# Evil and God's Toxin Puzzle

John Pittard
Yale University

*Note:* This is a *pre-refereed* version of the article. The final version, which contains non-trivial changes, is forthcoming in *Noûs*. Please cite the final version.

**Abstract:** I show that Kavka's toxin puzzle raises a problem for the "Responsibility Theodicy," which holds that the reason God typically does not intervene to stop the evil effects of our actions is that such intervention would undermine the possibility of our being responsible for overcoming and averting evil. This prominent theodicy seems to require that God be able to do what the agent in Kavka's toxin story cannot do: stick by a plan to do some action at a future time even though when that time comes, there will be no good reason for performing that action (and very good reason *not* to). I assess various approaches to solving this problem. Along the way, I develop an iterated version of Kavka's toxin case and argue that the case is not adequately handled by standard causal decision theory.

Kavka's "toxin puzzle" (Kavka 1983) presents us with a delightfully perplexing problem in the theory of practical rationality. In Kavka's story, an agent (let's call her Verity) is approached by an "eccentric billionaire" who offers to pay Verity one million dollars tomorrow morning if, at midnight tonight, she intends to drink a vial of a certain toxin tomorrow afternoon. Drinking the toxin would leave Verity feeling ill for a while, but Verity would gladly undergo a short period of nausea if it meant that she would receive such a large sum of money. Unfortunately for Verity, however, if Kavka is right she will not be able to intend to drink the toxin (at least not if she is rational), and thus will not be able to profit from the billionaire's proposition. To see why, note that the benefit attaches to the *intention* to drink and not to the act of drinking itself. If Verity succeeded in intending to drink the toxin at midnight, then tomorrow afternoon Verity would have no reason to actually ingest the toxin since she would have already received the million dollars earlier that morning (and since she would know that receiving the money was not in any way contingent on her actually drinking the toxin). And of course if she failed to form the intention and thus received no money, then she would have no reason to drink the toxin tomorrow afternoon. Since Verity knows all this, she knows that she will have no reason to drink the toxin tomorrow (and will have ample reason *not* to drink it) and thus she cannot rationally form the intention to drink the toxin. Or so Kavka claims. It *is* initially hard to believe that Verity would be unable to profit in this situation, and despite the straightforward line of reasoning offered by Kavka, some philosophers have contested his conclusion.[1]

In this paper, I argue that Kavka's toxin puzzle raises a thorny problem for what is arguably the most promising theistic response to the problem of evil. This response may be called the "Responsibility Theodicy," since it holds that one of God's aims in creating the world is the existence of creatures who are significantly responsible for much of the good in that world, and that this fact can explain the prevalence of genuinely pointless evil. This theodicy faces a significant and as yet unnoticed problem. For reasons that I will explain, the theodicy seems to require that God be able to do what Verity cannot do: stick by a plan to do some action at a future time even though when that time comes, there will be no good reason for performing that action (and very good reason *not* to). This poses a quandary for any advocate of the Responsibility Theodicy who (like myself) is sympathetic to Kavka's claim that Verity cannot rationally drink the toxin (and consequently cannot intend to do so). My aim in this paper is to consider whether the Responsibility Theodicy can be maintained *without* departing from Kavka's diagnosis of the toxin puzzle. I bring mixed news for the theodicist: there are various responses to the toxin puzzle that may be made on behalf of the Responsibility Theodicy, but none of these are philosophically unproblematic.

The paper has five sections. In section 1, I show what is at stake by sketching the Responsibility Theodicy and identifying its explanatory advantages over rival theistic explanations of evil. In section 2, I show why the Responsibility Theodicy implies that God faces a conundrum similar to the one Verity faces in Kavka's toxin story. In section 3, I canvass some possible solutions to the problem that are ultimately unsatisfactory, either because they involve contestable metaphysical or moral claims or because they significantly weaken the explanatory power of the Responsibility Theodicy. In section 4 I develop a "decision theoretic" solution that I judge to be the most promising answer to the problem of God's toxin puzzle. The decision theoretic solution depends on the fact that God's situation is analogous to an iterated toxin case rather than a one shot toxin case like the one discussed by Kavka. Unfortunately, if causal decision theory is correct, the iterated nature of God's toxin puzzle does not help: God still has no reason to "ingest" the toxin. But I argue that causal decision theory mishandles the iterated toxin case, and that the correct decision theory will support the decision theoretic solution. Finally, in section 5 I show that despite the advantages of the decision theoretic solution, the solution does place certain constraints on the evils that the Responsibility Theodicy could be used to explain.

## 1. The Responsibility Theodicy

According to the Responsibility Theodicy, it is a great good for creatures to be significantly responsible for the positive wellbeing experienced by others, and to exercise this responsibility through loving and morally courageous free actions (Swinburne 2004, chs. 10-11). Plausibly, the value of this responsibility is enhanced if creatures are responsible not only for contributing positive goods to the lives of others, but also for averting and overcoming evils (by opposing oppression, befriending the lonely, healing the sick, and so on). Without responsibility for overcoming such evils, the scope and moral seriousness of creaturely responsibility would be significantly curtailed (Swinburne 2004, 224-5).

Given certain plausible assumptions, God's consistently pursuing the good of serious creaturely responsibility makes it a near certainty that there will be a significant amount of evil. To see why, consider what must be the case in order for Lola to have the desired sort of responsibility for Cyril's wellbeing. Lola must be in a situation where she faces a free choice between some good action *G* that positively affects Cyril and some less good or bad action *B* that results in Cyril suffering some significant evil *E*. Could God give Lola the free choice between *G* and *B* and then cause

Lola to freely choose *G*? The Responsibility Theodicy sides with incompatibilists in holding that God could not do this, since Lola's *freely* choosing *G* is not compatible with her being causally determined to make this choice. Even if God merely made it overwhelmingly likely that Lola chooses *G* (for example by giving her an extremely virtuous moral character), this would arguably diminish the extent to which *Lola* can be responsible for the right choice. So if God desires that Lola be *significantly* responsible for Cyril's wellbeing, then arguably God must create a context where Lola's choosing *B* is not made overwhelmingly improbable by factors outside of Lola's control. Could God's decision to give Lola the choice between *G* and *B* be based on whether she would in fact choose *G*? Arguably not, since before Lola makes her choice, it is doubtful whether there is a knowable fact (or even a fact at all) about what she would *freely* choose if given the opportunity.[2] Finally, could God give Lola the choice between *G* and *B* but resolve to intervene if Lola chooses *B* in order to stop this action from resulting in evil *E*? No, not without forfeiting the possible good of Lola's being responsible for averting *E*. Lola's choosing *G* can be credited with averting evil *E* only if a different choice would have resulted in *E*.[3] Thus, the divine intention to intervene to stop the occurrence of *E* would prevent Lola from having morally serious responsibility for Cyril.

If the above line of reasoning is correct, then in order for God to pursue the good of Lola's positively exercising significant responsibility for Cyril, God must bring about a set of circumstances where there is *substantial* risk of *significant* evil. Thus, if significant creaturely responsibility (of a morally serious sort) is highly valuable, valuable enough to pursue even if such pursuit risks significant evil, then we have an explanation for why there is significant evil.[4] If God consistently pursues the good of significant creaturely responsibility, then the probability that there will be a large quantity of evil rapidly approaches 1.

A significant count in favor of the Responsibility Theodicy is that unlike "Defeat Theodicies," the Responsibility Theodicy does *not* affirm that every evil makes possible some greater good that depends on the evil for its realization.[5] If Lola chooses *B* and causes Cyril to suffer, that suffering may be completely pointless, with no upside whatsoever. God may permit *B* (and evil *E* that results) *not* because *E* is necessary for the attainment of some good, but only because the *possibility* of *E* was necessary to secure the possibility of some great good (namely, Lola's being responsible for rescuing Cyril from *E*). The proponent of the Responsibility Theodicy can hold that if God could have known in advance that Lola would choose *B*, then God would not have given her the free choice at all. For Lola's having a free choice to do good or evil might be of little to no *intrinsic* value, being significantly valuable only because it makes possible the *positive use* of free choice to significantly improve Cyril's wellbeing.

Because the Responsibility Theodicy does not require affirming that evils somehow serve the greater good, the Responsibility Theodicy does not threaten to clash with commonsense views about what attitudes we ought to take towards nasty evils. This is arguably a significant count in favor of the Responsibility Theodicy over Defeat Theodicies.[6] Defeat Theodicies do threaten to clash with these commonsense attitudes, since it is hard to see why we should hope that evils do not occur (and grieve when they do) if those evils contribute to some outweighing good that depends on some such evil for its realization.

## 2. God's Toxin Puzzle

If the Responsibility Theodicy is correct, then God faces a divine version of the toxin puzzle. Here's why. As explained in the last section, in order for God to pursue the good of Lola's being responsible for rescuing Cyril from *E*, God must ensure that *were* Lola to choose bad action *B*, God would *not* intervene in order to stop *E*'s resulting from *B*. But the same sort of reasoning that (according to Kavka) makes it irrational for Verity to drink the toxin would seem also to make it irrational for God to refrain from intervening in the event that Lola chooses *B*. For if Lola does choose *B*, then God will no longer have any reason to refrain from intervening to stop *E*. This is because the only reason for nonintervention was to make possible the good of Lola's rescuing Cyril from *E*, and once Lola has chosen *B* this good is no longer a possibility and thus cannot supply any reasons for acting. Moreover, God will have a very good reason in favor of intervening (namely, averting *E*). So if Lola chooses *B*, intervening to stop *E* is the only rational action for God. Since God always acts rationally, we can conclude that were Lola to choose *B*, God would intervene to stop *E* from occurring. Thus, Lola cannot be given significant responsibility for Cyril's wellbeing.

The divine situation does differ in some important respects from Verity's situation in the original toxin puzzle. First, God's "toxin," which is allowing *B* to cause evil *E*, can be "ingested" only if the hoped for good fails to obtain. Second, God's *intending* to ingest the toxin is not enough to secure the possibility of the hoped for good. Rather, it must be the case that God would ingest the toxin if presented with the possibility. An analogous toxin story would go as follows: The billionaire will flip a coin. If the coin lands heads, the billionaire will not give Verity any money but will present her with a toxin that causes nausea for a minute and which she may drink if she so chooses. If the coin lands tails, then the billionaire will not give Verity a toxin but will give her a million dollars if and only if it is the case that *had the coin landed heads*, Verity would have drunk the toxin. The logic rehearsed by Kavka suggests that Verity could not rationally drink the toxin if the coin landed heads, and that as long as Verity is rational she cannot profit from the billionaire's offer.

If for analogous reasons God cannot ensure that God would not intervene to stop *E* should Lola choose badly, then there would be no way for God to pursue the good of Lola's positively exercising significant responsibility for Cyril's wellbeing. In this case, the initially promising Responsibility Theodicy would prove to be incoherent.

## 3. Some problematic solutions

*A. Divine self-binding*

In the original toxin puzzle, Kavka stipulates that Verity can neither restrict the actions available to her when presented with the toxin (for example, by planting a chip in her brain that compels her to drink the toxin) nor change the costs and benefits associated with drinking the toxin (for example, by signing a contract that transfers ownership of her house to her worst enemy if she refuses to drink the toxin).[7]

Since God's situation is not subject to Kavka's stipulations, one might think that God's toxin puzzle is easily resolved. But some of the solutions available to a human agent who is not restricted by Kavka's stipulations might not be applicable to an omnipotent being. Consider first the proposal that God could restrict God's future self so as to make intervention to stop *E* impossible. While we human beings can frequently restrict the actions available to us in some situation (as Odysseus did, binding himself to the mast of his ship before passing through the province of the sirens), it is far from clear that God could act so as to make the divine will inefficacious at some future time.[8] Claiming that God can bind God's future self would involve contested metaphysical commitments that go beyond those already presupposed by the Responsibility Theodicy. A less metaphysically loaded solution would be preferable.

*B. Divine promises*

Maybe God could make intervention immoral by *promising* (to Godself? to a human being? to an angel?) not to intervene. Maybe. But serious doubts may be raised about this solution. If after Lola chooses *B* there is really nothing to be gained from allowing *E* to result (prior to consideration of God's promise), then it seems that no good and rational agent would hold God to a promise to not intervene. Since no good person would hold God to such a promise, breaking this promise would arguably not be morally problematic. Imagine that Verity, unrestricted by any of Kavka's stipulations, made a promise to her sister that she would drink the toxin the next day. After winning the million dollars, we can imagine Verity's sister encouraging her not to drink the toxin. After all, Verity already fairly won the money and drinking the toxin isn't necessary or helpful to anyone. Moreover, it will pain the sister to see Verity drink the toxin. Verity's protest that *she promised* her sister to drink it would not, it seems to me, be adequate to justify drinking the toxin. This gives us reason to doubt the "divine promise" solution to the divine toxin puzzle.

*C. The intrinsic badness of divine intervention*

If divine intervention in the natural order is intrinsically very bad (because nomological regularity is intrinsically very good), and if the intrinsic badness of God's intervening to stop *E* outweighed the badness of *E*, then God would not face a toxin puzzle after all. If Lola chose action *B*, God's intervening to stop *B* from leading to *E* would be worse than allowing *E* to take place. In this case, God would not intervene in the event that Lola chooses *B*. Thus, in choosing *G* Lola would be genuinely responsible for averting *E*.

Is it plausible that for every evil that has occurred, the divine intervention needed to avert that evil would have been intrinsically worse than the evil itself? I suspect that many will join me in thinking that this is not an especially plausible thesis, even if it isn't downright absurd.[9] The proponent of the Responsibility Theodicy has reason to hope for a more satisfying solution to the toxin puzzle problem.

### *D. Epistemic side effects of intervention*

The final solution to be considered in this section argues that if God intervened to stop Lola's bad choice from leading to *E*, this would undermine the epistemic conditions required for meaningful human responsibility in the future.[10] To illustrate the solution, suppose that on a trip to a nearby national park, Lola crosses a bridge spanning a deep gorge and notices that one of the planks in the bridge is weakened by significant rot that, if not replaced, could lead to someone's death. Upon reaching the other side, Lola considers notifying a park ranger about the dangerous bridge. She doesn't see any ranger around, notes that she's hungry and eager for dinner, and decides not to bother. Now suppose that the following day, young Cyril steps on the plank but is prevented from falling to his death by an imperceptible act of divine intervention that prevents the plank from collapsing. The following day a park ranger sees the problem and it is fixed before anyone gets hurt. According to the present solution, God's intervening in the envisioned way would likely result in a worse outcome than letting the boy fall to his death. The reason is that divine intervention is likely to have epistemic side effects that would in some way diminish the degree of meaningful creaturely responsibility in the future.

What epistemic side effects? Well, suppose that at some future time Lola finds herself in a situation where doing the conscientious thing really *would* help someone (since in this case God would not intervene to prevent the bad effects of Lola's negligence). Despite the fact that Lola now finds herself in a situation where she has significant responsibility, God's past act of saving Cyril might compromise Lola's ability to recognize the responsibility she currently has. Lola might remember that evening years ago when she neglected to tell anyone about the rotten plank and nothing bad came of it, and this might mislead her into thinking that acting in a conscientious manner is unlikely to make any difference. And of course others besides Lola might be misled into underestimating their responsibility as a result of God's saving Cyril. Had Cyril fallen to his death, the reporting of this event would no doubt have buttressed many people's appreciation of life's fragility and the importance of their making conscientious and caring choices. So divine intervention to stop the evil effects of bad choices may result in an evidential situation that leads people to a lower estimation of their responsibility. And this in turn may make people less likely to choose loving and courageous actions and may compromise the value that could otherwise be gained from putting creatures in a position of genuine responsibility.

There are various problems with this "epistemic side effects" solution to the divine toxin puzzle. First, the proposal arguably *presupposes* a solution rather than supplying one. The puzzle is that it seems that God cannot give Lola real responsibility for Cyril, since (i) for Lola to be responsible God must not intervene should Lola make the bad choice, and (ii) if Lola *does* make the bad choice, intervening is apparently God's only rational option. The proposal under consideration points to putative costs of intervention that may make it rational for God to refrain from intervening: intervention has epistemic side effects that compromise the value of *future* situations where Lola or other agents are in fact responsible for the wellbeing of another person. But note that this reply simply presupposes that there can be a future situation where some agent is genuinely responsible for the wellbeing of another person. This is not a legitimate supposition in this context, since the toxin puzzle would equally challenge the possibility of genuine responsibility in this future situation. If we took it for granted that Lola and other agents could be genuinely responsible in the future, then we could perhaps use this fact to explain why in the present case God should not intervene to save Cyril. But we cannot take it for granted that there could be creaturely responsibility in the future when attempting to explain why creaturely responsibility is not inevitably undermined on account of God's toxin puzzle.

Second, the proposed solution represents a significant step in the direction of Defeat Theodicies, and as a result significantly undercuts the advantage that the Responsibility Theodicy may be thought to have over such theodicies. The epistemic side effects solution says that if God does not intervene to rescue Cyril, it will be because the epistemic effects of such intervention impose a cost that is worse than what would be gained by rescuing Cyril. This means that *given that Lola made the bad decision*, Cyril's death is a welcome development on account of its epistemic effects.[11] This is an unpalatable implication that is not easily squared with commonsense attitudes towards evils. Confronted with the tragedy of Cyril's death, it seems entirely appropriate to wish that Cyril had stepped over the weak plank rather than directly on it. But according to the present solution, this would be to wish for an overall *worse* outcome. For Cyril's stepping over the rotten plank would presumably have the same epistemic effects as God's imperceptibly strengthening the plank, and Cyril's death is evidence that God determined that Cyril's demise was not as bad as the epistemic effects of his crossing the bridge alive. So if we accept this solution to God's toxin puzzle, it is arguably inappropriate to regret developments that allowed bad decisions to have evil consequences (even if it is appropriate to regret those bad decisions). Such troubling implications are exactly what the Responsibility Theodicist seeks to avoid in rejecting Defeat Theodicies.

Finally, even if the previous two problems can be successfully addressed, it remains the case that the plausibility of the Responsibility Theodicy is significantly reduced if the divine toxin puzzle can be addressed only by appealing to the epistemic side effects of divine intervention. If this solution to the puzzle is all we have, then God could grant responsibility only when the epistemic distortion that would result from divine intervention is worse than the evil that would otherwise result. Since it is very hard to believe that our responsibility is limited in this way (even if we cannot prove otherwise), the Responsibility Theodicist should seek a solution that could explain a greater level of responsibility. I turn to the development of such a solution in the next section.

## 4. The decision-theoretic solution

Like the epistemic side effects solution just considered, the decision-theoretic solution relies on the fact that the situation involving Lola and Cyril is one of many similarly structured situations. We might say that God will confront the decision to ingest the "toxin" of non-intervention not just once, but on many occasions. To see why this is so significant, consider again the revised (and more analogous) toxin story described in section 2, where if a coin lands heads Verity is presented with a toxin and if it lands tails she is given a million dollars if and only if she would have drunk the toxin had the coin landed heads. If Verity knows that she will be given this offer only once, then Kavka's reasoning supports the conclusion that Verity cannot make the relevant conditional true and thus cannot profit from the situation. But suppose that Verity knows that the billionaire will present Verity with this offer some large but indefinite number of times. In this case, does Verity have a reason that rationalizes drinking the toxin when the coin lands heads?

If evidential decision theory is correct, then Verity clearly *does* have a reason that rationalizes drinking the toxin when the coin lands heads. Suppose the first toss comes up heads. Since the situation Verity now faces is relevantly like the situation that she would face in future tosses that lands heads, Verity knows that the choice she makes now after the first toss is the same as the choice she *would* make in future tosses that lands heads. (This assumes that only one choice is rational, and that as an impeccable rational agent Verity knows that she will always make the uniquely rational choice.) So Verity can reason as follows: "If I drink the toxin now, then in subsequent tosses where the coin lands tails, it will be true that I *would* have drunk the toxin had the coin landed heads. So if I drink the toxin now, I will be paid $1 million in future tosses where the coin lands tails. If I do *not* drink the toxin now, then in subsequent tosses where the coin lands tails, it will *not* be true that I would have drunk the toxin had the coin landed heads. So if I do not drink the toxin now, I will be paid nothing in future tosses that land tails. Since I will be paid in future tosses if and only if I drink the toxin now, I have reason to drink the toxin."

Unfortunately, this reasoning is fallacious from the perspective of causal decision theory. While Verity's decision to drink the toxin may be conclusive *evidence* that the relevant subjunctive conditional will be true in future tosses, this does not mean that drinking the toxin *causes* the conditional to be true in future tosses. And according to the causal decision theorist, Verity's desire that this conditional be true can give her a reason to drink the toxin right now only if the truth value of this conditional *causally* depends on her drinking the toxin in this particular case. A mere evidential connection between the conditional and her present choice would not supply a reason for acting.

Still, for the evidential decision theorist, a rationale for drinking the toxin is available in the multiple coin toss case that is not available in the single toss case. And since God's situation is analogous to the multiple coin toss case, the evidential decision theorist can endorse a decision-theoretic solution to the divine toxin puzzle. God's reason for not intervening in the event that Lola chooses *B* is that if God refrains from intervening in *this* situation, then in similar future situations where Lola or someone else makes a *good* choice, they will be responsible for averting the evil that would have resulted from the bad choice since God would *not* have intervened to stop this evil consequence. But the justification for this unwieldy "if…then…" claim is purely evidential: God's choice to not intervene in the present case may be conclusive evidence that God would not intervene in other similar cases, but we have not yet provided any reason for thinking that God's choice in the present case *causally* contributes to its being the case that God would not intervene in similar situations.

Returning to Verity's iterated toxin case, do we have any reason to think that her drinking the toxin when the first toss lands heads could *cause* it to be the case that she would also drink it in future tosses? I tentatively suggest that we do.[12] My argument for the possibility of a causal relationship between Verity's drinking the toxin and the truth of the relevant subjunctive conditional (in future tosses) has two steps. First, I suggest that anytime a fully rational agent decides (on the basis of explicit reflection) to follow some action plan in circumstance *C*, the agent endorses following an equivalent action plan (or policy) in all circumstances that are (by her lights) relevantly similar to *C*.[13] In this context, another circumstance qualifies as being relevantly similar to *C* if it shares with *C* features that provide sufficient reason for following the action plan in question. Since a rational agent treats like cases alike, or at least is *willing* to treat like cases alike, a reflective decision to act reveals a rational commitment to the appropriateness of acting equivalently in a wider set of situations that are relevantly like the one at hand. Second, this endorsement may have effects even after the agent has carried out the action plan in circumstance *C*; in particular, assuming nothing has happened to change the agent's rational perspective or dispositions, this endorsement may make it the case that the agent would follow the same action plan in future circumstances that are relevantly similar to *C*. So in settling on a course of action, the agent may determine both what she will do in the present circumstance *and* what she would do in other similar circumstances (assuming there are no intervening changes in her rational perspective). If this is right, then Verity's decision (after the first coin lands heads) to follow the simple action plan "drink the toxin" constitutes an endorsement of this action plan in situations taken to be relevantly similar, and this endorsement may *cause* it to be the case that she would drink the toxin in some such situations in the future. Since by her lights any future toss where the coin lands heads is relevantly like this particular occasion, her following the action plan "drink the toxin" on this occasion could make the relevant subjunctive conditional true for future tosses as well. Verity therefore has a good reason to drink the toxin, since making the relevant conditional true could enable her to win money in future tosses. And since Verity has an overwhelmingly good reason to drink the toxin (and since she always acts rationally), it is true even before she drinks the toxin for the first time that she would drink it when a coin lands heads.

The same reasoning may be transposed to God's situation. If Lola chooses *B* and God follows the action plan "refrain from intervening," God thereby endorses following this action plan in future situations that God judges to be relevantly similar. This endorsement would arguably *make* it the case that God would refrain from intervening in relevantly similar situations. God therefore has a forward-looking reason to refrain from intervening to stop *E* when Lola chooses *B*. By not intervening in this case, God causes it to be the case that God would not intervene in similar situations in the future, thereby securing the possibility that in these future situations creatures can be responsible for averting great evils.[14]

Unfortunately, from the perspective of standard causal decision theory, the rationality of accepting the (literal or figurative) toxin is not established merely by showing that doing so *causes* it to be the case that in future situations the toxin would be accepted. What really matters from the perspective of causal decision theory is not what an action *causes* to be the case, but rather what *difference* the action makes, where the difference an action makes is characterized in terms of subjunctive (or "counterfactual") conditionals that specify what the outcomes would be were the agent to act in a certain way.[15] Of course the difference that my action makes is closely related to what that action causes. But these may come apart when some state of affairs that would be caused by my doing action *A* is causally overdetermined or would be caused by something else were I to refrain from doing action *A*. For example, the fact that my garden needs water before tomorrow's heat wave may not make it rational for me to water my garden today if I know that there is going to be a large rainstorm tonight. I can cause the desired state (the garden's receiving water before tomorrow), but doing so will make no difference as to whether the state comes about. Quite reasonably, causal decision theory says that when I know it will rain tonight, the value that attaches to the garden receiving water before tomorrow has no bearing on the question of whether I should water the garden today.

Turning now to Verity's decision whether to drink the toxin when the first coin lands heads, does her drinking the toxin right now make a difference as to whether she *would* drink the toxin in some future toss? It seems not. To see why, consider Verity's decision situation after a coin lands heads for a second time and she is presented with a toxin. If Verity is rational, then her decision whether to drink the toxin will supervene on facts about costs and benefits associated with drinking or not drinking. And whether she drank the toxin the *first* time the coin landed heads, and what actions she endorsed then, does not in any way affect the costs and benefits associated with drinking or not drinking on *this* occasion. Thus, whether she drank the toxin the first time, and what she endorsed on that occasion, will make no difference with respect to whether she will drink the toxin this time around. Even if her action in the first heads toss caused her action in the second heads toss, her decision in the first heads toss would not make a difference as to what happens in the second toss. What happens in the second toss will inevitably align with cost and benefit facts that Verity cannot influence at the time when she first drinks the toxin.

All this would seem to imply that Verity has no forward-looking reason to drink the toxin after the first toss lands heads. As the example of the garden and the coming rainstorm apparently shows, the goodness of some state of affairs does not give me a reason to contribute to bringing about that state of affairs if I know that this state of affairs would obtain whether or not I contribute to bringing it about. And if Verity has no forward-looking reason to drink the toxin, then drinking the toxin is surely irrational (since the immediate effects of drinking the toxin are purely negative). We thus have a powerful line of reasoning, reasoning that accords with causal decision theory, to the conclusion that Verity will not drink the toxin and will not profit in the iterated toxin case. And the same reasoning of course implies that God has no forward-looking reason to refrain from intervening to stop evils, and thus that the decision-theoretic solution fails.

A critical premise in the above argument is the following:

> **Difference-Making Principle:** The goodness of some state of affairs does not give an agent a reason to contribute to bringing about that state of affairs if she is certain that the state of affairs would obtain whether or not she contributes to bringing it about.

This premise, which follows from standard causal decision theory, is initially plausible. But I believe the principle is subject to compelling counterexamples. Putative counterexamples to causal decision theory developed by Andy Egan (2007) do, I think, hint at worries for the Difference-Making Principle (though they are not, strictly speaking, counterexamples to the principle).[16] But instead of diving into a discussion of Egan's cases, I will offer a case that is structurally similar to Egan's cases but that is a straightforward counterexample to the Difference-Making Principle. Here's the case:

> **Button Game**: Gavin and Hilda are put in different rooms, each of which contains a large red button. Before each leaves their room, they will choose whether or not to press the button. If at least one of them presses the button in their room, then Gavin and Hilda will each win $1 million. If neither presses the button, then they will win nothing. No communication is allowed before they both leave the room. There is, of course, a catch. Pressing the button will cause a mildly painful (but ultimately harmless) shock. Gavin and Hilda are told all of this information and left to make their choices. They know with certainty that they are both impeccable rational agents who will not fail to choose rationally. They also know that neither of them would like to experience the shock, but that for each of them the value of the money vastly outweighs the disvalue of the shock.

Should Gavin press the button in this case? The answer looks to be perfectly clear: of course he should! But if the Difference-Making Principle is true, then we must reject this intuitive verdict. For suppose that pressing the button is the rational choice. As an impeccable rational agent, Gavin will be certain of this. He will also be certain that Hilda will press the button in her room, since he knows her to be perfectly rational and in a symmetrical situation. But this means that Gavin is certain that he and Hilda would win their money whether or not he pressed his button. Applying the Difference-Making Principle, this implies that the goodness of winning the money does not give Gavin a reason to press the button. And since this is the only reason Gavin could have for pressing the button, and since Gavin has a reason to refrain from pressing the button (avoiding the shock), it follows that it is *not* rational for Gavin to press the button, contrary to our original supposition. So if we accept the Difference-Making Principle, we cannot coherently affirm the intuitive view that Gavin should press the button.[17]

This result strikes me as a very good reason for rejecting the Difference-Making Principle and standard causal decision theory, which has that principle as an implication. (Which alternative theory should be embraced is a question that I will not pursue here. There are, however, decision theories intermediate between the standard evidential and causal approaches that do deliver the intuitively correct verdict in the Button Game.[18]) Moreover, there are good reasons for thinking that the reasons why the Difference-Making Principle fails in the Button Game are also reasons that apply in Verity's iterated toxin case. In the Button Game, the fact that there is a cogent rationale for pushing the button would itself be evidence that another agent will bring about the desired outcome and thus that pushing the button makes no difference. Unlike the garden example, where the desired outcome is guaranteed by the rainstorm for reasons that are independent of whether I have a rational basis for watering the garden, in the Button Game it is the supposed reasonability of pushing the button that would ensure that Hilda would press her button (whatever Gavin may do), making the desired result causally overdetermined and thereby undermining the supposition that pushing the button is reasonable (given the Difference-Making Principle). The same structure applies in Verity's case. The supposed fact that she has an adequate forward-looking reason for drinking the toxin before her (namely, making the desired conditional true in future tosses) would also ensure that future Verity would drink *her* toxin (whatever present Verity may do), which in turn makes the desired conditional overdetermined, thereby undermining the supposition that drinking the toxin is rational (given the Difference-Making Principle). In both cases, the fact that rationality favors a certain option would make the desired outcome overdetermined. If we judge that pushing the button is rational in the Button Game case, despite the fact that Gavin would know that the desired result is overdetermined, then we should reject the argument given above which holds that Verity cannot rationally drink the toxin for sake of making true the relevant conditional since her drinking the toxin does not make a difference to whether or not the conditional is true.

Here, then, is what I think would happen in Verity's toxin case. The first time a coin lands heads and she is presented with the toxin, she drinks the toxin because in doing so she endorses the action plan of drinking the toxin in such situations, and this endorsement (which we might describe as Verity's "settling" on a certain course of action) makes it the case that Verity would drink the toxin in future tosses (should the coin land heads) and therefore ensures that Verity will win one million dollars should the next coin (or coins) land tails. Granted, Verity knows that if she was temporarily irrational and refrained from drinking the toxin, then the desired conditional would still be true in future tosses (as long as her rationality was fully restored and future rational lapses were not possible). So there is a sense in which her presently endorsing the "drink the toxin" plan does not make a difference, since the rationality of her future self suffices to make it true that in future tosses she would drink the toxin. But I've argued that the Difference-Making Principle is false and that it can be rational to act for sake of bringing about some desired end by means of some costly action even if it is known that the desired end would obtain whatever one does.

## 5. The solution's advantages and limitations

The advantages of the decision-theoretic solution over the first three solutions explored in section 3 is clear: it avoids the controversial metaphysical commitments of the self-binding solution, the controversial moral commitments of the divine promises solution, and the controversial axiological commitments of the solution that invokes the intrinsic badness of divine intervention.[19] It is worth taking a bit of time to spell out why the decision-theoretic solution is highly preferable to the (vaguely similar) epistemic side effects solution. Consider again young Cyril running across the bridge over the gorge. Suppose Cyril faces a free choice between sprinting over the bridge and merely jogging. If Cyril sprints, it so happens that he will not step on the rotten plank; if he jogs, he will step on the plank. Suppose further that God would not intervene to save Cyril in the event that he steps on the plank. It seems that proponents of the epistemic side effects solution are committed to saying that from the fact that God would not intervene to save Cyril, we can infer that it would be better if Cyril chooses to jog (resulting in his death) than if he chooses to sprint (thereby preserving his life). For on their view, God's non-intervention is explained by the fact that the epistemic distortion that would result from divine intervention is worse than Cyril's death. And since the same epistemic distortion would result if Cyril sprints and happens to avoid the rotten plank (without divine intervention), it seems to follow that Cyril's jogging and dying would be a better outcome than his sprinting and living. The decision-theoretic solution does not have this nasty implication. While God must refrain from intervening in order to preserve creaturely responsibility in future situations, nothing good would come of Cyril's death itself. God may legitimately hope that Cyril chooses to sprint, and the commonsense judgment that it would have been better if Cyril had luckily avoided the rotten plank is not threatened. Cyril's *death* is not a means towards any greater good and may rightly be recognized as tragic, even if divine intervention was ruled out as too costly.

While the decision-theoretic solution to the puzzle does have significant advantages over the proposals considered in the previous section, the decision-theoretic solution also has its own liabilities (even if we have no problem rejecting causal decision theory). If we rely on the decision-theoretic solution alone, the Responsibility Theodicy has trouble explaining the possibility of an evil that is temporally last or an evil that is much worse than all other potential evils. First, suppose God knows that choice *n* is the last choice involving significant responsibility. Given this knowledge, God has no future-oriented reason to refrain from intervening if the bad choice is made; so God would intervene if the bad choice is made; thus, *n* is *not* a choice involving significant responsibility, which contradicts our original supposition. It would seem, then, that if we rely only on the decision-theoretic solution, we must deny that there is ever a choice that God knows to be the last choice involving significant responsibility. Perhaps this is not a problem. Maybe the theist can simply deny that there is such a last choice. Such a denial is compatible with the view that there will be an end to the era of human responsibility for averting evil. For example, God could guarantee that there are 20 choices that are tied for last, rather than a single last choice. In this case, God's reason for not intervening in the instances where a bad choice is made would be to confer responsibility on the agents who *simultaneously* made a good choice. Alternatively, the theist could grant the possibility of a last significantly responsible choice but simply acknowledge that if there is one, God doesn't know in advance which choice it is (perhaps because this is determined by a free or random process that even God cannot predict in advance).

Another argument raises problems for the possibility that God could know of some choice that it is the costliest (if the bad choice is made). To see why, consider the iterated toxin case, but with the following twist. Normally, when the coin lands heads Verity is presented with a green toxin that causes nausea for one minute. But Verity is told that for exactly one coin toss, if the coin lands heads she will be given a red toxin that causes nausea for two minutes. The series of coin tosses commences, and for the first several tosses, when the coin lands heads Verity drinks the green toxin, and when it lands tails she is given a million dollars. But before the 23rd toss, the billionaire announces that this is the toss where the red toxin will be presented should the coin land heads. Upon hearing this, Verity knows that she will not earn any money should this particular coin land tails. To see why, consider the situation Verity would face if the coin lands heads and she is presented with the red toxin. Normally, Verity has reason to drink the green toxin since endorsing the policy of drinking the toxin allows her to make money in future coin tosses. But once she's already been presented with the red toxin, she knows that endorsing the policy of drinking the green toxin and refusing the red toxin has the same expected payoff in future coin tosses as endorsing the policy of always drinking the toxin. And the nauseating effect of the red toxin gives her a reason to prefer following the former policy. So if the red toxin is presented, Verity, being rational, will follow the "drink green toxins only" policy and refuse the red toxin. This is why she cannot profit from the coin toss where the red toxin would be presented.[20] Analogous reasoning suggests that the Responsibility Theodicy (combined with the decision-theoretic solution to the toxin puzzle) is unable to explain how God could permit some evil that God knows is much worse than any other potential evils. Again, this may not be a fatal problem for someone who prefers the decision-theoretic solution, but it is perhaps an uncomfortable result.

The degree of discomfort may be heightened by a third argument that combines the features focused on in the last two arguments (a potential evil's "lastness" and its unsurpassed badness). If we rely on the decision-theoretic solution alone, the Responsibility Theodicy cannot easily explain why human beings would have the power to bring about the extinction of the human species. Suppose that Quinn takes some action that, absent divine intervention, is guaranteed to lead to the extinction of the human species. According to the decision-theoretic solution, God's reason for *not* intervening to stop Quinn's act from destroying the species is that refraining from intervening ensures that, when God faces *relevantly similar* situations in the future, God would not intervene (thereby allowing for the possibility of morally significant freedom). But this application of the decision-theoretic solution may be incoherent. At the moment that God has refrained from intervening long enough to ensure that Quinn's action will result in the destruction of the human species, there arguably cannot be any future situations where God will face a situation that is relevantly like the decision situation God faced in responding to Quinn's bad choice. Given the imminent destruction of humankind, there may be no remaining human choices that are comparable in significance to the choice already made by Quinn. It is therefore doubtful that God's non-intervention in Quinn's case would serve the good of allowing for human responsibility in comparably significant situations. While there may be less significant free human choices remaining, God's non-intervention in past situations that were also less significant would suffice to make it the case that God would not intervene in these remaining situations. If this is right, then no future good would be served by God's refraining from intervening in order to stop Quinn from destroying humanity. So the Responsibility Theodicy and the decision-theoretic solution do not seem capable of explaining the commonsense view that human beings do have the power to destroy the species.[21] Note that the self-binding and divine promises solutions described in section 3 have no problem in accounting for our potential to bring about our own extinction. God could easily give human beings responsibility for human preservation if God could bind Godself to prevent divine intervention to save the species, or if God could take some action that would make such intervention highly costly or immoral.

While the decision-theoretic solution seems to me to be the most promising approach to the divine toxin puzzle, I've suggested that this solution does limit the explanatory power of the Responsibility Theodicy in certain (bearable but uncomfortable) ways. Of course there may be promising solutions to the divine toxin puzzle that I have not identified, and perhaps others will argue that one of the solutions discussed in section 3 is more promising than I have supposed. Whether or not one accepts my view on which solution is most satisfactory, I at least hope to have shown that the toxin puzzle raises a significant challenge for the Responsibility Theodicy, one that any adequate defense of the theodicy must address. And for those uninterested in the prospects of the Responsibility Theodicy, the iterated toxin case raises puzzling questions that are interesting in their own right. I've argued that a rational agent could profit in such a situation (even though this conflicts with standard causal decision theory). No doubt this claim merits further scrutiny.

## References

Edgington, D. (2011). Conditionals, Causation, and Decision. *Analytic Philosophy*, 52/2: 75–87.

Egan, A. (2007). Some Counterexamples to Causal Decision Theory. *The Philosophical Review*, 116/1: 93–114.

Gauthier, D. (1998). Rethinking the Toxin Puzzle. Coleman J. L. & Morris C. W. (eds) *Rational Commitment and Social Justice: Essays for Gregory Kavka*, pp. 47–58. Cambridge University Press: Cambridge, UK.

Gibbard, A., & Harper, W. (1978). Counterfactuals and Two Kinds of Expected Utility. Hooker A., Leach J. J., & McClennen E. F. (eds) *Foundations and Applications of Decision Theory*, pp. 125–62. D. Reidel: Dordrecht.

Hasker, W. (1992). The necessity of gratuitous evil. *Faith and Philosophy*, 9/1: 23–44.

Kavka, G. S. (1983). The toxin puzzle. *Analysis*, 43/1: 33–6.

Mavrodes, G. I. (1963). Some puzzles concerning omnipotence. *The Philosophical Review*, 72/2: 221–3.

McClennen, E. F. (1990). *Rationality and dynamic choice: foundational explorations*. Cambridge, UK: Cambridge University Press.

Murray, M. (2008). *Nature Red in Tooth and Claw : Theism and the Problem of Animal Suffering*. Oxford: Oxford University Press.

Swinburne, R. (2004). *The Existence of God*, 2nd edition. New York: Oxford University Press.

Wedgwood, R. (2011). Gandalf's solution to the Newcomb problem. *Synthese*, 190/14: 2643–75.

[1] See, for example, (Gauthier 1998; McClennen 1990).

[2] The responsibility theodicy thus presupposes "open theism," which holds (in opposition to theological determinism and Molinism) that God cannot know in advance what some creature

would freely do in various possible circumstances. (Or at least God cannot *always* know this in advance. On some more attenuated libertarian conceptions of free will, an act can count as free even if it is causally determined by the agent's character, as long as that character was formed on the basis of one or more "base case" free choices that were not causally determined by anything. Given this conception of free will, God could know in advance what some agent would do if a freely-formed character would determine her to do it, but God still could not know in advance what an agent would do in "base case" instances of free choice.)

[3] Here I suppose that the outcomes of Lola's actions are deterministic rather than probabilistic. A more careful discussion of Lola's responsibility would be required if we allow that *B* only results in *E*'s having a certain probability (one that is higher than the probability resulting from *G*). But I believe that this added complexity would not affect any of the central problems or solutions to be considered.

[4] More precisely, we have an explanation of why there is significant *moral* evil—evil that is the result of free choices of moral agents. Natural evil that does not result from the choice of creatures requires a different explanation.

[5] This point is emphasized in (Hasker 1992). Even if the Responsibility Theodicy is correct, it is possible that some evils are allowed for sake of goods that they make possible and which defeat them. But this is not essential to the Responsibility Theodicy, and there is no pressure to affirm that the majority of evils are defeated.

[6] Of course much more would need to be said in order to fully support the claim that Defeat Theodicies conflict with commonsense attitudes regarding evils and that this conflict constitutes a significant count against Defeat Theodicies. That is not the focus of this paper. Here, I only aim to give the reader a sense of the (putative) explanatory advantage of the Responsibility Theodicy.

[7] Additionally, Kavka stipulates that no significant bad comes from Verity's going back on some intention. Thus, she does not change the payoffs associated with refusing the toxin merely by forming the intention not to drink it.

[8] Presumably, the possibility sort of divine self-binding would be rejected by those who agree with Mavrodes (1963) and others that God could not make a stone that is too heavy for God to lift.

[9] For an interesting and helpful discussion of appeals in theodicy to the intrinsic goodness of nomological regularity, see (Murray 2008, chs. 5-6).

[10] Thanks to [NAME REMOVED] for urging me to consider this suggestion, and for helpful discussion.

[11] Of course this is compatible with it being the case that the *best* scenario would be the one where Lola makes the right decision and thereby prevents Cyril's death.

[12] [ACKNOWLEDGMENT REMOVED].

[13] To clarify, an "action plan" may involve a simple action (e.g., "drink the toxin"), or it may involve a decision procedure that has more than one possible outcome. So when we observe a tennis player counter a deep cross-court shot with a cross-court shot of his own, we cannot conclude that he endorses hitting a cross-court shot in *all* equivalent situations (even if we were to assume his shots are fully reflective actions). Rather, the action plan may be something like "typically counter with a crosscourt shot, but sometimes hit it down the line, and even more rarely hit a drop shot."

[14] It may be helpful to underscore the difference between this account and the account of the evidential decision theorist. Consider a revised scenario where Verity knows that after every coin toss, she will be given some time to drink the toxin (if applicable) and then her memory of the toss will be erased and her brain (and soul, if she has one!) will be completely "reset" back to the its state just before the coin was tossed. Because this process would break any causal connection between Verity's choice in the present toss and her choice in future tosses, a decision theory that ranks options according to their causal effects would have to concede that in this case Verity has no reason to drink the toxin when a coin lands heads. But the reasoning offered by the evidential decision theorist would not be affected.

[15] As Edgington (2011, 78) and others have observed, standard causal decision theory would be more aptly named "counterfactual decision theory."

[16] Egan's "Psychopath Button Case" involves Paul, who is highly confident that he is not a psychopath, deliberating over whether to press a button that would kill all living psychopaths. Paul would prefer to live in a world where all the psychopaths are killed, as long as this didn't involve *his* getting killed. But Paul's pressing the button would constitute very strong evidence that Paul is a psychopath, and thus that pressing the button would result in his death. Egan maintains (quite plausibly) that pressing the button is irrational. The case raises worries for the Difference-Making Principle. The good that motivates Paul's refraining from pressing the button is the good of staying alive; but if Paul does not press the button, Paul is highly confident that he is not a psychopath and thus that not pressing the button makes no difference to whether Paul stays alive. Because Paul is not *certain* that he is not a psychopath

(and thus not *certain* that his not pressing the button makes no difference to whether he stays alive), the Pyschopath Button Case is not, strictly speaking, a counterexample to the Difference-Making Principle. If we stipulated that Paul was certain, then it arguably becomes rational for him to press the button (since on standard Bayesian models, certainty that *p* is preserved even in the face of very strong evidence for not-*p*).

[17] From the fact that the causal decision theorist cannot affirm that Gavin should press the button, it does not follow that according to causal decision theory he should *not* press the button. Rather, I think the right think to say about this case is that causal decision theory does not issue a determinate verdict, at least not if we accept the supposition that Gavin is certain about the requirements of rationality and the fact that Hilda will conform to those requirements. Any reasons that establish the rationality of *not* pushing the button would also establish that Hilda will not press the button which in turn implies that Gavin rationally ought to push the button (since doing so would cause him to win money that he would otherwise forgo). Given standard causal decision theory, a rationale for *either* option is self-defeating. In this respect, applying causal decision theory to the Button Game exhibits a kind of decision instability that is reminiscent of the instability that characterizes the much-discussed "Death in Damascus" case (Gibbard & Harper 1978). But unlike the Death in Damascus case (and *like* Egan's cases), there does seem to be a unique rational decision in the Button Game.

[18] For proposals that deliver the correct verdict, see, e.g., (Edgington 2011; Wedgwood 2011).

[19] I myself feel that all of these commitments are more doubtful than the claim that it is rational for Verity to drink the toxins in the iterated toxin case. And this is true even though causal decision theory seems unable to support this verdict. Since causal decision theory supports a manifestly (to me!) irrational verdict in the structurally similar Button Game, the incompatibility of the decision theoretic solution with causal decision theory does not strike me as a significant cost at all. Of course others may weigh the balance of reasons differently.

[20] One might argue that my treatment of the red toxin case generalizes in a way that conflicts with my view that Verity could profit in the original multiple coin toss case. Here's how such an argument would go. Going back to this original multiple coin toss case, suppose the 14th coin toss lands heads and Verity is presented with a toxin. It seems that Verity could rehearse the following reasoning, which is analogous to the reasoning she gives in the red toxin case: "Now that I know that coin 14 landed heads, I also know that the policy of drinking every toxin *except* the toxin presented on the 14th toss will lead to the same payout as the policy of drinking all of the toxins presented to me. Since I have a reason to prefer the 'drink all toxins

except toxin 14' policy (the reason being that it will lead to a bit less nausea), I should adopt this policy and refrain from taking the toxin on this occasion." If this line of reasoning is rational, then we can predict that Verity will engage in this sort of reasoning on *every* occasion that she is presented with a toxin; and this means that she would not drink any toxin and therefore cannot profit from the arrangement. Since the envisioned line of reasoning is analogous to the reasoning she employs in the red toxin case, if we endorse this reasoning in the red toxin case then we should deny that Verity can profit in the multiple coin toss case. That is the argument against my position. Here's where the argument goes wrong. If an agent rationally does some action *x* in circumstances *C*, he thereby adopts a policy that generalizes to every situation that the he takes to be relevantly similar to *C*. So if Verity rationally does the action of "adopting the policy to drink all toxins except number 14," there must be some policy that she adopts that generalizes to every situation that she takes to be relevantly like her current situation. Consider the situation where coin toss 7 comes up heads and Verity is handed a toxin. This situation is (by Verity's lights) relevantly similar to her situation in toss 14. But Verity would *not* endorse the action of "adopting the policy to drink all toxins except number 14" in the event that toss 7 landed heads. This action would be irrational by her lights, since it would undermine her ability to win money in toss 14. So when Verity skips toxin 14, there must be a more general policy that she is also adopting that applies to all relevantly similar cases. And clearly there is. The policy she adopts may be put as follows: for any toss *n* that lands heads, adopt the policy of drinking all toxins except toxin *n*. *This* is the policy that produces equivalent results in all relevantly similar cases. But quite clearly, adopting this policy is not rational, since adopting it would make it the case that Verity would not drink any toxin, and would therefore eliminate the possibility of winning any money. Now turn to the red toxin case. When Verity is presented with the red toxin and decides not to drink it and to only drink the green toxins, she adopts a policy to act analogously in all relevantly similar cases. But since none of the other coin tosses are relevantly like her current situation (since none of them have the salient feature of being the uniquely worst outcome), Verity's action of skipping the red toxin does not amount to endorsing any actions in the other coin tosses. Skipping the red toxin has no bearing on the other tosses and therefore does not compromise Verity's ability to profit from the arrangement.

[21] Or at least they do not seem capable of explaining this fact if we assume that the existence of the human species is an unrepeatable event and that there are not other species of comparable moral significance that God interacts with.