**So why can’t you intend to drink the toxin?**

Fernando Rudy-Hiller

rudy@filosoficas.unam.mx

The Version of Record of this manuscript has been published on August 21, 2019 and is available in *Philosophical Explorations* <https://doi.org/10.1080/13869795.2019.1656280>. Please cite from published version.

**1. Introduction**

Philosophers of action usually think that sustained reflection on Gregory Kavka’s Toxin Puzzle promises to reveal crucial features of the attitude of intention itself. In particular, they think that it reveals metaphysical and/or rational constraints on the formation of prospective intentions through deliberation.[[1]](#endnote-2) The difficult question, and the source of disagreement, has been to pin down exactly what these metaphysical and rational constraints are. The scenario is this:

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, 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. (Kavka 1983, 33-4)

It is widely, though not universally,[[2]](#endnote-3) agreed that ordinary rational agents[[3]](#endnote-4) would be unable to form the required intention “in the normal way,” i.e., just by making a decision,[[4]](#endnote-5) despite the fact that they have excellent reasons for doing so. However, as Kavka points out, this is puzzling, since “You are asked to form a simple intention to perform an act that is well within your power” (1983, 35). The challenge is to explain what it is about the nature of intention that seems to make it impossible to respond to the billionaire’s offer in the required way. In this paper I will propose a novel account of the impossibility for ordinary agents to win the prize in the toxin scenario. I will argue that the following principle explains why this is so:

**Motivating-Explanatory Reason Principle (MERP).** An agent cannot intend at t1 to φ at t2 if she believes at t1 that she won’t have at t2 a *motivating* reason that can explain her intentionally φ-ing at t2.

Since any ordinary agent placed in the toxin scenario will believe at the time of deliberation that the next day at noon she won’t have any motivating reason that could explain her intentionally drinking the toxin, she cannot intend to drink it in the first place. Thus, I will argue that MERP is a genuine metaphysical constraint on the formation of intentions through deliberation and what ultimately explains what is going on in the toxin scenario.

Many authors have thought that the key to solving the puzzle lies in the fact that there is a close link between intentions and reasons for acting as intended. But, as we’ll see in sections 2 and 3, their conception of this link is problematic. Partly in response to this, other authors have argued that the basic link isn’t between intentions and reasons, but between intentions and beliefs about acting (or at least trying to act) as intended. I will argue, however, that the link intention-belief about trying doesn’t provide the ultimate explanation of the puzzle, because that link is a consequence of the more fundamental connection between intentions and motivating reasons that MERP highlights (subsection 4.1). I’ll then provide an argument in support of MERP and explain in detail what the latter requires of agents (subsection 4.2). Finally, I will show that my account of why ordinary agents can’t, but extraordinary agents such as Mele’s (1992) Ted can, form the required intention in the toxin scenario isn’t *ad hoc* but, on the contrary, provide valuable insights about often overlooked aspects of rational agency (subsection 4.3). I close by summing up the paper’s main findings (section 5).

**2. False starts**

All available explanations of the impossibility of forming the required intention in the toxin scenario appeal to the close link between intention and action. However, they differ on how to exploit this link in order to provide a solution to the puzzle. One popular way is this. Since intention is intimately connected to action, reasons for forming an intention must be reasons to act as intended. If so, the problem in the toxin scenario resides precisely in the fact that the reason the agent has for *forming* the intention to drink the toxin is not a reason to *drink it*, and thus it isn’t a reason *of the right kind* for forming the intention. I will call this diagnosis the Right Kind of Reasons (RKR) account (Hieronymi 2005, 2006; Shah 2008).[[5]](#endnote-6) The RKR account thus proposes the following metaphysical constraint on deliberation-based intentions:

**Right Kind of Reasons.** An agent can form through deliberation at t1 the intention to φ at t2 *only* by considering at t1 reasons for φ-ing at t2.

As several philosophers have argued (Pink 1991; Gauthier 1998b; Clarke 2008; Schroeder 2012; Morauta 2010), the problem with the RKR account is that it’s not only possible, but sometimes even rational, to form intentions in response to what Bratman (1999) calls “autonomous benefits”, that is, benefits related to the formation, rather than the execution, of a particular intention. For example, one can form an intention to φ not because φ-ing is decisively favored by one’s reasons butbecause one correctly believes that such an intention will remain firmer than any available alternative and one wants to avoid a mismatch between one’s present intentions and one’s future actions (Pink 1991, 352). Or one may form an intention to help someone at a later time because one correctly believes that so intending (and successfully communicating it) is the only way in which one will be helped now (Gauthier 1998b), regardless of whether one actually reciprocates. Or, finally, one can intend to φ at a later time because deciding now between φ-ing or Ψ-ing will allow one to successfully put the issue out of one’s mind and carry out an important task presently (Clarke 2008, 196; Morauta 2010, 219). Given that it seems possible and even rational to form intentions based on autonomous benefits, the RKR account can’t be the right explanation of the impossibility to intend to drink the toxin. The problem with the billionaire’s offer, then, can’t be that it forces the agent to consider autonomous benefits in deliberation.[[6]](#endnote-7)

A natural amendment to the RKR account is to argue that one can respond to autonomous benefits for forming an intention *as long as* one also has reasons in favor of acting as intended. Since in the toxin scenario there aren’t such reasons, this might explain the impossibility of winning the prize.[[7]](#endnote-8) Thus, it might be thought that the following principle provides a genuine metaphysical constraint on deliberation-based intentions:

**Normative Reason Principle (NRP)**. An agent cannot intend at t1 to φ at t2 if she believes at t1 that she won’t have at t2 a *normative* reason to φ at t2.

Kavka (1983, 35) himself thought that NRP offered the correct solution to his puzzle, and a number of other authors have endorsed this principle as well.[[8]](#endnote-9) This principle has the virtue of reconciling what is initially attractive about the RKR account—the tight link between intention and action—without overreaching and thus precluding the possibility of appealing to autonomous benefits in deliberation. However, as Mele (1992) has shown, NRP can’t be the final word about the puzzle because there are agents who, despite violating this principle, are clearly capable of forming the required intention. I’ll review Mele’s argument in some detail since my proposal will take its cue from the counterexample to NRP he presents.

**3. Mele’s account**

Mele’s argument against NRP relies on the case of Ted, an agent with the following pair of unusual characteristics. First, a strange curse has befallen him which causes him to drink any vial of toxin he encounters. The only thing that is up to Ted is whether to drink the toxin intentionally or unintentionally. Second, Ted’s intentions always persist and are executed unless he actively revokes them for a sufficient reason. According to Mele, the conjunction of these two unusual features makes it possible for Ted to win the prize offered by the billionaire, even though Ted will have, and believes he will have, no normative reason for drinking the toxin when the time comes—thus proving NRP false.

Here’s why Ted can win the prize. He knows, at the time of deliberation, that if the billionaire places the toxin near him tomorrow at noon he will inevitably drink it. The only thing that is up to him is whether to do so intentionally or unintentionally, and normally he is completely indifferent between either possibility. However, the billionaire’s offer constitutes an excellent reason for forming tonight the intention to drink the toxin intentionally tomorrow, so in this case Ted does prefer to form the intention to drink. Moreover, his intention will persist and be executed unless he actively revokes it for a sufficient reason, but he also knows that tomorrow at noon he won’t have any reason to revoke it, given that, whether he intends to or not, he will drink the toxin. It’s true that he is also aware that he will lack a reason to drinkit, but that doesn’t matter, given that his intention will still be in place when the time of drinking comes and so he will in fact drink.[[9]](#endnote-10) So given that Ted fully believes that if he intends to drink the toxin he will do it intentionally, and given that he has excellent reasons for so intending, in *his* case forming the required intention is no more difficult than it is for the rest of us to form garden-variety intentions concerning actions we know we will perform if we so intend.

I agree with Mele that this example shows that NRP is neither a metaphysical nor a rational constraint on intentions, for Ted not only is able to form the intention to drink the toxin while having no normative reason to drink it, but his intention is rational as well. Thus, being false, NRP can’t be what ultimately explains why ordinary agents can’t win the prize in the toxin scenario.[[10]](#endnote-11) Mele’s proposal is that what ultimately solves the toxin puzzle is what I’ll call the “Belief About Trying Principle”:

**Belief About Trying Principle (BATP).** An agent cannot intend at t1 to φ at t2 if she believes at t1 that she will not even try to φ at t2.[[11]](#endnote-12)

Mele claims that his argument against NRP provides direct support for BATP, for it shows that the reason why Ted *can* win the prize despite having (like the rest of us) absolutely no normative reason for drinking the toxin is that Ted (unlike the rest of us) lacks the conviction that he won’t even try to drink.[[12]](#endnote-13) Thus, it must be that what is really incompatible with intending to φ is not the absence of normative reasons in favor of φ-ing but, rather, the conviction that one won’t even try to φ when the time comes.[[13]](#endnote-14)

**4. A new account of the toxin puzzle**

I think we should resist Mele’s claim that the best explanation of the toxin puzzle must appeal to the agent’s beliefs about what she will attempt to do rather than to her reasons for acting. I’ll now offer a new account of the puzzle that shows how we can do this while avoiding the pitfalls of the RKR and the NRP accounts. The crucial difference between these accounts and mine is that, in my view, the essential connection doesn’t obtain between intentions and normative reasons but between intentions and motivating reasons. Since the latter view is compatible both with the absence of normative reasons for the intended action and with the possibility of considering autonomous benefits in deliberation, my proposal isn’t subject to the objections that beset the NRP and RKR accounts.

I’ll proceed by showing that Mele is mistaken in thinking that Ted’s case directly supports BATP as the ultimate explanation of the toxin puzzle. In effect, although Ted’s case is indeed a counterexample to NRP, it isn’t a counterexample to the following metaphysical constraint on deliberation-based intentions:

**Motivating-Explanatory Reason Principle (MERP).** An agent cannot intend at t1 to φ at t2 if she believes at t1 that she won’t have at t2 a *motivating* reason that can explain her intentionally φ-ing at t2.

If this principle is a genuine constraint on deliberation-based intentions, and if it turns out to be more basic than BATP, it follows that this principle, and not BATP, is what ultimately solves the toxin puzzle. My argument will proceed in three steps: first I explain why Ted satisfies MERP (4.1), then I defend MERP directly by offering reasons to accept it as a genuine metaphysical constraint on intentions (4.2) and, finally, I address several questions raised by my account (4.3).

***4.1 Step one: why Ted satisfies MERP***

Mele convincingly argues that Ted can form the intention to drink the toxin, and later intentionally drink it, while having no reason to do so. It’s clear that the sense of “reason” Mele has in mind is “normative” or “justifying” reason: there is no justification for Ted’s intentionally drinking the toxin, since by that time either he has won the prize or not and he lacks an intrinsic desire to drink toxin. I’ll argue, however, that Ted does have a reason to drink the toxin, but it isn’t a normative reason; it’s, rather, a *motivating* reason that *explains* his intentionally drinking the toxin. And it’s precisely because at the time of deliberation Ted knows that this motivating reason will be available to him at the time of action that he can form the intention to drink in the first place. This suggests that Ted lacks the conviction that he won’t even try to drink *because* he lacks the conviction that he will have no motivating reason for intentionally drinking. In other words, Ted satisfies Mele’s BATP *because* he satisfies MERP.

To begin with, let me say a word about the distinction between normative and motivating reasons. Following Michael Smith (1994, 95), I understand it as indicating two different ways in which actions can be rendered intelligible. On the one hand, normative reasons make actions intelligible by *justifying* them, i.e., by showing that they are desirable or required from the perspective of a particular “normative system” (rationality, morality, prudence, etc.). Thus, normative reasons are truths about which actions are desirable or required by specific normative systems.[[14]](#endnote-15) On the other hand, motivating reasons make actions intelligible by *explaining* them, i.e., by uncovering their causes. Of course, there are many different ways of explaining actions in causal terms; in the case of motivating reasons, the relevant causes are the agent’s motives for acting as she did.[[15]](#endnote-16) Thus, motivating reasons render actions intelligible by showing what moved the agent to act (which is usually, but not always, the pursuit of a goal. See below). It follows that motivating reasons are psychological states that play an explanatory role in the production of action (Smith 1994, 96). While the default assumption is that the relevant psychological state is a belief-desire pair (Davidson 1980), I think it’s more appropriate to cast it more generically as a cognitive attitude-conative attitude pair, so as to not prejudge in advance the kinds of cognitive and conative attitudes that can constitute motivating reasons. This will be important below.

With this distinction between normative and motivating reasons at hand, let’s ask: What motivating reason could Ted possibly have for intentionally drinking the toxin, given that doing so will gain him nothing and will only make him ill?

Here’s my proposal. Consider the following: given his strange disposition, Ted knows that, when the time of drinking arrives, he will drink the toxin either intentionally or unintentionally. Normally, he lacks both normative and motivating reasons for drinking either way, but this time around is different, since he has formed an intention to drink and has no (normative) reason to revoke it, because, again, he will end up drinking the toxin regardless of what he intends to do. Thus, he knows that his intention to drink will persist and issue in his intentionally drinking the toxin. Given these features of Ted’s case, he has a ready explanation of why he does so. What I want to claim is that this explanation uncovers a motive Ted has for drinking, and that this motive amounts to the motivating reason for which he acts. To see why this is so, consider the following schema for the explanation of actions:

Agent *S* φ-s intentionally because *m*

where “*m*” stands for the motive the agent has for acting as she does (which is normally a further goal of hers). In Ted’s case, the schema takes this form:

Ted drinks the toxin intentionally because *m*

What’s “*m*” here? Presumably, “Ted formed the intention to drink it and has no reason to revoke it” or, alternatively, “Ted decided to drink intentionally and has no reason not to do it.” My claim is that the motive Ted has for drinking the toxin is, in effect, that he has the intention of doing so and no reason to revoke it, and that this motive constitutes the motivating reason that explains why Ted drinks the toxin intentionally. Moreover, it’s natural to suppose that, at the time of deliberation, Ted will be aware that this explanation will be available to him at the time of action; that is, Ted knows that he will have an answer to the question “why am I intentionally drinking this?” in terms of such motivating reason.[[16]](#endnote-17) And I suggest that it’s *because* he knows at the time of deliberation that he will have an answer to this question that he lacks the conviction that he won’t even try to drink when the time comes—thus fulfilling Mele’s BATP—and is therefore able to form the intention to drink.[[17]](#endnote-18) By contrast, it’s precisely because we ordinary agents would lack an answer to this question,[[18]](#endnote-19) and because we would be aware of this fact at the time of deliberation, that we are incapable, unlike Ted, to form via deliberation the intention to drink the toxin. These considerations not only show that Ted actually satisfies MERP—since he doesn’t believe he lacks motivating reasons that can explain his intentionally drinking the toxin—but, more importantly, they also show that MERP is a more fundamental constraint on intentions than BATP. So if MERP proves defensible in its own right, it, and not BATP, would provide the ultimate explanation of the toxin puzzle.

***4.2 Step two: The rationale behind MERP***

I readily admit that my suggestion that Ted’s strangely-formed intention constitutes a motivating reason to drink the toxin, and my claim that the fact that he can explain his action in these terms but we don’t ultimately solves the puzzle, aren’t easy pills to swallow. To convince you that this isn’t an *ad hoc* story that works (if it works at all) only in this case, I’ll now provide the rationale behind my proposal.

My central claim throughout—the claim that MERP embodies—is that *sensitivity* to whether actions can be made intelligible in terms of motivating reasons is constitutive of deliberation-based intentions. Given that, as explained above, there is a conceptual link between motivating reasons and explanations of actions, the claim is that sensitivity to explanations (of the appropriate kind, i.e., in terms of the agent’s motives) for the intended action is constitutive of deliberation-based intentions. Note that I talk of sensitivity to, rather than conscious awareness of, explanations. The point of this distinction is to make clear why the constitutive condition on intention I’m postulating isn’t cognitively taxing.[[19]](#endnote-20) In effect, MERP doesn’t say that it’s constitutive of intentions that the agent has available during deliberation a motivational explanation for her future action, much less that she is occurrently aware of one; it only says that intentions can’t be formed in the presence of a belief that there is no such explanation at all for her intended action. This is why MERP isn’t cognitively taxing: it doesn’t require agents to be actively on the lookout for explanations of their actions; rather, it only requires them to be sensitive to *absences* of possible explanations in terms of motivating reasons. This is a pretty modest and garden-variety psychological ability: the ability to detect when the pursuance of a course of action would be unintelligible given that one lacks, and believes one will lack at a future time, appropriate motivations for it.

I do think, however, that this sensitivity to absences of motivational explanations is accompanied by the ability to become aware of such explanations if need be. This ability is often manifested in offering (to oneself or others) *post hoc* explanations for one’s conduct, but it can also be exhibited during deliberation. This is exactly what happens in the case of Ted, who, given the unusual nature of the billionaire’s proposal, must take a moment to reflect on what he has been asked to do. But, even in this case, it should be clear that Ted won’t have to go through an unusually cognitively taxing process of detecting motivations and explanations for his action. Rather, it’s the ordinary thinking process normal agents engage in when, for some reason, doubts arise as to whether a certain course of action is intelligible or makes sense to them.

At this point, however, one may rightly wonder whether the notions of intelligibility and explanation at play in Ted’s case are the same notions employed in ordinary contexts. Two worries are pressing here.[[20]](#endnote-21) The first worry is that, in ordinary contexts, the notion of intelligibility seems to be centrally linked to normative, rather than motivating, reasons. The second worry is that, even when we explain actions purely in terms of motives, pointing to one’s motivations not always makes actions intelligible in any robust sense, for example if one’s action is the product of an addictive craving, depression, an incorrigible desire to do the bad thing, a simple whim or, in Ted’s case, a very peculiar intention. So even if *some* motivational explanations make actions intelligible, not *all* of them do.

In response to the first worry, it’s important to recall that the distinction introduced above between normative and motivating reasons marks two different ways of rendering actions intelligible, that is, two ways of answering “why-questions” (i.e., “why are you doing this?”). Why-questions can be answered by invoking normative reasons, in which case the answer renders the action intelligible by *justifying* it. But they can also be answered by appealing to motivating reasons, in which case the action is rendered intelligible by *explaining* the motives behind it—what moved the agent to act. (Of course, if all goes well, normative and motivating reasons work in tandem, i.e., what justifies the agent’s action is importantly related to her motives for acting [Smith 1994, 131-2].) It’s plausible to suppose that agents usually answer why-questions by appealing to the normative reasons that, in their view, support their actions, but this isn’t always so. For instance, one can try to understand why one did something without trying to justify oneself for doing it, and the same is true regarding the actions of others. Think, for instance, in cases in which people are anxious to know the motives behind a heinous crime because this knowledge would render it intelligible by explaining why it happened, i.e., what moved the criminal to act, without justifying it at all. So even if it were true that, statistically speaking, why-questions are more frequently answered by offering justifications rather than explanations, what is crucial for my purposes is that both are legitimate and distinguishable ways of rendering actions intelligible.

Now even if the last point is conceded, one may still worry that sensitivity to explanations for one’s actions of the kind MERP requires is just a byproduct of the “normative vigilance” well-functioning rational agents exercise by nature, i.e., their disposition to ensure that their actions are supported by (normative) reasons.[[21]](#endnote-22) In other words, sensitivity to explanations of actions would be parasitic to sensitivity to justifications. Two things are pertinent to note in response. First, even if this were the case, it’s crucial to bear in mind that I’m not concerned with the question of what the defining features of rational agency are (one of which may indeed be normative vigilance), but with the narrower task of discovering the constitutive features of deliberation-based intentions, and in fact with the even narrower task of identifying which of these constitutive features explains the toxin puzzle. So while it’s very plausible to suppose that well-functioning rational agents usually exercise normative vigilance, for our current purposes the key question is whether they must necessarily do so in order to form intentions through deliberation. And the answer to the latter question is clearly “no”: in addition to the evidence provided by Ted’s case, phenomena like weak-willed intentions (Wallace 2001) and intentions formed in pursuit of the bad (Stocker 1979) offer excellent reasons for rejecting the idea that, in intending to φ, agents must see φ-ing as justified. Thus, exercising normative vigilance isn’t constitutive of intentions. Second, and more importantly, clear-eyed pursuers of the bad and clear-eyed akratics can still be sensitive to motivational explanations for their actions—thus exhibiting what we can call “motivational vigilance”—despite failing to be normatively vigilant, which suggests that these are in fact independent agential capacities.

Regarding the second worry mentioned above, about whether explanations of actions necessarily make them intelligible, it’s essential to be clear exactly what kind of intelligibility is at stake here. I’m *not* claiming that intelligibility, understood as a robust form of self-understanding involving either wholeheartedness (Frankfurt 1988b) or a coherent self-conception (Velleman 2006), is constitutive of deliberation-based intentions. Rather, I’m concerned with a thinner notion of intelligibility related to the explicability of actions in terms of motives, regardless of whether the agent endorses them, finds them good, or sees them as contributing to a unified picture of herself. Although thin, this is a perfectly legitimate form of intelligibility, one that is often invoked in everyday life—think again of wanting to understand a person’s motives for performing a heinous crime. Also, it must be noted that this thin notion of intelligibility as motivational explanation is the basis of more robust notions of intelligibility involving deep self-understanding. In effect, in order to wholeheartedly endorse one’s motives or to see them as constituting a coherent self-conception, one has to know first what those motives are. This is important because it shows that the type of intelligibility that, in my view, explains why Ted can form the intention to drink the toxin is an ordinary notion of intelligibility present in a broad swath of cases and not only in the toxin scenario.

Now when one’s own actions are concerned, it’s true that the thin kind of intelligibility just described may fail to yield robust self-understanding. In other words, coming to know one’s motives for performing a certain action can leave oneself as mystified as before, or even more if one finds the motives themselves quite incomprehensible, such as in cases of desiring the bad or cases of whimsical, weak-willed, or addictive actions. But this possibility doesn’t create trouble for my account because, again, I don’t claim that it’s constitutive of deliberation-based intentions that agents acquire robust self-understanding in forming and acting on them. All I claim is that it’s constitutive of deliberation-based intentions that agents don’t believe that there aren’t any motivational explanations for their future actions.[[22]](#endnote-23) And this seems to be the right result, since weak-willed and whimsical agents, addicts, depressives and the like can all form future-directed intentions despite the fact that they often fail to understand themselves in a robust sense, provided that they still exhibit motivational vigilance.

Of course, even if all of the foregoing is correct and it’s true both that the requirement on intentions MERP postulates isn’t cognitively taxing and that the notion of explanation invoked by it is a legitimate one, we still need positive reasons for accepting MERP as a genuine metaphysical constraint on deliberation-based intentions. What can be said in its defense?

A plausible defense is two-pronged. It appeals both to a modest requirement of self-knowledge and to the possibility of intending actions one doesn’t see as justified. Regarding self-knowledge, when an agent deliberates and forms future-directed intentions, she is (at least implicitly) conscious of the fact that she herself is going to implement them and, consequently, she has to be sensitive to whether she could see herself performing the intended action, i.e., whether it makes sense to her. (As I explained above, this doesn’t require her to be actively on the lookout for possible rationalizations for her actions but, more modestly, simply to be alert to the possibility of the absence of possible rationalizations.) This is because, even though intentions are conduct-controlling attitudes, the kind of control they exert is *rational* control: it isn’t as if in forming an intention the agent goes into “autopilot mode,” executing the intended action through a ballistic process that can’t be stopped once initiated.[[23]](#endnote-24) Rather, she retains the ability to evaluate whether the course of action she has set for herself continues to make sense to her up to the time of implementation (Bratman 1999, 72-3). But, and here’s where the second prong comes in, “making sense” shouldn’t be understood necessarily in terms of normative reasons, given that it’s a fact about human agency (and not only about Ted) that people can form intentions even though they are perfectly aware that their intended action isn’t all-things-considered justified, and even if they are aware that it lacks justification altogether. This suggests that the constitutive form of rational control that intentions embody shouldn’t be construed as normative vigilance but as motivational vigilance. That is, to be able to form and sustain intentions, agents must be capable of detecting when their prospective actions couldn’t be explained (or fail to make sense) in terms of their motivations. This entails that the basic, constitutive link between intention and action goes through the agent’s motivations rather than through her normative reasons. If this is right, then a metaphysical constraint on deliberation-based intentions must be that intentions can’t be formed in the presence of a belief that this basic link between intention and motivation has been severed. And this is exactly what MERP postulates, so this is why we should accept it as a genuine metaphysical constraint on intentions.

Returning to Ted, one may doubt whether we really observe in his case the link between intention and motivation that, I just claimed, is constitutive of deliberation-based intentions. One could object that while it’s true that Ted will have an answer to the question “why am I intentionally drinking this?”, the explanation of his action will appeal to his dispositions (drinking any vial of toxin he encounters, never dropping an intention for insufficient reasons) rather than to his motivations. So it might be thought that motivating reasons needn’t figure in Ted’s explanation for his drinking the toxin intentionally. Since we can concede to Mele that Ted really is capable of forming the intention to drink, this would show that MERP isn’t a genuine metaphysical constraint, for Ted would be able to form the required intention even if he believes that he lacks motivating reasons for drinking.[[24]](#endnote-25)

In response, it must be noted that if Ted really is capable of *intending* to drink, as opposed to *predicting* he will drink, the explanation he gives himself of why he will do so can’t appeal solely to his dispositions, in particular to his sub-agential disposition to drink any vial of toxin he encounters. Generalizing this point: if the only explanation an agent has for her future φ-ing is an explanation in terms of a causal mechanism that bypasses her agency, then the agent can’t intend to φ but only predict that she will φ at some point in the future. The reason for this goes back to the kind of control intentions afford, i.e., rational control. As I explained above, this kind of control entails that agents must be sensitive to whether their behavior makes sense in terms of, at least, motivating reasons. But sub-agential dispositions aren’t themselves motivating reasons, so if they are the sole explanation for one’s action then one lacks rational control with respect to it. On the other hand, predicting one’s conduct not only doesn’t presuppose rational control over the predicted action, but is in fact compatible with *certainty* that one will lack rational control then and there. Therefore, if Ted really is capable of intending to drink and not only of predicting he will drink, the explanation for his action can’t appeal solely to his sub-agential disposition but must include rational considerations as well. But this doesn’t mean that his peculiar sub-agential disposition to drink toxin plays no role whatsoever. It does: awareness of this disposition is what *enables* Ted to form the required intention by ensuring that he won’t have a sufficient reason to drop it later on (I expand on this point in 4.3 below).

Finally, I have argued that Ted’s intention to drink is itself his motivating reason for drinking (and thus the rational consideration that explains his drinking). Now it’s true that Ted doesn’t drink the toxin as a means to accomplish something else he wants (for the money is already in his bank account at the time of action), nor does he drink it for its own sake (he isn’t a masochist). Since motivating reasons typically have either of the two contents just mentioned, i.e., they represent the action as a means or as an end, Ted’s motivating reason is peculiar. But, if what I argued above is correct, since he is nevertheless capable of forming the intention to drink and not only of predicting he will drink, there must be an explanation in terms of reasons for his action. Given that Ted lacks any desire, either intrinsic or instrumental, for drinking the toxin, then the only option is that what explains his intentionally drinking it, and thus what constitutes his motivating reason for doing so, is his intention to drink itself. Typical motivating reasons are constituted by belief/desire pairs but, in Ted’s case, the desire to drink is replaced by his intention to drink. This is why in his case the intention, plus the belief that the stuff in front of him is toxin, *constitutes* the motivating reason in question. Intention, like desire, is an attitude with world-to-mind direction of fit, so it’s plausible to suggest that in certain cases it can occupy the “conative attitude spot” usually reserved for desires in the constitution of motivating reasons.[[25]](#endnote-26)

Although peculiar, Ted’s motivating reason isn’t unique. Ordinary agents can have as their sole motivating reason for acting an intention of theirs, and this motivating reason may fail to represent the intended action either as a means or as an end. Acting out of sheer willfulness or blind self-assertion—doing something merely because one has formed an intention to do it, for no particular reason—aren’t paradigms of rational action, but are nevertheless realistic possibilities for human agents (Wallace 2001, 8). So we do well to broaden our psychology of intention and action to encompass such possibilities, independently of discussions about the toxin puzzle.

***4.3 Step three: Some further questions about Ted***

Even if you accept MERP as a genuine constraint on intentions, and accept as well my contention that it explains why Ted can, but ordinary agents can’t, form the required intention in the toxin scenario, some further questions remain. The gist of them is whether key features of Ted’s case are genuine (and not merely conceivable) possibilities that can be generalized to more standard agents. I’ll address three such questions.[[26]](#endnote-27)

The first concerns whether it’s really possible for an agent to intend to φ when she believes that she will φ regardless of whether she intends it or not. Call this *the problem of* *intending the inevitable*. Ted correctly believes that he will drink the toxin regardless of what his intentions are, and so one might think that his intention to drink can’t be the cause of his drinking. Moreover, and unless he is self-deceived, he will be aware of this fact at the time of deliberation, and so it might seem that he can’t intend to drink after all; he can only predict he will.

Mele himself considers this objection and offers what I see as a sensible answer. He argues that even if we concede that intending to φ necessarily involves a presumption of openness regarding one’s φ-ing, this doesn’t preclude Ted from forming the required intention because, while it’s true that whether he drinks the toxin at all isn’t under his control, “whether he drinks it intentionally or unintentionally is, in his opinion, subject to his control. This, he believes, is open, and up to him” (Mele 1992, 180). The general lesson we can draw is this: intentions are incompatible *not* with the belief that the action in question is inevitable, but with the belief that it’s inevitable that the action will be performed unintentionally. Although I think this is correct, the problem with Mele’s answer is that, in appealing again to the highly contrived case of Ted, it may fail to convince the objector that it’s possible for more standard agents to intend to do something they regard as inevitable. So let me present a different and more realistic example.[[27]](#endnote-28)

*Laura’s Tourette’s*. Laura suffers from Tourette syndrome. In her case, the urge to curse is particularly acute during boring philosophy talks. Whenever the urge reaches a certain point, Laura starts cursing unintentionally. However, she can monitor very reliably how her urge to curse grows in intensity as the talk progresses. Today she is in the midst of a boring talk and she is unable to leave the room (it’s too crowded for her to reach the door). At some point she realizes that she will start cursing unintentionally within one minute. Before that happens, and for whatever reason (perhaps in order to gain a sense of control or to make the talk stop at an appropriate juncture), she decides to start cursing immediately. She then intentionally shouts “F\*\*\*!”.

This case shows that one can intend to do something even when one believes that one will end up doing it regardless of whether one intends it or not, provided that one also believes that it’s up to one whether the action occurs intentionally or unintentionally.[[28]](#endnote-29) It also shows that one’s intentions can be the cause of behavior one believes will inevitably come about. Finally, if what I have argued above is correct, for Laura to be able to intend to curse intentionally her action must be explicable in terms of reasons rather than in terms of her sub-agential dispositions. This condition is clearly fulfilled in the example: Laura can explain her cursing in terms of reasons (gain a sense of control, make the talk stop) rather than in terms of her syndrome.

The second and related worry is what we might call *the problem of* *domesticating alien dispositions*. The worry is that trying to form an intention to φ when one knows that one’s φ-ing will inevitably come about due to one’s sub-agential (and sub-rational) dispositions involves a high degree of self-deception, because in such cases trying to form the intention to φ reflects a transparent attempt to “domesticate” what would otherwise be a sub-rational or “alien” disposition into an instance of rational agency.[[29]](#endnote-30) One might wonder whether an agent can accomplish this feat via clear-eyed deliberation.

I’ll say two things in response. The first is that, in Ted’s case, forming the intention to drink the toxin isn’t an attempt to transform his sub-rational disposition to drink vials of toxin into an exercise of rational agency. The intention itself is eminently rational because it will get Ted the money; moreover, given the idiosyncratic features of his agency, this can be easily accomplished via deliberation without any degree of self-deception. As I said above, the sub-rational disposition plays the role of an enabler of his intention, and Ted can be aware of this fact without his awareness undermining his ability to rationally form and sustain the intention.[[30]](#endnote-31) Moreover, since this disposition isn’t itself his motivating reason for drinking, he needn’t be confusing a sub-agential disposition with a motivating reason.

The second thing to note is that it does seem possible, and sometimes even rational, to form certain intentions in order to try to domesticate one’s alien dispositions. The case of Laura is again a good example. In one version of the case, what she does is precisely to form the intention to curse in order to gain a sense of control over her cursing, thus transforming her sub-rational disposition into an (admittedly pyrrhic) exercise of rational agency. This move needn’t involve self-deception on Laura’s part; it would only do so if she were actually unable to intentionally perform the action that her sub-rational disposition would in any case produce (which isn’t so in Laura’s case). Moreover, in the alternative version of the case, Laura’s cursing intentionally is more robustly rational, because she curses in order to embarrass the speaker, thus forcing her to stop, and thereby making everybody in the audience happy that the talk is over.[[31]](#endnote-32) And she can rationally accomplish these ends without losing sight of the fact that in all likelihood she wouldn’t have taken this course of action had she lacked the sub-agential disposition to curse. Therefore, an ordinary agent (at least more ordinary than Ted) can form through clear-eyed deliberation a certain intention partly as a means of transforming the operation of a sub-agential and sub-rational disposition into an instance of rational agency.

The third and final worry targets my claim that Ted’s sub-rational disposition plays the role of an enabler of his intention to drink the toxin. The worry is that there seems to be something deeply problematic about an intention the formation and persistence of which is mediated by a disposition that bypasses the agent’s rationality. Call this *the problem of* *sub-rational enablers of intention*. This can be cast either as a worry about the *possibility* of forming an intention in these circumstances or as a worry about the *rationality* of the intention thus formed. I hope my answers to the previous two problems have sufficiently shown that the metaphysical worry about possibility can be satisfactorily addressed. Now I will say something more to address the rationality worry, although Laura’s case has already touched on it.

The key observation is that it’s neither distinctive of Ted nor a problematic feature in general to have one’s intentions regularly enabled and sustained by dispositions that bypass one’s rational agency. Imagine, for instance, a strong-willed agent who always follows through with her intentions when she has conclusive reasons to do so. So far, this looks like a model of rationality. Now suppose we learn that her ability to do this is mediated by (or grounded on) a disposition that bypasses her rational agency: she suffers terrible pangs of conscience whenever she contemplates the prospect of diverting from a course of action she has good reasons to follow through with. Indulging in a bit of psychoanalysis, let’s suppose further that this disposition originated in her because she was conditioned as a child to honor her avowed commitments on pain of being severely shamed by her parents. She then internalized this ruthless regime and became habituated to always implement her intentions unless she had sufficient reasons not to (just as Ted), since the prospect of dropping an intention for an insufficient reason made her experience almost unbearable guilt.

In this case, the agent’s ability to follow through with her intentions is mediated by a disposition that bypasses her rational agency, but this isn’t rationally problematic as long as her intentions are *also* formed and sustained for good reasons, just as Ted’s intention to drink was formed for the excellent reason of winning the prize and was sustained by the good reason that dropping the intention would gain him nothing (since he would drink anyway). Sub-rational dispositions are rationally problematic only when they cause the agent to adopt intentions that are less well supported by her reasons.[[32]](#endnote-33) Imagine a person who always walks to her destinations instead of taking the subway because she suffers from claustrophobia. Given her disposition to panic in confined spaces, she can be counted to stick to her intention to walk regardless of the distance or the weather, but in her case this is rationally problematic given that sometimes her intentions will run counter to the balance of reasons (arriving on time, not getting wet).

Someone might still object that there is a worry about self-knowledge lurking here, for it might be thought that if the strong-willed agent gained vivid awareness of her sub-rational disposition she would lose the ability to rationally form and sustain her intentions. Thus, the seeming rationality of her intentions hinges on either self-ignorance or self-deception, and the same can be said about the cases of Ted and Laura. To counter this objection, and to reinforce the contention I made above that vivid awareness of the mechanisms that enable the formation and stability of one’s intentions doesn’t necessarily undermine their rationality, consider one last example:

*Myrna’s blocker*. Myrna, an academic, found herself constantly distracted by her e-mail, for she had the habit of checking it every five minutes. This noxious routine prevented her from forming reliable intentions about meeting deadlines. Fortunately, one day she discovered a clever online application that allows the user to block access to her e-mail for a certain number of hours and that, once installed and configured, can’t be modified or removed from the device. When Myrna got this application, her procrastination problems disappeared and she was then able to form and sustain effective intentions about meeting deadlines.

Three important lessons can be learned from this example: i) an agent’s ability to rationally form and rationally sustain intentions can be mediated by a mechanism that bypasses her agency; ii) this mechanism needn’t even be part of the agent; and iii) the agent can be fully aware of the mechanism’s operation without this awareness undermining either the possibility or the rationality of the intentions enabled by it. Since this case is structurally identical to the cases of Ted, Laura, and the strong-willed agent with respect to the presence of a sub- (or non-) agential mechanism that enables the formation and stability of certain intentions, these lessons apply to them as well.

**5. Conclusion**

I hope the foregoing discussion has shown that the picture of agency that supports my treatment of Ted’s case, together with the lessons I have drawn from it, not only withstand close scrutiny but, more importantly, are generalizable to more standard agents. This last point is crucial because, as I mentioned at the outset, the expected payoff of delving into the toxin puzzle is precisely to learn something about the nature of intention itself.

One of the main lessons of the paper is that, although there must certainly be a close link between intention and action, the nature of the link isn’t as straightforward as one would initially have thought. One can take the toxin puzzle precisely as a challenge to discover the *essential* nature of this link. We can thus grant Hieronymi and Shah that *in general* agents form intentions by focusing exclusively on the reasons that support available courses of action; and we can grant Kavka that even when this isn’t the case and one takes into account autonomous benefits in deliberation, *in general* one’s actions will still be supported by normative reasons. We have seen, however, that this isn’t necessarily so, so the RKR and NRP accounts can’t explain the puzzle. In order to solve it, we need to find a link between intention and action that can’t be broken without thereby making it impossible to form an intention.

Mele claims that the basic link doesn’t go through the agent’s reasons for action, but through her beliefs about what she will try to do. While I haven’t argued directly against the necessity of this link,[[33]](#endnote-34) I do have argued at length that the most basic link between intention and action goes through the agent’s motivations. If this is correct, we have an explanation of why, whenever an agent believes that this basic link has been severed, she can’t form the corresponding intention—just as MERP indicates. So while it’s true that Ted, unlike the rest of us, doesn’t believe that he won’t even try to drink the toxin, this isn’t the ultimate explanation of why he can, but we don’t, form the required intention. Rather, the ultimate explanation is that he, unlike the rest of us, doesn’t believe that he won’t have an appropriate motivational explanation for drinking the toxin. I thus conclude that MERP, being a more fundamental metaphysical constraint on intentions than Mele’s BATP, adequately captures the essential link between intention and action and thus ultimately explains the inability of ordinary agents to win the prize in the toxin scenario.[[34]](#endnote-35)

**Acknowledgments**

I would like to express my deep appreciation for the outstanding comments, suggestions, and objections made by the referees for this journal and which greatly improved the paper through several revisions.

**Notes**

1. . A metaphysical constraint on intentions is a constraint that identifies conditions the violation of which makes it impossible for an attitude to count as an intention. In turn, a rational constraint is one which identifies conditions the violation of which makes it impossible for an agent to *rationally* form an intention, although it may remain possible for the agent to acquire the intention through non-rational processes (by hypnosis, say). In this paper my main interest is on metaphysical constraints on intention, although toward the end I will address some issues about rationality as well. [↑](#endnote-ref-2)
2. . Gauthier (1998a) and McClennen (1990, 227-31) are the most prominent dissenters. [↑](#endnote-ref-3)
3. . Part of what this implies is that the agent isn’t a masochist who enjoys suffering for its own sake, for if we admit this possibility, then the fact that drinking the toxin will make him suffer will be a reason for him to drink. It is ruled out as well the appeal to the independent value that steadfastness has for some people (Kavka 1983, 35). See Sobel (1994, 249) for an argument that this value can make it rational to intend to do something that is, previously to the formation of the intention, irrational. [↑](#endnote-ref-4)
4. . Kavka (1983, 34) also stipulates that the agent in the toxin scenario cannot form the intention by using “gimmicks” such as promising someone that she will drink the toxin, hiring a hitman to kill her if she doesn’t drink it, hiring a hypnotist to implant the intention, etc. [↑](#endnote-ref-5)
5. . Other philosophers who endorse the idea that reasons for intention must be reasons for action include Anscombe (1963, 90), Davidson (1985, 213-4), and Goetz (1998). [↑](#endnote-ref-6)
6. . These brief remarks of course don’t suffice to conclusively refute the RKR account. The important point for present purposes is just that, given the seeming feasibility of intending in response to autonomous benefits, a correct explanation of the toxin puzzle must be capable of accommodating this apparent fact. [↑](#endnote-ref-7)
7. . A related suggestion, due to Michael Bratman (1999), is that considering autonomous benefits in deliberation is acceptable as long as the agent expects that executing the intention will be rational. Bratman calls this constraint on intention formation via deliberation “the Linking Principle” and claims that it—coupled with a “no-regret condition” to the effect that the agent wouldn’t later regret to have stuck with her prior intention—is the key for solving the toxin puzzle (see also Farrell [1989, 288]). Although I think that Bratman’s Linking Principle can be falsified, I lack the space here for making that argument. For present purposes, it suffices to say that Mele’s case of Ted discussed below casts doubt on it, since Ted rationally forms an intention the implementation of which isn’t straightforwardly rational (Mele [1992, 188-9] discusses whether it’s rational for Ted to implement his rationally-formed intention and concludes that it might not be). See also Clarke (2008, 210) for another potential counterexample to the Linking Principle (he calls it “the coherence principle”). In note 34 below I say some more about Bratman’s no-regret condition. [↑](#endnote-ref-8)
8. . See for example Gauthier (1998a, 48) and Goetz (1998, 205). However, Gauthier thinks that, if you form the intention to drink the toxin, then you will actually have a reason to drink it, namely that doing so is part of the best course of action *available* to you (he assumes that forming the intention to drink and then not drinking isn’t a feasible course of action). For criticism of Gauthier’s test for assessing the rationality of plans, see Bratman (2013). [↑](#endnote-ref-9)
9. . Ted’s case doesn’t violate Kavka’s constraint mentioned in note 4 above, which rules out the use of gimmicks (like promises or assassins) that would give agents normative reasons for actually drinking the toxin. The role of Ted’s peculiar dispositions in the example is to ensure the stability of his intention to drink, not to provide him with normative reasons for drinking. In other words, these dispositions are *enablers* of his intention without being normative reasons themselves. See note 14 and subsection 4.3 below for more on this. [↑](#endnote-ref-10)
10. . One may object that NRP may be false for Ted but not for ordinary agents like us, and so it may still be the correct explanation of why *we* cannot form the intention to drink the toxin. The problem is that even ordinary agents may find themselves in a Ted-like situation in which it’s not only possible but even rational to intend to φ even while having no normative reason to φ. For a case in point, involving the adoption of conditional deterrent intentions, see Mele (1992, 193n.11). More importantly, ordinary phenomena like “desiring the bad” (Stocker 1979) and weak-willed intentions (Wallace 2001) also show that it’s possible to intend to do something for which one lacks (and believes one lacks) normative reasons. I’ll discuss at length the significance of such phenomena for understanding the nature of intention in section 4.2 below. [↑](#endnote-ref-11)
11. . This constraint is negative in the sense that it only requires the *absence* of a belief about oneself not even attempting to φ (see Bratman [1987, 38] for a similar constraint). The metaphysical principle on intentions I present and defend in the next section also imposes a negative constraint on belief, but of a different kind. By contrast, many philosophers have argued that intentions require the *presence* of certain beliefs, e.g., Hampshire and Hart (1958), Grice (1971), Velleman (2007), and Setiya (2007). In 4.2 I explain why is relevant that a belief constraint on intentions is negative rather than positive. See also note 22 below. [↑](#endnote-ref-12)
12. . Beside Mele (1992, 189), other philosophers have appealed to BATP to solve the toxin puzzle. See for instance Clarke (2008, 211) and Heuer (2014, 313). [↑](#endnote-ref-13)
13. . It’s clear that Mele intends BATP to be read as a metaphysical, and not merely as a rational, constraint on intention (see Mele 1992, 189). Heuer (2014, 313) also construes this principle as a metaphysical constraint, while Clarke (2008, 211) opts for interpreting it merely as a rational constraint. [↑](#endnote-ref-14)
14. . This makes clear why Ted lacks a normative reason for drinking the toxin. Normative reasons justify actions by showing that they are desirable or required from the perspective of a particular normative system, which in Ted’s case would be prudence or self- interest. But when the time of drinking comes, actually drinking the toxin won’t be in Ted’s self-interest at all, since doing so is neither a means for getting something he wants (drinking isn’t required for winning the prize) nor something he wants for its own sake (Ted lacks an intrinsic desire to drink toxin). Therefore, that there is no justification, and thus no normative reason, for him to drink the toxin. [↑](#endnote-ref-15)
15. . By contrast, causal explanations couched purely in terms of, say, neurophysiological processes don’t reveal the agent’s motives for acting as she did. [↑](#endnote-ref-16)
16. . Smith (1994, 104) emphasizes that motivating reasons explain actions in teleological terms, that is, “by making [what] they explain intelligible in terms of the pursuit of a goal.” But Ted’s explanation of his drinking the toxin isn’t teleological in this sense, so the natural question is whether this explanation amounts to a motivating reason. I address this worry at the end of 4.2 below. [↑](#endnote-ref-17)
17. . This doesn’t contravene what Mele (1992, 173) calls “Condition C” on an agent’s winning the prize, i.e., that the agent must be convinced at the time of forming the intention that he has (and will have) no reason to drink the toxin. It is clear that the sense of “reason” at stake here is normative reason, not motivating reason. [↑](#endnote-ref-18)
18. . Why? Because, barring special cases like an agent who is intrinsically motivated to do what makes her ill, ordinary agents will lack (and know will lack) any motivation for drinking the toxin intentionally and so will lack an explanation of the appropriate kind for doing so. [↑](#endnote-ref-19)
19. . Thanks to an anonymous referee for pressing the objection to which this paragraph is responding. [↑](#endnote-ref-20)
20. . Thanks to an anonymous referee for raising these worries. [↑](#endnote-ref-21)
21. . I owe the term “normative vigilance”, as well the phrasing of this sentence, to an anonymous referee. [↑](#endnote-ref-22)
22. . I hope this makes clear why my account of intention isn’t “cognitivist” in the sense of accounts of the sort defended by Setiya (2007), Velleman (2007), and others. Unlike cognitivist views, I don’t think that intentions are a certain type of belief or must necessarily be accompanied by certain beliefs. Rather, as I explained above, in my view the belief constraint on intentions—the constraint MERP imposes—is purely negative: it just requires the absence of certain beliefs. In addition, the conditions on intelligibility and explanation I impose on intention formation are much less demanding that the requirement of self-understanding that, in Velleman’s view, is the “constitutive aim” of action. See Velleman (2000, 2006). [↑](#endnote-ref-23)
23. . Here I put aside intentions regarding the formation of what Elster (1979) calls “precommitments”, by which an agent embarks in a course of action that does involve (to varying degrees) relinquishing control over her future φ-ing. The clearest example is that of Ulysses, who orders his sailors to tie him up to the mast so that he can hear the sirens’ song without throwing himself to the water. Notice that this case isn’t a counterexample to MERP, since at t2 Ulysses doesn’t perform an intentional action at all in refraining from jumping to the water. [↑](#endnote-ref-24)
24. . Once again, I thank an anonymous referee for raising this worry. [↑](#endnote-ref-25)
25. . To avoid confusions, it’s important to keep in mind that my claim that intentions can constitute reasons is restricted to *motivating* reasons. Many philosophers think that intentions can’t be reasons, but it’s clear that they are concerned with *normative* reasons (see e.g., Broome 2001). Other philosophers have argued that at least some intentions can in fact constitute normative reasons. See for instance Mele (1992, 181) and Sobel (1994, 249). [↑](#endnote-ref-26)
26. . I thank two anonymous referees for raising these objections. [↑](#endnote-ref-27)
27. . An anonymous referee offered a variant of this case, although with a different purpose in mind. [↑](#endnote-ref-28)
28. . An anonymous referee suggested that the assumption that the same type of action obtains regardless of whether the behavior comes about intentionally or unintentionally is mistaken. That is, one can deny that Ted’s drinking or Laura’s cursing are the same type of actions when they are performed intentionally or unintentionally, despite the indiscernibility of the respective overt behaviors. I take that my argument that Ted and Laura are able to form their respective intentions even in the face of their beliefs about the inevitability of certain overt behaviors works equally well even if the referee is right about this. [↑](#endnote-ref-29)
29. . I take this last sentence almost verbatim from the comments provided by one of the anonymous referees. [↑](#endnote-ref-30)
30. . See the third problem discussed below for further defense of this claim. [↑](#endnote-ref-31)
31. . A famous example of an agent who transforms her alien dispositions (in her case, desires) into exercises of rational (free) agency is Frankfurt’s willing addict (Frankfurt 1988a, 24-5). [↑](#endnote-ref-32)
32. . And they are *metaphysically* problematic (i.e., problematic concerning the possibility of forming certain intentions) when they crowd out other (rational) explanations for the conduct in question. This isn’t the case in the example I offer immediately in the text, since the claustrophobic person does have an explanation in terms of reasons for always walking to her destinations, e.g., that by doing so she avoids suffering disabling fear and public embarrassment. These reasons, however, needn’t make her actions all-things-considered rational. [↑](#endnote-ref-33)
33. . Anscombe (1963, 94) suggests that it’s possible to intend to do something even while fully believing that one won’t do it. See also Holton (2009, 50). It is a further question, which I will leave open here, whether it’s possible to intend to do something even while fully believing that one won’t even *try* to do it. [↑](#endnote-ref-34)
34. . A referee worries that my treatment of the puzzle in this paper may obscure the central issue that originally motivated Kavka to develop the toxin scenario, namely to object to Gauthier’s theory of rational commitment according to which an agent can rationally commit to a course of action that provides some benefits to her even when actually executing the plan won’t be in her best interest and may actually be directly against it (for instance a government issuing a threat of retaliation in case of a nuclear attack). On the referee’s view, the adequate way to capture Kavka’s insight in developing the puzzle is through Bratman’s (1999) no-regret condition: an agent can’t rationally commit to a course of action if she fully expects that at “plan’s end” she will regret following through with it (which is plausibly the case both in the toxin scenario and in the nuclear deterrence example). I’ll make two briefs points in response. First, as I said in note 1 above, my approach to the toxin puzzle focuses on discovering metaphysical constraints on intention, whereas Bratman’s treatment focuses on rational constraints. These shouldn’t be taken to be competing approaches; rather, it can be argued that they complement each other. Just as a hint of how this argument might go: MERP isn’t opposed to Bratman’s no-regret condition; on the contrary, it seems that the former partially grounds the latter: if one believes one will lack a motivating reason for φ-ing at a certain time, then one can plausibly expect regretting φ-ing afterwards (or course, MERP isn’t the whole story about regret, since one can certainly regret performing an action for which one *does* have motivating reasons, such as in cases of giving in to temptation). Second, it’s true that my treatment of the puzzle focuses on the synchronic aspect of it, i.e., on whether the agent can form the required intention at the time of deliberation, whereas Bratman’s treatment focuses on its diachronic aspect, i.e., on the stability (or lack of it) of the intention to drink in the interval between intention formation and execution. Again, however, these two approaches needn’t be seen as incompatible but rather as complementary.

    **Notes on contributor**

    *Fernando Rudy-Hiller* is a researcher at the Institute of Philosophical Research (UNAM) in Mexico City. He works on moral responsibility and practical rationality and has published several papers on these topics.

    **References**

    Anscombe, E. 1963. *Intention*. Cambridge, MA: Harvard University Press.

    Bratman, M. 1987. *Intention, Plans, and Practical Reason.* Cambridge, MA: Harvard

    University Press.

    Bratman, M. 1999. “Toxin, Temptation, and the Stability of Intention,” Chap. 4 in *Faces of*

    *Intention*. Cambridge: Cambridge University Press.

    Bratman, M. 2013. “The Interplay of Intention and Reason.” *Ethics* 123 (4): 657-672.

    Broome, J. 2001. “Are Intentions Reasons? And How Should We Cope with

    Incommensurable Values?” In *Practical Rationality and Preference*, edited by C. Morris and A. Ripstein, 98–120. Cambridge: Cambridge University Press.

    Clarke, R. 2008. “Autonomous Reasons for Intending.” *Australasian Journal of*

    *Philosophy* 86 (2): 191-212.

    Davidson, D. 1980. “Actions, Reasons, and Causes.” Chap. 1 in *Essays on Actions and Events.* Oxford: Oxford University Press.

    Davidson, D. 1985. “Replies to Essays I-IX.” In *Essays on Davidson: Actions and Events*, edited by B. Vermazen and M.B. Hintikka, 195–229. Oxford: Clarendon Press.

    Elster, J. 1979. *Ulysses and the Sirens*. Cambridge: Cambridge University Press.

    Farrell, D. 1989. “Intention, Reason, and Action.” *American Philosophical Quarterly* 26

    (4): 283-295.

    Frankfurt, H. 1988a. “Freedom of the Will and the Concept of a Person.” Chap. 2 in *The Importance of What We Care About*, Cambridge: Cambridge University Press.

    Frankfurt, H. 1988b. “Identification and Wholeheartedness.” Chap. 12 in *The Importance of What We Care About*. Cambridge: Cambridge University Press.

    Gauthier, D. 1998a. “Rethinking the Toxin Puzzle.” In *Rational Commitment and Social Justice*, edited by J. Coleman and C. Morris, 47-58. Cambridge: Cambridge University Press.

    Gauthier, D. 1998b. “Intention and Deliberation.” In *Modeling Rationality, Morality, and*

    *Evolution*, edited by P. Danielson, 41-54. Oxford: Oxford University Press.

    Goetz, S. 1998. “Reasons for Forming an Intention: A Reply to Pink.” *Mind* 107: 205-

    213.

    Grice, P. 1971. “Intention and Uncertainty.” *Proceedings of the British Academy* 57: 263-

    79.

    Hampshire, S. and H. Hart. 1958. “Decision, Intention, and Certainty.” *Mind* 67: 1-12.

    Heuer, U. 2014. “Intentions and the Reason for Which We Act.” *Proceedings of the*

    *Aristotelian Society* 114 (3): 291-315.

    Hieronymi, P. 2005. “The Wrong Kind of Reason.” *Journal of Philosophy* 102: 437-57.

    Hieronymi, P. 2006. “Controlling Attitudes.” *Pacific Philosophical Quarterly* 87: 45-74.

    Holton, R. 2009. *Willing, Wanting, Waiting*. Oxford: Oxford University Press.

    Kavka, G. 1983. “The Toxin Puzzle,” *Analysis* 43: 33–6.

    McClennen, E. 1990. *Rationality and Dynamic Choice*. Cambridge: Cambridge University

    Press.

    Mele, A. 1992. “Intentions, Reasons, and Beliefs: Morals of the Toxin Puzzle.”

    *Philosophical Studies* 68: 171-94.

    Morauta, J. 2010. “In Defense of State-Based Reasons to Intend.” *Pacific Philosophical*

    *Quarterly* 91: 208–228.

    Pink, T. 1991. “Purposive Intending.” *Mind* 100: 343-359.

    Schroeder, M. 2012. “The Ubiquity of State-Given Reasons.” *Ethics* 122 (3): 457-488.

    Setiya, K. 2007. “Cognitivism about Instrumental Reason.” *Ethics* 117 (4): 649-67.

    Shah, N. 2008. “How Actions Governs Intention.” *Philosophers’ Imprint* 8 (5): 1-19.

    Smith, M. 1994. *The Moral Problem*. Malden, MA: Blackwell.

    Sobel, H. 1994. “Useful Intentions.” Chap. 12 in *Taking Chances*: *Essays on Rational Choice.* Cambridge: Cambridge University Press.

    Stocker, M. 1979. “Desiring the Bad: An Essay in Moral Psychology.” *Journal of*

    *Philosophy* 76 (12): 738-753.

    Velleman, D. 2000. “Introduction.” Chap. 1 in *The Possibility of Practical Reason*, Oxford:

    Oxford University Press.

    Velleman, D. 2006. “From Self-Psychology to Moral Philosophy.” Chap. 10 in *Self to Self*.

    Cambridge: Cambridge University Press.

    Velleman, D. 2007. “What Good Is a Will?” In *Action in Context*, edited by A. Leist and H. Baumann, 193-215. Berlin: de Gruyter.

    Wallace, J. 2001. “Normativity, Commitment, and Instrumental Reason.” *Philosophers’*

    *Imprint* 1 (3): 1-26. [↑](#endnote-ref-35)