* intending to drink the toxin. So even if intending to drink the toxin and then changing my mind were more rational than intending to drink the toxin and then drinking the toxin, the fact that the former does not constitute a course of action makes it possible for the latter to remain the (more) rational course of action.

The second objection against the inference of (21) from (13) and (20) concedes that the Course is more rational than the Alternative Course and therefore that it is more rational to intend to drink the toxin *and* drink the toxin than not to intend to drink the toxin to begin with. But it simply does not follow from this comparison that drinking the toxin is more rational than refraining from drinking the toxin. The comparison of courses of action does not dictate the comparison of actions themselves. Even if the Course is more rational than the Alternative Course, refraining from

drinking the toxin may still be more rational than drinking the toxin. Indeed, when we compare these two actions independently of the courses of action in which they are situated, the result is quite simple: while drinking the toxin will cost me a day of severe pain and not benefit me at all, refraining from drinking the toxin will not cost me anything. Contrary to the Rational Self-Interest Argument, then, it is still in my rational self-interest to refrain from drinking the toxin rather than drink the toxin.

In response to this second objection, Gauthier argues that it is simply improper to measure, as the act-utilitarian or 'orthodox economist' does,[36] the rationality of drinking the toxin against the rationality of the alternative action – i.e. refraining from drinking the toxin. Rather, when evaluating the rationality of drinking the toxin, we should first situate it within the larger course of action of which it is a part (again, the Course) and then measure the rationality of this entire course of action against the rationality of the Alternative Course. Once we do this, we find that even if drinking the toxin costs me more than refraining from drinking the toxin, it is still more rational. For (a) drinking the toxin is part of the Course, which is more rational than the Alternative Course; and (b) there are significant outcome-oriented reasons for carrying out more rational courses of action in general.[37]

According to Gauthier, the reason that the Alternative Course rather than the alternative action itself offers the proper contrast is because, as (b) above implies, rule-utilitarianism produces greater overall utility than act-utilitarianism.[38] Pursuing utility-maximizing *courses of action* produces greater overall utility in the long term than pursuing utility-maximizing *actions*, even if such pursuit sometimes requires performing actions that are likely to produce less utility than alternative actions. So the only reason that I would have to reconsider and/or abandon my plan to drink the toxin would be if, at the time that I am faced with the choice of drinking or not drinking the toxin, I have reason to believe that I will do worse by carrying out my plan to drink than I would have if I had not planned to drink in the first place. According to Gauthier, I do *not* have such a reason to reconsider and/or abandon my plan to drink the toxin because I still reasonably believe that I will do better to carry it out, given that it made me $1m richer, than I would have if I had not formed it in the first place, which would have left me just as poor as before.

3.3. GAUTHIER FAILS TO ELIMINATE THE POSSIBILITY OF RATIONAL IRRATIONALITY

In this section, I will argue that Gauthier fails to show that whether or not drinking the toxin is rational is determined by whether or not the Course is rational. He fails, in other words, to rule out the possibility of *rational irrationality* – the Course's being rational and the drinking's being

irrational.[39] And the most likely reason he fails is because there just *is* no good reason to think that rational irrationality is impossible.

Gauthier must either assume or demonstrate the following proposition:

(24) If the Course is rational, then my drinking the toxin is also rational.

Gauthier needs (24) because it (plus Gauthier's argument that the Course is rational) is necessary to derive what he ultimately hopes to show:

(10) My drinking the toxin is rational.

(24) is clearly controversial. Some philosophers, including myself, reject it. So Gauthier may not simply assume it. He must demonstrate it. Now, considering the matter a priori, there are two general ways to defend (24). Gauthier might assume or argue either:

(25) Bottom-up rationality. The rationality of the Course's constituent parts, including my drinking the toxin, makes the Course rational; or

(26) Top-down rationality. The rationality of the Course makes its constituent parts, including my drinking the toxin, rational.

But Gauthier is not entitled to rely on either proposition.

Gauthier is not entitled to rely on (25) for two different reasons. First, it begs the question. It simply *assumes* that the constituent parts of the Course, and therefore my drinking the toxin, is rational. And Gauthier may not assume that my drinking the toxin is rational, since this point (i.e. (10)) is precisely what is in question in the first place. Second, we have already seen in the last section that Gauthier embraces a proposition that turns out to be incompatible with (25):

(27) The rationality of the Course derives from the fact that its outcome is overall more beneficial to me than the outcome of the Alternative Course.

If (27) is the case, then (25) is not. If the rationality of the Course derives from the comparative favorability of its outcome, then it does *not* derive from the rationality of its constituent parts.

So Gauthier must opt to defend (24) with (26) instead of (25). And, indeed, this is precisely what Gauthier does.[40] But, as it turns out, Gauthier is no more entitled to rely on (26) than he is on (25). For (26) conflicts with a proposition that Gauthier's (27) entails by parity:

(28) The standard for measuring the rationality of my *drinking* the toxin is *its* outcome as compared with the outcome of the alternative action – i.e. *refraining* from drinking the toxin.

(26) conflicts with (28) in the same way that (25) and (27) conflict. While (26) says that the rationality of my drinking the toxin derives from the rationality of the Course, (28) says that the rationality of my drinking the toxin derives from something else – namely, its outcome as compared with the outcome of the alternative action.

In order to get around this problem, Gauthier would have to challenge the inference of (28) from (27). He would have to show:

(29) The transmission of rationality from the Course to my drinking the toxin *outweighs* the parity principle. The rational status of my drinking the toxin derives from the rational status of the Course *rather than* from its outcome as compared with the outcome of the alternative action of refraining from drinking the toxin.

In other words, Gauthier would have to *defend* his position that the rationality of drinking the toxin should be measured *not* by this action's comparative outcome with not drinking the toxin but rather by comparison of the Course (in which drinking the toxin is situated) with the Alternative Course.

As it turns out, Gauthier does not explicitly provide any arguments for (29). Instead, either Gauthier is simply assuming (29) (consciously or unconsciously) or Gauthier believes that another one of his assumptions or conclusions supports (29). If the former, Gauthier's position is weak. For (29) is by no means obviously true. In fact, it is more reasonable to presume that it is false – i.e. that (27) entails (28) – unless proven otherwise. Put another way, we may presume that (27) entails (28) unless Gauthier can show that (29) is true. Therefore Gauthier may not simply *assume* that (29) is true.

If the latter, the only one of Gauthier's assumptions/conclusions that might plausibly be thought to support (29) is Gauthier's position that rule-utilitarianism is superior to act-utilitarianism. (See sections 2.3 and 3.2.) Again, according to Gauthier, I do better overall to adhere to rational courses of action than to abandon them even if the former sometimes require me to perform actions that yield less utility than alternative actions.

But closely related as the contents of the two propositions are, the proposition that I might do better overall to adhere to rational courses of action does *not* entail the proposition that the rational status of the actions within each of these courses of action is determined by the rational status of the courses of actions themselves. For the former proposition is consistent with the negation of the latter proposition. It is consistent to maintain that I will do better overall by adhering to rational courses of actions even

if some of these rational courses of action consist in part of irrational actions.

Gauthier might respond that this point *begs the question* against (26). For it assumes that rational irrationality – i.e. a rational course of action's consisting at least in part of irrational sub-parts – is possible when (26) proposes the very opposite: that the rationality of the course of action is fully 'organic,' automatically 'spreads' to its constituent parts, and therefore that rational irrationality is impossible. But, first, Gauthier *acknowledges* the possibility of rational irrationality.[41] To this extent, Gauthier is guilty of inconsistency. Second, even if Gauthier had *not* so acknowledged, he is *committed to* the possibility of rational irrationality. For we have already seen that Gauthier endorses (27), that (27) entails (28), and that (27) and (28) together suggest that the rational status of both the Course and my drinking the toxin depend on the respective outcomes of each. So a divergence of outcomes would lead to a divergence of rational statuses. And a divergence of rational statuses would amount to rational irrationality. Indeed, contrary to Gauthier, TP itself proves this point. For in TP, while the outcome of the Course is comparatively favorable, in which case the Course is rational, the outcome of my actually drinking the toxin is comparatively unfavorable, in which case my actually drinking the toxin is irrational.[42]

Although Gauthier does not offer any other arguments for (29), one might argue that there is still an argument available to him. According to this argument, (29) must be true because it is in my rational self-interest to adopt (29). And it is in my rational self-interest to adopt (29) because adopting (29) will help me to form the overall more beneficial course of action, the Course. This argument fails, however, because the pragmatic benefits of adopting (29) do not help to show that (29) is true. As Gauthier himself argues, some propositions may be false despite the utility or rationality of adopting them.[43]

We may conclude that Gauthier has failed to give us a good reason to accept (29) – i.e. to accept the proposition that the rational status of my drinking the toxin derives from the rational status of the Course rather than from its outcome as compared with the outcome of its alternative – namely, refraining from drinking the toxin. And (29) was really Gauthier's last, best hope. To see why, we need to retrace the steps of this section. Once again, Gauthier needs (29) because he accepts:

(27) The rationality of the Course derives from the fact that its outcome is overall more beneficial to me than the outcome of the Alternative Course

and (29) is the only possible way of blocking (27)'s entailment (by parity) of:

(28) The standard for measuring the rationality of my *drinking* the toxin is *its* outcome as compared with the outcome of the alternative action – i.e. *refraining* from drinking the toxin.

Given Gauthier's failure to establish (29), his acceptance of (27) commits him to (28). This commitment is a problem for Gauthier because (28) is inconsistent with:

(26) The rationality of the Course makes its constituent parts, including my drinking the toxin, rational

and (26) is one of the only two ways in which Gauthier may establish:

(24) If the Course is rational, then my drinking the toxin is also rational.

So Gauthier must abandon (26) as a means of establishing (24). But, besides (26), the only other means by which Gauthier may establish (24) is:

(25) The rationality of the Course's constituent parts, including my drinking the toxin, makes the Course rational.

And (25) not only begs the question in favor of Gauthier's ultimate conclusion:

(10) My drinking the toxin is rational

but also conflicts with (27), which – again – Gauthier accepts. So Gauthier must also abandon (25) as a means of establishing (24). Without (25) or (26), Gauthier has no hope of establishing (24). And without (24), even if we concede that Gauthier has established the antecedent of (24) (i.e. the Course is rational), Gauthier cannot establish (10) – again, his ultimate conclusion.

4. Conclusion

Kavka's Argument suggests that I cannot intend to drink a pain-inducing toxin, even if I have an outcome-oriented reason to form the intention, if I have no outcome-oriented reason to drink the toxin and a (substantial) outcome-oriented reason not to drink the toxin. The Rationalist Solution, however, suggests that I *can* form the intention to drink the toxin. For even though I do not have an outcome-oriented reason to drink the toxin,

I still *do* have a good reason to drink the toxin – namely, the *backward-looking* reason that (a) I previously formed the intention to drink the toxin, (b) the intention was rational, and (c) my intention correctly anticipated the circumstances surrounding my drinking the toxin.

(a) through (c) constitute a sufficiently strong reason for drinking the toxin only if it can be shown that the rationality of the intention *transfers* to the action itself. Two arguments in favor of this proposition are the 'Self-Promise Argument' and David Gauthier's 'Rational Self-Interest Argument.' According to the Self-Promise Argument, my rational intention to drink the toxin constitutes a promise to myself and this promise imposes an obligation upon me to fulfill it. According to Gauthier's Rational Self-Interest Argument, my rational intention to drink the toxin recommends a course of action (the 'Course') of which drinking the toxin is an essential part.

But both the Self-Promise Argument and the Rational Self-Interest Argument fail. The Self-Promise Argument fails because my rational intention to drink the toxin construed as a self-promise imposes at most a practical, not a moral, obligation upon me to act in accord with it. And this practical obligation is outweighed by *another* practical obligation of mine, my forward-looking reason to minimize my pain. The latter outweighs the former because it will help me to avoid rather than suffer a day of pain. The Rational Self-Interest Argument fails for much the same reason. The fact remains that, after I have won the $1m, I do better to abandon the Course and refrain from drinking the toxin than to stick with the Course and drink the toxin. For while I win $1m either way, changing my mind spares me the day of pain that following through would cost me.

Gauthier's response to this latter objection is that the rationality of drinking the toxin should be determined *not* by comparing it with the alternative action – namely, refraining from drinking the toxin – but rather by first situating it within the Course and then comparing the rationality of the Course with the rationality of the alternative course of action – namely, not intending to drink the toxin at all (the 'Alternative Course'). But this response fails. For it ultimately relies on the assumption that if the Course is rational, then my actually drinking the toxin is rational. And Gauthier is not in a position either to assume or defend this proposition. He cannot assume it just because it is so controversial. Those who endorse the possibility of rational irrationality, such as myself, reject it. Nor may he defend it. For the only two ways in which he might try – namely, by showing either that the rationality of a course of action determines the rationality of its constituent parts (intention and action) or vice versa – are both incompatible with Gauthier's commitment to the proposition that the Course is rational because it is more beneficial overall than the Alternative Course.

I conclude, then, that the Rationalist Solution fails. And because the Rationalist Solution embodies two different propositions, this failure is actually twofold. First, the Rationalist Solution fails to show that drinking the toxin is rational. As a result, we may still hold on to what seems anyway to be the more intuitively plausible view – namely, that drinking the toxin is *ir*rational. Second, as a result of the first failure, the Rationalist Solution fails to show that it is possible for me to intend to drink the toxin and thereby win the $1m. So whereas the Rationalist Solution argues from the rationality of drinking the toxin to the possibility of intending to drink the toxin, it is still open to some to argue in precisely the reverse (Kavka's) direction – from the irrationality of the drinking the toxin to the impossibility of intending to drink the toxin.[44,45]

Louisiana State University Law Center

NOTES

[1] See also den Hartogh, 2004, pp. 10–11; Gauthier, 1984b, p. 161; 1994, pp. 696, 697; Kavka, 1978, p. 292. Another formulation of (5) that might be acceptable is:

(5*) I cannot intend to do what I know that I will not do.

[2] Another way to formulate Kavka's Argument is in terms of McClennen's distinction between 'myopic' and 'sophisticated' planning. See McClennen, 1990, chs. 9, 11, 13. See also Bratman, 1998b, pp. 64 ff.; Finkelstein, 2001, pp. 58 ff.; Gauthier, 1996, pp. 222–25; 1997a, pp. 8–12. According to Kavka's Argument, it is impossible for me to remain myopic tonight. It is impossible for me tonight to remain concerned only with my current preference to win the $1m by intending to drink the toxin. Instead, I cannot help but be sophisticated. That is, I cannot help but concern myself tonight also with the preference that I expect to have tomorrow afternoon – namely, my expected preference to refrain from drinking the toxin. And given this concern, I cannot form the intention tonight after all.

[3] Bratman (1996, pp. 55–56; 1998a, 57, pp. 62–63) and Farrell (1989) appear to be sympathetic to Kavka's Argument. Hinchman (2003, p. 43) says that I 'cannot form the intention to drink without ignorance, manifest irrationality or external measures.'

[4] It has been suggested to me that what is counter-intuitive is not that I cannot form the intention to drink *per se* but that I cannot form the intention *even though it is rational*. I agree, and will argue below, that this intention *is indeed* rational. So I have no real problem with this alternative formulation.

[5] A possible fourth strategy might proceed to ~(6) through (11); the assumption that it is if it is rational for me to drink the toxin, then I can drink the toxin; and the assumption that it is rational for me to drink the toxin. But given that it is more cumbersome than the other three approaches, I regard it as unworthy of discussion.

[6] See Gauthier, 1984b, p. 159; 1994, pp. 707–10; 1998a; 1998b, pp. 50–53; Harman, 1998; Holton, 2004, pp. 528–30; and McClennen, 1990, pp. 226–31. Bratman (1987, pp. 105–06; 1998a, pp. 59 ff.; 1998b, pp. 66, 74–76) offers an objection to Gauthier's position as well as his own positive attempt to solve TP, an attempt which does not clearly follow any one of the three strategies outlined above. Bratman seems tacitly to assume throughout his discussions of TP that if it is rational for me to intend to A, then I can intend to A. But if he is indeed

making this assumption, then he is simply begging the question in favor of the Rationalist Solution. I make a similar point in section 2.2.

[7] The Rationalist Solution is best known in the literature as the theory of 'resolute planning.' The resolute approach is supposed to constitute a way around (or through) the false dichotomy between being 'myopic' and being 'sophisticated.' See note 2. For discussions of different versions of resolute planning, see Bratman, 1998b, pp. 64 ff.; Gauthier, 1996, pp. 222 ff.; 1997a, pp. 13 ff., 20 ff.; Holton, 2004; Finkelstein, 2001, pp. 59 ff.; McClennen, 1990, chs. 11, 12.

[8] If it turns out that Gauthier is wrong and drinking the toxin is irrational, it remains an open question whether or not the irrationality of drinking the toxin makes it – and therefore my forming the intention to drink – impossible.

[9] Similarly, David Gauthier, one of the principal proponents of the Rationalist Solution, maintains that I must believe tonight not merely that I have *a* reason to drink the toxin tomorrow but that I will *continue to believe tomorrow* that it is in my rational self-interest for this reason to prevail. See Gauthier, 1994, pp. 694–98, 708–09; 1998a, pp. 56–57. I develop Gauthier's argument further in section 3.2.

[10] See Gauthier, 1984a, p. 483.

[11] Kavka (1984, p. 155) calls these consequences of my intention its 'direct effects.' See also Broome, 2001, pp. 102, 103–04. Notice, I am speaking of intention here as a *psychological cause* of my drinking the toxin. This treatment contrasts, though does not necessarily conflict, with the suggestion in section 2.4 by the proponent of the Rationalist Solution that my intention to drink the toxin may itself be regarded as a good *reason* to drink the toxin. Bratman (1996, pp. 44 ff.) describes yet a third alternative to both of these views of intentions but ultimately sides with the *causal theory* (pp. 51–52). For a thorough discussion of the nature of intentions in general, see Bratman, 1987, pp. 3 ff., 15 ff., 56–59.

[12] Gauthier (1998a, pp. 48, 49) makes the similar point that it is better to intend to drink and to drink than not to intend to drink at all. See section 2.2.

[13] For discussions of rational irrationality, see Gauthier, 1994, pp. 697 ff.; 1996, pp. 239–40; 1997b, pp. 27–28, 36–37; Kavka, 1978, p. 293; Mele, 1995, p. 184; Parfit, 1984, pp. 9, 12–17, 46; 2001, p. 86; Schelling, 1980, pp. 16–18. Goldstein (2003, p. 244) seems to think that rational irrationality is impossible. I agree with Holton (2004, p. 512), Kavka (1984, pp. 156–57), and Parfit (2001, pp. 91–92) that rational irrationality is possible.

[14] See Gauthier, 1984a, pp. 480, 481, 487–88; 1994, 709; 1998a, p. 55; 1998b, pp. 43 ff.; Kavka, 1978, p. 301; Lewis, 1984, pp. 141, 143. Bratman (1987, pp. 103–104) concludes from this argument that I have good reason not to intend to drink the toxin but only to cause myself to have the intention to drink the toxin, which he regards as different.

[15] One might wonder why only two options are presented here, that the third option of intending tonight to drink the toxin and then refraining from drinking it tomorrow is not also included. But as will be explained in section 3.2, even if this option constitutes a coherent sequence of events, it does not constitute a coherent plan or 'course of action.' I cannot at the same time both plan or intend to drink the toxin and intend to refrain from drinking the toxin. I cannot at the same time both intend to drink the toxin and change my mind.

[16] According to Gauthier (1998a, p. 49), the fact that the rationality of my intention does not derive from my action itself or from any reasons supporting my action is the very source of the puzzle in TP.

[17] Kavka's term. See Kavka, 1978; 1984.

[18] Setiya (2003, pp. 369–70) seems to reject the possibility of having reasons to intend that do not also support the intended action.

[19] Another kind of backward-looking reason is 'sunk costs' – i.e. resources that an agent has previously expended in pursuit of a goal or project and that are no longer useful or

retrievable. People often regard sunk costs as a good, if not sufficient, reason for continuing to expend further resources in pursuit of the same goal or project. For an insightful discussion of whether or not this attitude is rational, see Kelly, 2004.

[20] See Gauthier, 1984a, pp. 479–80, 483, 486; 1984b, p. 159. McClennen (1990, p. 228) correctly suggests that Gauthier takes RAP 'to be a basic principle governing future intention.' Bratman (1987, pp. 105–06; 1998b, pp. 61 ff.) rejects RAP, even though he endorses two similar propositions – the 'Intention-action principle' (1987, pp. 54–55) and the 'linking principle' (1998a, p. 55; 1998b, p. 62). Kavka (1984, pp. 156–57) tries to refute RAP with TP itself. Although he distinguishes between several different versions of RAP, Parfit (1984, pp. 37–40; 2001, pp. 82 ff., 91–92) ultimately rejects it.

[21] See Farrell, 1989, p. 288.

[22] See Gauthier, 1994, p. 709.

[23] See Gauthier, 1994, p. 709; 1996, pp. 218, 241; 1998a, p. 52; Setiya, 2003, pp. 359 ff. Bratman (1987, pp. 34, 109) suggests that intentions may constitute 'special kinds of reasons – *framework reasons* – whose role is to help determine the relevance and admissibility of options.' Bratman (1996, p. 45 ff.) also suggests that the beliefs behind my intentions may themselves constitute reasons. Broome (2001) argues that the mere fact that I have previously formed an intention does not by itself constitute a reason for acting on that intention. See also den Hartogh, 2004.

[24] See Bratman, 1996, p. 53; Gauthier, 1994, p. 709; 1996, pp. 217–18, 220, 230 ff., 240.

[25] See, e.g. Gauthier, 1994, p. 708. This is also Parfit's interpretation of Gauthier. See Parfit, 2001, p. 88.

[26] See Broome, 2001, p. 102. Copp (1986, p. 15), Kavka (1978, p. 291), and Lewis (1984, p. 143) make similar points with regard to deterrent intentions. See also Finkelstein, 1999, p. 326. Gauthier (1984b, p. 160) refers to Lewis's position in this regard as 'schizophrenic.'

[27] Bratman (1987, pp. 24 ff., 42 ff., 78, 86–87), Farrell (1989, p. 291), Gauthier (1996, pp. 218, 241), and Holton (2004, pp. 509–10, 513–16) discuss similar bootstrapping problems.

[28] See Bratman, 1987, pp. 101–06; Holton, 2004, p. 528; Kavka, 1978, p. 293.

[29] Gauthier never explicitly proposes the Self-Promise Argument, but he arguably motivates it in two different ways. First, Gauthier frequently speaks of honoring assurances. See, e.g. Gauthier, 1984a, p. 475; 1994, pp. 704, 707, 712–13; 1996, pp. 234, 242–43; 1998b, p. 45. To be sure, Gauthier uses this language in the context of inter-personal coordination. But he discusses inter-personal coordination primarily to defend a point that he equally applies to *intra*-personal coordination – namely, that counter-preferential, even negative-utility-yielding, actions may still be rational if they execute rational, autonomous-benefit-yielding intentions. See section 3.2. Second, Gauthier (1994, p. 709; 1996, pp. 218, 241; 1998a, p. 52) suggests that intentions may constitute reasons for action. And one of the most plausible ways in which an intention may constitute a reason for action is by imposing upon me some obligation, even if minimal and defeasible, to honor it.

[30] Gauthier (1997a, p. 18) makes a similar point in the course of arguing that rational agents may not be motivated to act resolutely on their previous intentions in the same way that they are motivated to act resolutely on their assurances to others:

> [W]hen [a rational agent] turns her attention to the effect of her manner of deliberation on the realization of her own concerns over time, independently of her interactions with others, she will not view intrapersonal connections as if they were interpersonal relations. Certainly she will not suppose that her posterior self might regard her prior self as an end in itself . . . [T]he agent, at any given time, has no interest in preferences or concerns that she once embraced but holds no longer. She need not, and normally does not, regret having had concerns that she no longer embraces. She may indeed see her past life as

having been enriched by those concerns . . . and she may value the present self that she recognizes as having been significantly shaped by them. But their practical significance belongs to her past life, not her present one. To give them standing in her current deliberations would make no sense to her.

[31] See (5) in Kavka's Argument (in the Introduction).

[32] We are excluding from consideration *moral* reasons, which are inapplicable in the context of TP, and considering only reasons deriving from (rational) self-interest. As Gauthier (1997b, p. 27) says, '[O]ur concern is with rationality and not morality.' See also Gauthier, 1994, pp. 693 n.6, 693 n.7, 704 n.17; 1996, pp. 242–43; 1997b, p. 27; 1998b, p. 43. For a more elaborate account of Gauthier's moral theory and its relation to rationality, see Gauthier, 1975; 1984a, p. 494; 1986.

[33] See Gauthier, 1984b, p. 159; 1994, pp. 707–10; 1998a. There are two omissions from this distilled composite of Gauthier's various discussions of TP and rational planning more generally. First, Gauthier's points about inter-personal coordination are omitted because (a) TP concerns primarily intra-personal coordination and (b) the lessons that Gauthier draws from them apply – and are meant to apply – to intra-personal coordination as well. See, e.g., Gauthier, 1994, pp. 692–97, 702–07; 1996, p. 231; 1997b, pp. 26–27, 29–30, 32–35; 1998a, pp. 50–51, 53; 1998b, pp. 42–50. Second, Gauthier's argument for the proposition that resolute planning is preferable to sophisticated planning is also omitted here. See notes 2 and 7. According to Gauthier, the former yields a more favorable cost/benefit ratio than the latter. While resolute planning requires executing the Course through sheer force of rational will, sophisticated planning may yield the same benefits as resolute planning but only at the greater cost of pre-committing myself to carrying out the plan. See Gauthier, 1994, pp. 708–09; 1996, pp. 222 ff.; 1997a; 1997b.

[34] Actually, four objections. The third objection suggests that (22) is not necessary, that I can cause myself through means other than rational reflection – namely, through rational irrationality – to acquire the belief that it is in my self-interest to drink. This issue (rational irrationality) will be discussed in section 3.3; see also note 13. The fourth objection suggests that the inference of (23) from (22) is invalid because it falsely assumes that endorsement of a given proposition after rational reflection guarantees the truth of that proposition.

[35] See Gauthier, 1994, pp. 695, 696; 1997b, p. 33; 1998a, p. 48; 1998b, p. 45.

[36] See Gauthier, 1984a, p. 488; 1996, pp. 221–22, 238.

[37] See Gauthier, 1994, pp. 704–09; 1996, pp. 231 ff., esp. 232, 241; 1998a; 1998b, pp. 44–53. See also Gauthier, 1984a, pp. 487–89 (advocating a similar position in the context of deterrence); Gauthier, 1986, pp. 182–83 (advocating a similar position in terms of dispositions as opposed to actions). For useful summaries of Gauthier's position, see Bratman, 1998a, pp. 57–59; 1998b, pp. 65–66. Finkelstein (2001, pp. 57 ff.) rejects Gauthier's notion that while plan-adoption is 'preferentially constrained', my performance or execution of the plan may not be. See also Bratman, 1987, pp. 23–27, 68–69. Parfit (1984, esp. chs. 7, 8) discusses this issue in terms of the desires of my present self and the desires of my future self. See also Gauthier, 1986, pp. 36–38.

[38] Bratman (1998a, pp. 58–59; 1998b, p. 61) points out that Gauthier's account is 'two-tier' (see also Finkelstein, 2001, p. 60) and 'seems similar in structure to versions of rule-utilitarianism.' (Incidentally, Bratman (1987, pp. 64, 68 ff.) suggests that his own approach to 'nonreflective (non)reconsideration of a prior intention' is also two-tier and 'analogous in structure to certain versions of rule-utilitarianism' as well. See also Holton, 2004, p. 510.) Gauthier (1984a, p. 488) manifests his adherence to the rule-utilitarian principle that certain self-destructive actions can still be rational when he says: 'the actor who assesses the rationality of his actions only from now, from the point at which the question of performance

arises, may expect a lesser overall utility than the actor who assesses the rationality of her actions in the context of policies, who adjusts performances so that the probability-weighted sum of their utilities is greatest.' See also Gauthier, 1996, pp. 236–237; 1997b, p. 34; Parfit, 2001, p. 87. Despite Gauthier's largely rule-utilitarian approach, he still makes a dig against utilitarianism in general in Gauthier, 1984b, p. 161.

[39] See notes 13 and 34.

[40] See Gauthier, 1994, pp. 701, 705, 708–09; 1996, pp. 239–40; 1998a, pp. 48, 50.

[41] See Gauthier, 1994, pp. 697–702; 1996, pp. 237, 239–40; 1997b, pp. 27–28, 36–37.

[42] It has been suggested to me that if the Course really does not supervene on its parts, then it might be possible for it to be rational *even if* not only my action but also my intention to drink the toxin are *both* irrational. It is not clear, however, that such a situation is possible. For it is doubtful that the following two conditions are compatible: (a) the intention to drink is irrational *and* (b) the Course can be reasonably expected to produce an outcome more favorable than the outcome of the Alternative Course.

[43] See Gauthier, 1998a, 51–52:

> Do I reflect on the benefits of so believing, even at the cost of performing, in relation to not believing? This does not lead me to adopt the belief. For believing is believing *true*, and reflecting on the benefits of believing that I have reason to drink the toxin seems quite irrelevant to determining whether the belief is true. It may seem plausible to claim that if I would benefit from forming an intention, despite the cost of executing it, then I have reason to form and (if all turns out as I expect) carry out the intention. But it does not seem plausible to claim that if I would benefit from adopting a belief, despite the cost of acting in accordance with it, then I have reason to adopt and (if all turns out as I expect) act on the belief.

To use the language in Hieronymi (2005), Gauthier is suggesting in the passage above that only 'constitutive' or content-based reasons for believing a proposition such as (29), not 'extrinsic' or utility-based reasons for causing myself to believe (29), can warrant a belief in (29). See also Parfit, 2001, pp. 88–90 (agreeing with Gauthier on this point); and the fourth objection in note 34. *Cf.* Kelly, 2002 (arguing that the expected utility of having a given belief does not necessarily make that belief *rational*).

[44] See Farrell, 1989.

[45] I would like to thank Howard McGary, Harry Silverstein, and Roy Sorensen for helpful comments on much earlier drafts of this paper. I would also like to thank three anonymous referees at this journal for helping me to make further improvements on a more recent draft.

REFERENCES

Bratman, M. E. (1987). *Intentions, Plans, and Practical Reason*. Cambridge, MA and London: Harvard University Press.

Bratman, M. E. (1996). 'Planning and Temptation,' in his (1999) *Faces of Intention*. Cambridge and New York: Cambridge University Press, pp. 35–57.

Bratman, M. E. (1998a). 'Following Through with One's Plans: Reply to David Gauthier,' in Danielson (1998), pp. 55–65.

Bratman, M. E. (1998b). 'Toxin, Temptation, and the Stability of Intention,' in Coleman and Morris (1998), pp. 59–83.

Broome, J. (2001). 'Are Intentions Reasons? And How Should We Cope with Incommensurable Values?' in Morris and Ripstein (2001), pp. 98–120.

Coleman, J. L. and Morris, C. W., eds. (1998). *Rational Commitment and Social Justice: Essays for Gregory Kavka*. Cambridge and New York: Cambridge University Press.

Copp, D. (1986). 'Introduction: Deterrence and Disarmament,' *Canadian Journal of Philosophy* Supp. 12, pp. 1–21.
Danielson, P. ed. (1998). *Modeling Rationality, Morality, and Evolution*. Oxford: Oxford University Press.
den Hartogh, G. (2004). 'The Authority of Intention,' *Ethics* 115, pp. 6–34.
Farrell, D. M. (1989). 'Intention, Reason, and Action,' *American Philosophical Quarterly* 26, pp. 283–95.
Finkelstein, C. (1999). 'Threats and Preemptive Practices,' *Legal Theory* 5, pp. 311–38.
Finkelstein, C. (2001). 'Rational Temptation,' in Morris and Ripstein (2001), pp. 56–80.
Gauthier, D. (1975). 'Reason and Maximization,' *Canadian Journal of Philosophy* 4, pp. 411–33.
Gauthier, D. (1984a). 'Deterrence, Maximization, and Rationality,' *Ethics* 94, pp. 474–95.
Gauthier, D. (1984b). 'Responses to the Paradox of Deterrence: Afterthoughts,' in MacLean (1984), pp. 159–61.
Gauthier, D. (1986). *Morals by Agreement*. Oxford: Clarendon Press.
Gauthier, D. (1994). 'Assure and Threaten,' *Ethics* 104, pp. 690–721.
Gauthier, D. (1996). 'Commitment and Choice: An Essay on the Rationality of Plans,' in F. Farina, F. Hahn and S. Vannucci (eds) *Ethics, Rationality, and Economic Behaviour*. Oxford: Oxford University Press, pp. 217–43.
Gauthier, D. (1997a). 'Resolute Choice and Rational Deliberation: A Critique and a Defense,' *Noûs* 31, pp. 1–25.
Gauthier, D. (1997b). 'Rationality and the Rational Aim,' in J. Dancy (ed.) *Reading Parfit*. Oxford: Blackwell, pp. 24–41.
Gauthier, D. (1998a). 'Rethinking the Toxin Puzzle,' in Coleman and Morris (1998), pp. 47–58.
Gauthier, D. (1998b). 'Intention and Deliberation,' in Danielson (1998), pp. 41–54.
Goldstein, L. (2003). 'Explaining Boxing and Toxin,' *Analysis* 63, pp. 242–44.
Harman, G. (1998). 'The Toxin Puzzle,' in Coleman and Morris (1998), pp. 84–9.
Hieronymi, P. (2005). 'The Wrong Kind of Reason,' *Journal of Philosophy* 102, pp. 437–57.
Hinchman, E. S. (2003). 'Trust and Diachronic Agency,' *Noûs* 37, pp. 25–51.
Holton, R. (2004). 'Rational Resolve,' *Philosophical Review* 113, pp. 507–35.
Kavka, G. S. (1978). 'Some Paradoxes of Deterrence,' *Journal of Philosophy* 75, pp. 285–302.
Kavka, G. S. (1983). 'The Toxin Puzzle,' *Analysis* 43, pp. 33–6.
Kavka, G. S. (1984). 'Responses to the Paradox of Deterrence: Deterrent Intentions and Retaliatory Actions,' in MacLean (1984), pp. 155–59.
Kelly, T. (2002). 'The Rationality of Belief and Some Other Propositional Attitudes,' *Philosophical Studies* 110, pp. 163–96.
Kelly, T. (2004). 'Sunk Costs, Rationality, and Acting for the Sake of the Past,' *Noûs* 38, pp. 60–85.
Lewis, D. (1984). 'Devil's Bargains and the Real World,' in MacLean (1984), pp. 141–46.
MacLean, D, ed. (1984). *The Security Gamble: Deterrence Dilemmas in the Nuclear Age*. Totowa, NJ: Rowman & Allanheld.
McClennen, E. F. (1990). *Rationality and Dynamic Choice*. Cambridge: Cambridge University Press.
Mele, A. E. (1995). *Autonomous Agents: From Self-Control to Autonomy*. New York: Oxford University Press.
Morris, C. and Ripstein, A., eds (2001). *Preferences, Principles and Practices: Essays in Honor of David Gauthier*. Cambridge: Cambridge University Press.

Parfit, D. (1984). *Reasons and Persons*. New York: Oxford University Press.
Parfit, D. (2001). 'Bombs and Coconuts, or Rational Irrationality,' in Morris and Ripstein (2001), pp. 81–97.
Schelling, T. C. (1980). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
Setiya, K. (2003). 'Explaining Action,' *Philosophical Review* 112, pp. 339–93.