ALFRED R. MELE

INTENTIONS, REASONS, AND BELIEFS: MORALS OF THE TOXIN PUZZLE

(Received 31 January, 1992)

In garden-variety instances of intentional action, according to a popular account, agents intend to perform actions of particular kinds, their intentions are based on reasons so to act, and the intentions issue in appropriate behaviour.[1] On this account, the reasons that give rise to our intentions are *reasons for action.* Interesting questions for this view are raised by cases in which an agent seemingly has a reason to *intend* to do something while having no reason to *do* it. Can such reasons to intend issue in appropriate intentions? If so, can these intentions issue in corresponding intentional actions — even though the agent has no reason to perform those actions? If these questions are properly given an affirmative answer, at least one popular thesis in the philosophy of action is false. One could not properly "define an intentional act as one which the agent does *for a reason.*"[2] A popular thesis about the *explanation* of intentional actions would be false as well — namely, that explaining an intentional action (qua intentional) requires reference to reasons for *action.*

My point of departure in this paper is a puzzle — Gregory Kavka's toxin puzzle (1983) — in which agents seem to have an excellent reason to intend to A while having no reason at all to A. Generally, commentators on the puzzle have set their sights on questions about *rational* intentions. However, the puzzle raises difficult questions about intending itself and about the nature of intentional action. Showing that this is so is easy. Answering the question is more challenging.

The bulk of Kavka's puzzle, further details of which emerge shortly, runs as follows:

> You have just been approached by an eccentric billionaire who has offered you the following deal. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects. . . . The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight,

Philosophical Studies **68**: 171—194, 1992.

you *intend* to drink the toxin tomorrow afternoon. He emphasizes that you need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives, if you succeed. . . . All you have to do is . . . intend at midnight tonight to drink the stuff tomorrow afternoon. You are perfectly free to change your mind after receiving the money and not drink the toxin. (The presence or absence of the intention is to be determined by the latest 'mind-reading' brain scanner. . .) (1983, pp. 33—34)

Much of what I shall have to say revolves around an attempt to construct an agent who is capable of winning the money. My central character, Ted, is a bizarre fellow. If *only* an agent as peculiar as Ted can succeed, that itself would have significant implications for our understanding of the connections, in normal human beings, among intentions, reasons, and intentional actions. So would the very possibility of a successful agent. To say at this juncture, however, whether the money can actually be won — or even to say any more about the implications of there being a winner — would spoil the fun.

1. DEAD ENDS

If you want the billionaire's money and believe that you can get it by intending to drink the toxin, then you have a reason to intend to drink the toxin, on a familiar conception of reasons. However, you have no reason, as you know, to drink the toxin, since the money will be in your bank account long before the time for drinking comes, if you succeed in forming the intention. So, given the desire and belief just mentioned, it looks as though you have a reason to intend to do something that you have no reason to do. (Practical reasons are often understood as psychological items — roughly, desire/belief complexes. On a competing conception, practical reasons are abstract items — for example, propositions. For my purposes, there is no need to take a stand on this issue. If reasons are psychological items, then by "reasons" I mean reasons; if, alternatively, reasons are abstract items, "reasons" in this paper means psychological states of reason-having.)

Most of us would want to intend to drink the toxin. But can we intend to drink it, knowing that there will be no reason to drink it? You might suppose that you can intend at midnight to drink the toxin tomorrow afternoon, even though you would be convinced at midnight — on the grounds that you will have no reason to drink it and excellent

reasons not to drink it — that you will not drink it when the time arrives. After all, you might think, you have an excellent reason to form an intention to drink the toxin; and, it seems, it is much easier to form an intention than it is to do many things that you have done for considerably less powerful reasons.

Some will urge, however, that your thinking is incoherent. Surely, a person who is convinced that she will not drink the toxin does not *intend* to drink the toxin, although she might, perhaps, pretend to intend — even to herself. Indeed, knowing yourself as you do, you know that you won't even make an effort to drink the toxin. But intending, whatever it may be, is such that agents who intend to A are not simultaneously convinced that they won't even try to A. So, at least, it will be claimed. And even if there are Freudian intentions to A, harbored in the unconscious mind while the conscious mind is convinced that one would never endeavor to A, you have no idea how to form a Freudian intention.

It might have occurred to you that you can *give* yourself a reason for drinking the toxin. You might, for example, sign a contract that commits you to turning over your most prized possessions to your worst enemy if you do not drink the toxin. Perhaps, having signed such a contract, you would intend to drink the toxin. Unfortunately, Kavka's billionaire has placed certain restrictions on the avenues by which you may form or acquire the intention: "arrangement of such external incentives is ruled out, as are such alternative gimmicks as hiring a hypnotist to implant the intention, forgetting the main relevant facts of the situation, and so forth" (p. 34).

You might cast about for avenues not proscribed by the billionaire. But in so doing, it is important not to lose sight of the puzzle. In at least one version of the case, the question, at bottom, is this: *Given that you are convinced at midnight that tomorrow (like today) you will have no reason to drink the toxin*, can you intend at midnight to drink it tomorrow? To keep things simple, let us suppose that the billionaire explicitly states that the intention must be paired with this conviction if you are to receive the money. (I shall call this conviction constraint "condition *C*.")[3]

You might think that you can cause yourself to become so irrational that tomorrow you will drink the toxin while lacking a reason to drink it

and having very good reasons not to drink it. And you might set out to construct a plan for making yourself irrational, so that, at midnight (while still rational), you can intend to drink the toxin tomorrow even though you are convinced at midnight that tomorrow you will have no reason to drink it and excellent reasons not to drink it. But you are not irrational yet! You know (or think you know) that even irrational agents do not perform intentional actions for no reason at all: They have their reasons for what they do, when they act intentionally. So are you thinking that you can make yourself so irrational that tomorrow you will *unintentionally* (or nonintentionally) drink the toxin? How is that supposed to work? Do you suppose that there is a significant nomic or statistical correlation between being irrational and unintentionally drinking toxins? Of course not.

Those wishing to explore strategies involving unintentional action, might do better to set irrationality aside. Perhaps tonight you can do something that will have the result tomorrow that you take the toxin to be, not the toxin, but tea, your favorite afternoon drink. If, tonight, you can so arrange things that at midnight you will be confident that the toxin will be in your favorite afternoon drink tomorrow and confident, as well, that by tomorrow you will have forgotten all about the toxin, then at midnight you can, while satisfying condition *C*, intend to drink the toxin tomorrow. The content of your intention may be described roughly as follows: "Tomorrow afternoon, I drink the toxin unintentionally while sipping my customary afternoon tea." You will, of course, have a reason tomorrow to drink tea, and in your tea will be the toxin; but you will have no reason at all to drink the toxin. And that is quite plain to you at midnight.

But wait; doesn't this involve some proscribed *forgetting*? Yes, it does. If the strategy works, you will have forgotten where the toxin is.

It is harder than one might have thought to form or acquire the intention while satisfying condition *C* and complying with the other constraints identified by the billionaire. Why is that? Kavka's own diagnosis centrally involves the claim that intentions are "dispositions to act that are based on *reasons to act* — features of the act itself or its (possible) consequences that are valued by the agent" (p. 35). Thus, "you cannot intend to act as you have no reason to act, at least when you have substantial reasons not to act. And you have (or will have

when the time comes) no reason to drink the toxin, and a very good reason not to. . . ."

This universal claim about intentions cannot, however, be established by reflection on Kavka's puzzle. For not all the work is done by truths about intention, reasons, and the like: There are the billionaire's stipulations as well. Were it not for the proscription against forgetting, for example, you could win the money in the way just described — you could, while satisfying the conviction condition, *C*, intend to drink the toxin tomorrow. Indeed, the "forgetting" case falsifies the idea that all possible intentions to *A* are "based on reasons to [*A*]." In that case, knowing that you have no reason to drink the toxin tomorrow, you nevertheless form an intention to do so.[4]

2. A NEWS FLASH!

In the forgetting case, if things go according to plan, you drink the toxin unintentionally. The case leaves open a weaker version of Kavka's moral. Perhaps all possible intentions to *A* such that in executing them one would *intentionally A* are based on reasons to *A*. Although we cannot find in reasons for *A*-ing a necessary basis for all possible intentions to *A*, we might find in them a necessary basis for all intentions of the sort just identified. If the money cannot be won in the toxin case, the explanation might well derive from this weaker moral. Perhaps the billionaire's conditions preclude winning the money in the absence of an intention of a kind capable of issuing in an intentional toxin drinking.

Here is a news flash! Once upon a time, a man, Ted, did receive a million dollars from Kavka's billionaire for intending to drink a toxin under the conditions laid down. So, at least, a magazine article alleged. Ted, it turns out, had a toxin problem of his own. Whenever a liquid toxin was nearby, he was sure to drink it. At first, he would mistake the toxins for something familiar — water or iced tea, for example. Later, when he became more careful about what he drank, he developed a serious case of somnambulism and started drinking toxins in his sleep, dreaming that he was sipping harmless beverages. Eventually, Ted discovered that an evil genius was determined to bring it about, by hook or by crook, that he would drink whatever liquid toxins were

nearby — provided that his drinking them would not prove fatal. (He was convinced, more precisely, that the genius would cause him to drink any such toxins unless the genius knew that Ted would *intentionally* drink the toxin: What mattered to the genius was that Ted drink the toxins — he placed no special value on *causing* Ted to do this.) Nor would he permit Ted to remove such toxins from his environment without drinking them. (Readers inclined to suppose that nothing of interest can be learned from "far-fetched" cases of this sort are encouraged to keep an open mind. The worry is addressed in Section 7.)

One day, it came to pass that the billionaire, who knew nothing of Ted's history, offered him the toxin deal. "Nothing easier," Ted thought: "Now that it is here, I'm going to drink the wretched stuff anyway; so I might as well intend to drink it — that way, not only will I get sick, I'll get a million dollars as well." The magazine reported that, at midnight, Ted intended to drink the toxin the following afternoon, even though he knew that he had then, and would have the following afternoon, absolutely no reason to drink the toxin. (He was convinced as well that he would have the usual excellent reason not to drink it — namely, that it would make him quite ill.)

If Ted was in fact able to win the money while we are not, what, exactly, accounts for that? A salient difference between Ted and the rest of us is that he has excellent grounds for confidence that he will drink the toxin. We do not. Indeed, because we are convinced that we will have no reason to drink the toxin when the time comes — and because our histories are quite unlike Ted's — we are confident that we will *not* drink the toxin.

Now, a number of philosophers have argued that there is a *belief* constraint on intending. Predictably, there is little agreement about just what the constraint is. For illustrative purposes, I shall mention a recent, modest proposal: One cannot intend to A while believing that one will not, or probably will not, A.[5] If there is such a constraint on intending (or on non-Freudian intentions, at least), we have a partial basis for a diagnosis of our difficulty as prospective prize winners in the toxin deal. We find it hard to intend to drink the toxin because we find it difficult not to believe that we (probably) will not drink the toxin. Probably, in our case, we find it difficult not to believe this because we

realize both that we lack (and will continue to lack) a reason to drink the toxin and that we have (and will continue to have) excellent reasons not to drink it. But one cannot infer from this alone that intentions — all of them — "are based on *reasons to act.*" Ted's intention, if in fact he has it, is not so based. Nor was the intention of the toxin drinker in the "forgetting" case.

3. TED'S PUZZLE AND ITS RESOLUTION

Can Ted really have intended at midnight to drink the toxin the following afternoon? As soon as he thought that he had formed the intention, he began to entertain serious doubts. His excellent reason for intending notwithstanding, Ted was unconvinced that he now intended to drink the toxin. An erstwhile philosophy major, he recalled a lesson that he learned long ago from Davidson (1980, p. 264) and Goldman (1970, p. 76):

PA. An agent intentionally A-s only if he A-s for a reason.

Knowing that he would have no reason tomorrow to drink the toxin, Ted concluded that he would not *intentionally* drink the toxin. And, he wondered, "Given that I am convinced that I won't intentionally drink the toxin, can I really intend to drink it? Doesn't intending to A preclude simultaneously believing that one won't intentionally A?"

When Ted recalled that his preferred belief constraint on intention made no mention of A-ing *intentionally*, he took some comfort in that thought. The constraint that he endorsed was that an agent who intends to A does not believe that he (probably) will not A; and that leaves it open whether the agent believes that he will not *intentionally* A. He himself did not believe that he would not drink the toxin, even though he found himself believing that he would not drink it intentionally. Moreover, Ted mused, he is not peculiar in this respect. His brother Fred, who once was offered the toxin deal, managed to slip the toxin into his favorite afternoon drink and later to forget all about the toxin — but not before he formed the intention to drink the toxin on the following afternoon while sipping his tea. While intending to drink the toxin, Fred believed both that he would not intentionally drink it and

that he would drink it. (Of course, Fred did not win the money, since he violated the prohibition against forgetting.)

Ted was not comforted for long, however. Reflecting on the point that he, unlike Fred, did not intend to drink the toxin unintentionally, Ted began to worry again. Knowing that he differed from Fred in this way, and convinced as well that he had now, and would have tomorrow, no reason to drink the toxin, could he really intend to drink it? Given these convictions and the lesson that he had learned from Davidson and Goldman (*PA*), Ted believed that he would not drink the toxin intentionally. He concluded that, inevitably, he would drink it *unintentionally*, as usual. And, he thought, setting aside intentions like Fred's (and Freudian intentions), forming an intention tonight to *A* tomorrow would commit one to rejecting the hypothesis that it is inevitable that one will not intentionally *A* tomorrow. He was sure that one cannot be as fatalistic as that about the inefficacy of one's intentions.

The story does not end here, of course. Ted recalled that his sister Wilma claimed to be so constituted that, once she formed an intention, it was sure to remain intact and unforgotten unless she either executed it or performed an intentional act of intention abandonment. Now, if, as Ted had learned in school, one cannot *A* intentionally in the absence of a reason for *A*-ing, Wilma cannot intentionally abandon an intention unless she has a reason for abandoning it. Of course, if Wilma were somehow able to form the intention to drink the toxin in order to win the money, she would have a great reason, Ted reflected, for abandoning the intention when the time for drinking came. But, he thought, *he* would have no such reason. His abandoning the intention to drink the toxin would gain him nothing, since he would drink it whether he intended to or not (and he has no aversion specifically to drinking toxins *intentionally*: that bother him no more than drinking toxins unintentionally). "So what if I am like Wilma?", Ted wondered. "If, today, I were to form the intention to drink the toxin tomorrow, I would be confident that I would have no reason to abandon the intention tomorrow, and, hence, that I would not abandon it. And, given that, I could conclude, after all, that I would drink it intentionally: My unabandoned intention would in all probability function like other intentions of mine for actions that I can easily perform and issue in an intentional toxin drinking." Ted's next thought was that he is, in fact,

like Wilma: He, too, never loses an intention unless he performs an intentional act of intention abandonment!

Ted, a reflective fellow, realized that he was faced with a dilemma of sorts. Owing to peculiar facts about himself and his situation and to his acceptance of *PA* (the thesis that an agent intentionally *A*-s only if he *A*-s for a reason), he can be confident that if he forms an intention to drink the toxin he will not abandon it, in which case (barring unforeseen obstacles) he will intentionally drink the toxin when the time comes. On the other hand, if *PA* is true, he will not intentionally drink the toxin, since he will have no reason to drink it. And if he is confident that he will not intentionally drink it, he is saddled with the problem confronted earlier.

Midnight was drawing near; Ted needed to think hard. It occurred to him that, *independently* of *PA*, he might be warranted in believing (*B*) that he would intentionally abandon an intention to drink the toxin only if he had a reason to abandon it. (According to the magazine article, this thought triggered reasoning that led Ted to suspect that *PA*, but not *B*, rested on an assumption about agency that did not apply to him. Unfortunately, the article was vague about this; and I shall try to reconstruct Ted's reasoning later.) So, Ted decided, he would try to form the intention and then think about the issue further. He said to himself, "Look, I have a great reason to form the intention; so I'll just form it — I'll settle upon drinking the toxin tomorrow at 12:01 p.m. in my kitchen. I'll stir the nasty stuff into a milk shake to make the taste more bearable and then I'll down it. Done." And at that point Ted felt entirely settled upon drinking the toxin (as settled as he had ever been on any course of action); and he was confident that he would drink it intentionally, knowing that he would have no reason to abandon his intention. Moreover, it was clear to him that he would have no reason tomorrow to drink the toxin. Thus, he was convinced, he satisfied the conviction condition, *C*; and he was confident that he had not violated any of the billionaire's additional conditions. Ted's worries were over.

4. WORRIES

Our worries, however, are not. If Ted intentionally drinks the toxin, then either *PA* (the claim that an agent intentionally *A*-s only if he *A*-s

for a reason) is false or, when the time for drinking comes, Ted has a reason for drinking the toxin. But if Ted has such a reason, and if it does not just come out of the blue, we should, in all fairness to the billionaire, expect Ted to realize that the reason will be present at the pertinent time, in which case he is ineligible for the prize — he will not have satisfied condition *C*. Is *PA* false? Can't Ted win the prize after all? Or what?

Some relatively mundane motivation for qualifying *PA* is provided by consideration of "subsidiary" intentional actions — e.g., routine actional parts of "larger" intentional actions — and by certain cases of double effect.[6] But these issues may safely be set aside for present purposes; for the difficulties that they raise leave open the following less demanding version of *PA*:

PA1. Every *intended* intentional action is done for a reason

And *PA1* poses essentially the problem raised in the preceding paragraph. If *PA1* is true, and Ted, intending to drink the toxin, intentionally drinks it, he drinks it for a reason.

A pair of worries require attention. First, can an agent intend to do something that he is convinced he will do whether or not he intends to do it? If not, then Ted cannot intend to drink the toxin after all. Second, is Ted's intention to drink the toxin itself a reason for drinking it? If so, then perhaps, in fairness to the billionaire, we should suppose that Ted realizes this, in which case he does not satisfy conviction constraint *C*.

Some philosophers have claimed that the intending agent makes a presumption of openness: The agent presumes that she has options, that more than one course of action is open to her (e.g., Donagan 1987, ch. 5). Suppose that this claim, literally interpreted, is true. Does it follow that Ted, given the conditions of the case, cannot intend to drink the toxin? Not at all. For, as Ted sees things, it is open to him to drink the toxin intentionally or unintentionally. If he intends to drink the toxin, there is every reason to suppose that he will drink it intentionally. If he does not intend to drink it, he believes (correctly, we may suppose), he will drink it anyway — but unintentionally. *Whether* he drinks it at all is not under his control. But whether he drinks it intentionally or unintentionally is, in his opinion, subject to his control. This, he believes, is open, and up to him.

Is Ted's intention to drink the toxin a *reason* for drinking it? Frequently, at least, intentions to *A* are not reasons for A-ing. Why, you ask me, did I turn my computer on a moment ago? "Because I intended to," I reply. My answer, assuming that I am trustworthy, does provide some information. It suggests that my turning my computer on was not an accident, or otherwise unintentional. But it does not identify a *reason* for which I switched it on. The fact that I intended to turn my computer on leaves open the matter of the reason(s) for which I so acted. My reason for turning it on, as it happens, was that I wanted to show my daughter how my computer worked and believed that I could do that only if I turned it on.

In some cases, however, an intention to *A* is perhaps plausibly viewed as a reason for *A*-ing. Consider two varieties of wants. Some things we want extrinsically, owing to their believed conduciveness to our doing, acquiring, etc., something else that we want. Other things we want intrinsically, for their own sakes or as ends. Thus, I might want to give John five dollars in order to repay a debt, thereby doing what (I think) is right; and I might want to do what is right, not for the sake of anything else, but as an end in itself. Intrinsic desires to *A* are candidates for reasons to *A* (Mele 1988; 1992, ch. 6). Similarly, I might, in a particular case, *intend* to do what is right, regarding my doing what is right, not as a means to something else, but as an end: I might *intrinsically* intend to do what is right. *Perhaps* my intrinsically intending to do what is right is a reason for doing what is right. So, perhaps, some intentions — intrinsic ones — are reasons for action. But this does not at all suggest that Ted's intention to drink the toxin was a reason for drinking it. That intention plainly is not an intrinsic intention. Ted does not regard his drinking the toxin as an end, as something to be done or pursued for its own sake. Rather, viewing his drinking the toxin neither as a means (or otherwise conducive) to an end of his nor as an end in itself, Ted has no reason to drink the toxin.

5. RECONSTRUCTING TED'S REASONING

It is time to reconstruct the reasoning on which the magazine article was vague. The first stage was Ted's forming a hypothesis about the grounds for *PAI*, the thesis that all intended intentional actions are

done for reasons. (He had already noticed that subsidiary actions and double effect pose problems for *PA*; but he saw as well that these problems did not provide him with an exploitable loophole.) His hypothesis was that *PA1* is based in part on the following pair of assumptions about intentional behavior.

> *A1.* An intended intentional *A*-ing is explicable *as an intentional A-ing* only in terms of a reason for *A*-ing that the agent has at the time.
>
> *A2.* All intentional actions are explicable *as* intentional actions.

Although Ted accepted *A2*, he wondered about *A1*. Ted endorsed another element of a standard theoretical package that includes *PA1* (and *PA*):

> *A3.* If an agent were explicitly and consciously to intend at *t* to *A* then and the intention were, in the absence of causal deviance and exceptional luck, to issue in an *A*-ing at *t*, that would suffice for the *A*-ing's being intentional; not only would the intention play a role in the production of the *A*-ing, but the role that it played would be such as to guarantee that the *A*-ing was intentional.

So, Ted thought, if it is possible that, in the absence of a reason to *A*, and agent nevertheless explicitly and consciously intends at a time to *A* then, it is in principle possible as well that an agent who has no reason to *A* nevertheless intentionally *A*-s — even granting *A2*.

At this point, Ted naturally turned to the question why one should think that intending to *A* depends upon having a reason for *A*-ing. He considered two hypotheses.

> *A4.* Given *A3*, the possibility of intending to *A* while having no reason to *A* would open the door to the possibility that *PA1* is false; and since *PA1* is a conceptual truth, that is a door that cannot be opened.
>
> *A5.* Forming an intention to *A* of a kind suitable to produce

> an intentional *A*-ing is causally dependent on having a reason to *A*. (In the absence of a reason to *A*, one might manage to form what amounts to an intention to *A* unintentionally, as in the forgetting case. But such an intention, if things go according to plan, will result in an unintentional *A*-ing.)

Ted realized that since invoking *A4* in support of *PA1* would be viciously circular, he was entitled to set that hypothesis aside. Concerning *A5*, he reasoned as follows. In normal agents, the conviction that one has (and will have) no reason to *A* and that one has (and will continue to have) a reason not to *A* renders any reasons that one might have for intending to *A* incapable of producing an intention to *A* of the kind at issue — as long as the conviction remains in place. While possessed of these convictions, the normal agent will be convinced as well (barring self-deception and the like) that *A*-ing is not a genuine option for her; and that, Ted thought, precludes forming an intention to *A*. But suppose that an agent is convinced that she will *A* whether she intends to or not and that *A*-ing is something that can, in principle, be done either intentionally or unintentionally. Suppose further that she does not intrinsically prefer *A*-ing unintentionally to *A*-ing intentionally (nor vice versa) and that she recognizes that neither is a *means* to anything that she desires. Neither of the two courses of action, then, in her opinion is intrinsically or instrumentally better than the other. Still, she must, she is convinced, follow one or the other of these paths.

Now, human agents, like Buridan's ass, occasionally find themselves in situations in which their two *best* options are equally good. Having no reason to prefer one to the other, they somehow manage to pick — sometimes on the basis of a randomizing procedure (e.g., a coin toss). And to "pick," in the relevant sense, is to form an intention. Assuming that the options have intrinsic or instrumental value for the agents, each option is supported by reasons, which is not the case in the scenario that Ted was considering. Still, Ted thought, although it is not up to one whether one *A*-s in a case in which one will *A* whether one intends to or not, that seems to leave it open whether one *A*-s intentionally or unintentionally. And if unintentional and intentional *A*-ing are equal with respect to intrinsic and instrumental value (zero in both cases,

when *A*-ing is Ted's prospective toxin drinking), one might pick on other grounds. In Ted's own case there is an excellent reason for "picking" intentional toxin drinking that is not also a reason to drink the toxin (nor to drink it intentionally). And if one can pick on the basis of extrinsic grounds — e.g., a coin toss — in other cases, surely Ted can pick on the basis of the $1,000,000 prize in his situation. Further, since to pick (in the sense at issue) a course of action is to form an intention, Ted can form an intention to drink the toxin. His forming the intention would not be causally inexplicable; for, in his case, nothing renders his reason to intend impotent vis-a-vis the formation of an intention to drink the toxin. That reason can bear a causal burden of the kind normally carried by reasons for *action* in the production of intentions.

PA1 rests partly on *A5*. But *A5* is false in some cases in which an agent will *A* whether he intends to or not. In those cases, one can form an intention fit to generate an intentional *A*-ing without having a reason to *A*; and, armed with such an intention, one can intentionally *A* without having a reason to *A*. Ted went to bed completely confident that the money would be his in the morning.

6. MORALS AND RATIONALITY

Given the various twists and turns in the preceding discussion, the morals that I have drawn require emphasis. What have we learned? First, even if, (1) typically, agents convinced that they have, and will have, no reason to *A* do not intend to *A*, the forgetting case and Ted's case show that we should not seek to explain this by supposing (2) that it is an essential feature of intentions to *A* that they are based on reasons to *A*. Again, Ted's intention to drink the toxin has no such basis. Second, the truth of (1) can be accounted for by the following conjunction: (3) typically, agents convinced that they have, and will have, no reason to *A* are convinced as well that they will not even attempt to *A*; and (4) setting aside Freudian intentions, an agent, *S*, intends at *t* to *A* at *t** (which may or may not be identical with *t*) only if *S* is not convinced at *t* that she will not even attempt at *t** to *A*. Thus, even if our convictions about what we will not even try to do are typically products of our convictions about our reasons, we need not suppose that intending to *A* is conceptually more tightly tied to having reasons for *A*-ing than to the

absence of a conviction that one will not even attempt to *A*. Third, an agent who has no reason to *A* and who is convinced at *t* that this is so may nevertheless perform an intended intentional *A*-ing at *t*.

Commentators on Kavka's toxin puzzle, as I have mentioned, generally set their sights on questions about rationality. A collection of questions may be raised about Ted in this connection. (Q1) Was it rational of Ted to intend at midnight to drink the toxin the following afternoon? (Q2) Was it rational of Ted at the time of action to intend to drink the toxin then? (Q3) Was it rational of Ted to drink the toxin? I consider each in turn. The investigation will pay dividends even beyond the sphere of rationality.

Q1. At midnight, Ted had an excellent reason to intend then to drink the toxin the following afternoon. Given that he was convinced (rightly, we may suppose) that he would drink the toxin whether he intended to or not, Ted had little or no reason *not* to intend at midnight to drink the toxin the next day. Thus, on the whole, Ted had better reasons for intending at midnight to drink the toxin the following afternoon than for not so intending. Hence, other things being equal, it was rational of Ted at midnight to intend then to drink the toxin the following afternoon.

Q2. At the time of action, Ted had no reason to drink the toxin and seemingly no reason to intend to drink it. On the face of it, he had a reason not to drink it — perhaps a pro-attitude toward his not being ill in conjunction with a belief that drinking the toxin would make him ill. Further, if he had a reason not to drink it, and if any reason not to *A* is a reason to *intend* not to *A*, Ted had a reason to intend not to drink the toxin. (Similarly, if any reason not to *A* is a reason *not to intend* to *A*, Ted had a reason not to intend to drink the toxin). And if Ted had a reason to intend not to drink the toxin (or a reason not to intend to drink it) while having no reason to intend to drink it, it was *not* rational of Ted at the time of action to intend to drink the toxin then. *However*, if, at the time of action, Ted had a reason to intend not to drink the toxin (or a reason not to intend to drink it), wouldn't he have had a reason to abandon his intention to drink the toxin? And his having a reason to abandon the intention would throw a big wrench into the works; for, in that case, Ted could have been expected to realize that he would have this reason (or a reason to this effect), which would reopen questions

raised earlier concerning the possibility of Ted's winning the money. A closer look at details relevant to Q2 is required.

Three questions need attention. First, is it possible that someone who is unable to *A* at *t* and who is convinced at *t* that he is unable to *A* at *t* nevertheless has a reason to *A* at *t*? (Here, "*A*" should be understood as ranging over "not doings" as well as doings: We can ask, for example, whether Ted, given the details of the case, can have a reason *not to drink* the toxin). Second, is every reason to *A* (or not to *A*), where *A* is an action, a reason to intend to *A* (or to intend not to *A*)? Third, is every reason not to A, where *A* is an action, a reason *not to intend* to A?

Having a pro-attitude toward *B*-ing while believing that one's *A*-ing is required for one's *B*-ing might seem sufficient for possessing a reason for *A*-ing. But is it? Suppose that Sam wants to win a certain million dollar prize and believes that he can do so only if he swims the English Channel, starting immediately. Does he have a reason for swimming the Channel, starting immediately? (Henceforth, "starting immediately" will be omitted for stylistic reasons, and should be supplied by the reader.) Suppose that Sam does not know how to swim, that he knows that he lacks this skill, and that he is convinced, consequently, that he is unable to swim the Channel. Once again, does Sam have a reason for swimming the Channel?

Here, opinions will differ. Some will assert, and others will deny, that Sam has a reason to swim the Channel. How might the issue be decided? Consider the following plausible assumption:

> *AR*. *S*'s possessing a reason to *A* suffices for *S*'s possessing a reason to attempt to *A*, unless *S* believes that her attempting to *A* would make it less likely that she *A*-s than would otherwise be the case. (Let us understand an attempt to *A* as an effort — no matter how slight — to *A*.)

Now, other things being equal, Sam has no reason to attempt to swim the Channel: There is, we may suppose, no premium on an attempt itself (no reward for trying) — in which case, given that he is convinced that he cannot swim the Channel, Sam has no reason to try to do so. Applying *AR* to our present question, we get the result that Sam has no reason to swim the Channel. This is not to deny, of course, that Sam *would* have a reason to swim the Channel, if he were to believe that he had a chance of success. But on what might be termed an ability-

sensitive conception of reason-having, Sam, given what he knows about himself, does not have a reason to swim the Channel.

On the same conception, Ted has no reason not to drink the toxin, given the details of the case (including his conviction that he cannot avoid drinking it). And if Ted has no reason not to drink the toxin, he has no reason to intend not to drink it, provided that there is no premium on his so intending (no reward for the intention itself). Whether Ted has a reason *not to intend* to drink the toxin — or a reason to abandon his intention to drink it — is, unfortunately, a distinct question, and requires discussion.

After the money is in his bank account, Ted, I shall grant, has no reason to intend to drink the toxin.[7] Typically, in virtue of not having a reason to intend to A, an agent would have a reason not to intend to A — or, alternatively, it would be *reasonable* of her not to intend to A. But, normally, an agent who has no reason to intend to A is not, at the time, possessed of an intention to A: Her reason for not intending to A is not a reason for intention-*revision*. In Ted's case, however, matters are different. Supposedly, he has an intention to A, even though he has no reason (any longer) for having that intention. Thus, any reason that he might have for not intending to drink the toxin would amount to a reason for revising what he intends. But, given the details of the case, any reason that Ted might have for the pertinent revision would rest on a reason not to drink the toxin. And on an ability-sensitive conception of reason-having, Ted has no reason not to drink the toxin. So, assuming that conception, Ted has no reason not to intend to drink the toxin.

Suppose that someone rejects AR and insists that reason-having is to be understood in such a way that agents sometimes have reasons to A, where A is an action, even though they know that they are unable to A. Given such a view, is it plausible that every reason to A (or not to A), where A is an action, is a reason to intend to A (or not to A) — or that every reason not to A is a reason not to intend to A? Consider Sam again; and suppose that just as there is no premium on his trying itself, there is none on his intending itself (that is, his intending to swim the Channel). Then even though, on the view in question, Sam has a reason to swim the Channel, he has no reason to *intend* to swim the Channel. For, Sam knows that the best such an intention can do is to initiate an

unsuccessful attempt to swim the Channel. Given the details of the case, there would be no point at all in intending to swim the Channel. But if not every reason to A (where A is an action) is a reason to intend to A (and not every reason not to A is a reason to intend not to A), then a proponent of the conception of reason-having presently under consideration who insists not only that Ted has a reason not to drink the toxin but also that he has a reason to *intend* not to do so needs to explain why we should agree. I do not see how the needed explanation can be forthcoming. For, other things being equal (e.g., there is no payoff of any sort for Ted's *intending* not to drink the toxin), there is no point in Ted's intending not to drink the toxin, since, as Ted knows, he will drink the toxin whether he intends to or not.[8] Nor, for similar reasons, is there any point in Ted's abandoning his intention to drink the toxin.

I conclude that Ted has, at the time of action, no reason not to intend to drink the toxin, and that his intending (at the time) to drink the toxin, therefore, is not positively irrational. However, I have granted that Ted no longer has a reason to intend to drink the toxin either: Owing to his having no reason to abandon the intention, he is simply stuck with it. If that is right, we can say that his having the intention at the time is not positively rational either and, perhaps, that it is nonrational.[9]

*Q*3. Now, for Ted's action: Was it rational of him to drink the toxin? Ted, I have argued, has no reason to drink the toxin. Does he have a reason *not* to drink it? On one of the conceptions of reason-having just sketched (the ability-sensitive one), he has no such reason. And in the absence of a reason not to drink the toxin, Ted's drinking it would not be positively irrational. On the other conception, Ted has a reason not to drink the toxin (since he has a pro-attitude toward his not being ill and believes that drinking the toxin would make him ill). Still, given that he cannot, as he believes, avoid drinking the toxin, there is, other things being equal, no point in his attempting to act on the basis of that reason. And it is not irrational of agents to refrain from attempts that they know to be pointless. Of course, one might claim that simply in virtue of Ted's having a reason not to drink the toxin while having no reason to drink it, his intentionally drinking it is irrational. Provided that we keep in mind the conception of reasons at work in this claim, I

have no objection. On this conception, as I have explained, one may (i) have a reason to A (or not to A), where A is an action, without having a reason to intend to A (or to intend not to A), and (ii) have a reason not to A without having a reason not to intend to A.

7. SUMMARY, REMOTENESS, AND IMPLICATIONS

In showing that the million dollars can be won, I resorted to a highly contrived case. My need to resort to a case of this sort is instructive. The problem that we (relatively) normal agents encounter in winning the money rests ultimately on the following point. Setting aside strategies for bringing it about that one unintentionally A-s, scenarios in which unintentional A-ing is likely, subsidiary actions, and double effect, normal agents who are convinced that they will have no reason to A when the time for A-ing comes, will also be convinced, if they are the least bit reflective, that they probably will not A, and that they will not even try to A. And having this latter conviction at t, on a plausible conception of intention, is incompatible with intending at t to A at t^*.[10] Given the constraints laid down by the billionaire, I needed a case in which the latter conviction is absent *even though* the agent is convinced that he would have no reason to drink the toxin. The possibility of a case such as Ted's shows that the money can be won. More importantly, it shows that not all intentions to A need be, in Kavka's words, "dispositions to [A] that are based on *reasons to* [A]" and that intentions to A that are not so based can issue in intentional A-ings.

Much of this paper was driven by the questions raised in my introduction. The answers that I have defended, briefly put, are as follows. An agent can have a reason to intend to A independently of having a reason to A. Such a reason can issue in and "rationalize" an intention to A even though the agent has, and continues to have, no reason to A; and such an intention can issue, in turn, in an intentional A-ing. These conclusions constitute constraints on any comprehensive theory of practical reasons. S's possessing reasons — even *effective* reasons — for the various positive practical attitudes toward a prospective A-ing by S (intending, wanting, etc.) does not in all cases depend upon S's possessing a reason for A-ing. Indeed, as I have argued, one can rationally intend to A in the absence of any reason to A. On the

positive side, the discussion points to a relatively tight connection between reasons for *A*-ing and reasons for intending to *A*, and to the priority of the former in normal cases. If the connection were only a loose one, a much less bizarre character than Ted could have won the $1,000,000.

To my mind, the most striking consequence of the arguments advanced here is that even *PA1* — a scaled down version of the traditional idea (*PA*) that an action's being intentional requires its being done for a reason — is *false*. (*PA1*, again, asserts that every *intended* intentional action is done for a reason.) Of course, strikingness is not only interest-relative but scheme-dependent. Like Ted, I thought that I had learned in school that *PA* (or at least something close) is true — indeed, a fundamental truth in the philosophy of action. I regarded it as a requirement for the explicability of intentional action, *qua intentional*. Others might have found it more surprising that an agent who knew that he had (and would have) no reason at all to *A* might nevertheless consciously and straightforwardly intend — even *rationally* intend — to *A*.

Some, however, might be insulated from surprise by methodological presuppositions. Thinking that cases instantiated only in relatively remote possible worlds are philosophically barren (or barren for the purposes of this paper, at least), they might insist that Ted can teach us nothing. Naturally, I reject the premise.[11] Since a fully developed argument for so doing cannot be advanced at the end of an already lengthy paper, I shall offer a brief, if indirect, defense. The defense will illuminate some of the implications of my results for traditional views about action.

Donald Davidson holds that "it is (logically) impossible to perform an intentional action without some appropriate reason" (1980, p. 264). He maintains as well that "an intention to act (or to refrain from acting) requires . . . a desire or pro-attitude toward outcomes or situations with certain properties, and a belief that acting in a certain way will promote such an outcome or situation" (1987, p. 40): In short, and in part, an intention to *A* requires a reason for *A*-ing. In Davidson's words, his "story about how beliefs and desires cause an action" — a story encompassing these two theses — "is arrived at . . . by reflecting on the nature of beliefs and pro-attitudes on the one hand, and on the nature

of action on the other" (ibid). This, of course, is a traditional approach, and a fruitful one.

Now, if there *might* be more to "the nature" of belief, desire, intention, action, and the like than can be uncovered without the assistance of thought experiments taking us to relatively remote possible worlds, then such experiments are properly included among the tools of the traditional approach. And on what grounds, one should wonder, can it be shown that the antecedent of the preceding sentence is false? After all, Ted harbors beliefs and desires, engages in practical reasoning, forms intentions, and acts. Careful, systematic reflection on the nature of action and the intentional attitudes cannot reasonably exclude from consideration a being like Ted. It is reflection on Davidson's own subject matter that reveals the falsity of significant elements in his story.

Much of the rest of Davidson's story might well be true. And Ted's case, partly in light of the apparent *need* to resort to a scenario so remote to find a winner, can be interpreted as speaking against *radical* revision of Davidson's account of the interrelationships among intentions, reasons, and actions (more on this shortly). My own inclination is to seek to contain the damage done to the Davidsonian picture by what Ted has taught us. But, in any case, it is worth knowing that intentional actions and intentions can be pried apart from reasons for action — and pried apart in a manner that allows some such intentions to be rational.

At one time, intentions attracted little attention in the philosophy of action. In Davidson's "Actions, Reasons, and Causes," a paper that did much to revive causal theories of action-explanation, we are told that such expressions as "the intention with which the driver raises his arm" refer to "no thing at all, neither event, attitude, disposition, nor object"; *reasons* shouldered the explanatory load (1963; 1980, p. 13). This has changed (owing partly to subsequent work of Davidson's); and the change is a healthy one.[12] But a consequence of the addition of intentions to a pristine belief/desire model of the explanation of action has generally been overlooked. Intentions, in some cases, may usurp the place of reasons for action, rendering them unnecessary for the explanation of certain intentional actions (*qua* intentional). The considerable ink required to show even that *Ted* could win the money indicates that the range of *intended* intentional actions not done for reasons is a small one — indeed, a range limited to instances in which agents are

convinced that they will *A* whether they intend to or not. If that is so, *PA1* is true of all cases in which the agent does not see the pertinent action as something that he will do regardless what he intends.

This last point may be reinforced. Normally, intended intentional actions are performed either as means to (or constituents of) ends or as ends themselves. *A*-ing as a *means* to an end is a paradigm case of *A*-ing for a reason. Doing something, *A*, as a *constituent* of an end, *E*, requires having *E* as a goal and taking *A*-ing to be a constituent of *E*. The goal-having and the taking together constitute a reason for *A*-ing; and the *A*-ing will be done *as* a constituent of an end only if it is done for some such reason. Finally, *A*-ing as an *end* requires the agent's having his *A*-ing as a goal I have argued elsewhere that an agent's having his *A*-ing as a goal or end is *itself* a reason (not always a *good* reason) for *A*-ing and that so-called "intrinsically motivated" actions — actions done "for their own sakes" — are done for such reasons (Mele 1988; 1992, ch. 6). If that is right, intended intentional actions are *not* performed for reasons only if they are performed neither as means to (or constituents of) ends nor as ends themselves. I was dubious about there being any such intended intentional actions — until I discovered Ted. One strange feature of his toxin drinking is that it does not fall into any of the normal categories for intended intentional actions. Those categories have seemed to many to be exhaustive — at least in the domain of *intended* intentional actions. The seeming exhaustiveness has a partial source in the assumption that agents do not see their prospective actions as things that they will do regardless of their intentions.

Ted has forced interesting modifications in a traditional account of intentional action; but he has left the main thrust of the account intact, and has even provided indirect support for the latter. That is a comforting thought. For if nothing *like* the traditional account were correct, intentional behavior would be even more far-fetched than Ted is.[13]

NOTES

[1] Donald Davidson, for example, contends that the "relation between the reasons an agent has for acting and his intention" is that "reasons cause the intention 'in the right

way' "; further, if the intention is effective, then "the intention, along with further events (like noticing that the time has come), causes the action 'in the right way'," or at least is "a causal factor in the development of the action" (1985, p. 221).

[2] Goldman 1970, p. 76. Goldman cites Anscombe 1958, p. 9. Cf. Davidson's claim that "it is (logically) impossible to perform an intentional action without some appropriate reason" (1980, p. 264).

[3] This constraint renders a certain kind of "integrity" useless in getting the money. Perhaps strong-willed people with a heroic commitment to "honoring" their intentions, whatever their intentions might be, could intend tonight to drink the toxin tomorrow, counting on the commitment to carry the day. But such people would have a reason to drink the toxin tomorrow, if they were to form the intention tonight — a reason partly constituted by (or based on) their commitment to honoring their intentions. And, insofar as they are counting on the commitment to carry the day, they are not (barring self-deception and confusion about the nature of reasons) convinced that they will have no reason tomorrow to drink the toxin.

[4] Incidentally, the forgetting case also falsifies another thesis sometimes advanced in the philosophy of action — namely, that any agent who intends to *A* intends to *A* *intentionally* (see, e.g., Ginet 1990, pp. 35–36).

[5] See Mele 1989 and 1992, ch. 8. An agent who believes that he (probably) will not *A* might, nevertheless, *hope* to *A* or intend to *try* to *A*; but these states are distinguishable from intentions to *A*.

[6] In Mele 1992, ch. 6, I argue that, at most, the considerations at issue force a modification of *PA* that preserves its essential spirit. (I also address "actions done for their own sakes" or as ends, and argue that they are done for reasons constituted by desires of a certain kind.)

[7] Harman has argued that "beliefs and intentions are subject to the following [principle]: One is justified in continuing fully to accept something in the absence of a special reason not to" (1986, p. 46). If this "principle of conservatism" is correct, Ted is justified in continuing to intend to drink the toxin even after the money is in his bank account. Acceptance of Harman's principle would make my task concerning *Q2* considerably easier.

[8] Recall that Ted is indifferent between drinking the toxin intentionally and drinking it unintentionally.

[9] Alternatively, if Harman's "principle of conservatism" (n. 7 above) is correct, Ted is rational (or justified) in continuing to have the intention.

[10] Again, I am setting aside Freudian intentions.

[11] In the context in which the toxin puzzle was spawned — investigation of the (ir)rationality and (im)morality of nuclear deterrence — a being like Ted might not be so far-fetched (as Gregory Kavka pointed out to me). Imagine that in the near future "our" president is convinced that, given the state of the world and predictable effects of our suffering a large-scale nuclear attack, our retaliating in kind would be inevitable, regardless of what he might endeavor to do at the time. (Communication would be difficult; surviving generals, he believes, would ignore any orders that he might give to refrain from retaliating anyway, if indeed he survived; and so on.) Convinced as well that there is a substantial positive payoff for forming and making known a deterrent intention for "us" to retaliate in kind to such an attack (though not for the retaliation itself), he might rationally form that intention, just as Ted rationally formed the intention to drink the toxin.

[12] The need for the change is a central theme in Mele 1992.

[13] I am grateful to Richard Foley, John Heil, Gregory Kavka, Bob Maydole, Brian McLaughlin, and Paul Moser for their comments on a draft of this paper.

REFERENCES

Anscombe, G. 1958, *Intention*. Ithaca: Cornell University Press.
Davidson, D. 1963, 'Actions, Reasons, and Causes,' *Journal of Philosophy* 60, pp. 685—700. Reprinted in Davidson 1980.
Davidson, D. 1980, *Essays on Actions and Events*. Oxford: Clarendon Press.
Davidson, D. 1985, 'Replies to Essays I—IX,' in B. Vermazen & M. Hintikka, eds., *Essays on Davidson*. Oxford: Clarendon Press, pp. 195—229.
Davidson, D. 1987, 'Problems in the Explanation of Action,' in P. Pettit, R. Sylvan, & J. Norman, eds., *Metaphysics and Morality: Essays in Honour of J. J. C. Smart*. Oxford: Basil Blackwell, pp. 34—49.
Donagan, A. 1987, *Choice*. London: Routledge & Kegan Paul.
Ginet, C. 1990, *On Action*. Cambridge: Cambridge University Press.
Goldman, A. 1970, *A Theory of Human Action*. Englewood Cliffs: Prentice-Hall.
Harman, G. 1986, *Change in View*. Cambridge: MIT Press.
Kavka, G. 1983, 'The Toxin Puzzle,' *Analysis* 43, pp. 33—36.
Mele, A. 1988, 'Effective Reasons and Intrinsically Motivated Actions,' *Philosophy and Phenomenological Research* 48, pp. 723—731.
Mele, A. 1989, 'Intention, Belief, and Intentional Action,' *American Philosophical Quarterly* 26, pp. 19—30.
Mele, A. 1992, *Springs of Action*. New York: Oxford University Press.

Department of Philosophy
Davidson College
Davidson, NC 28036
USA