The Toxin Puzzle
Author(s): Gregory S. Kavka
Reviewed work(s):
Source: *Analysis*, Vol. 43, No. 1 (Jan., 1983), pp. 33-36
Published by: Oxford University Press on behalf of The Analysis Committee
Stable URL: http://www.jstor.org/stable/3327802
Accessed: 24/10/2012 16:37

that *p* via a reliable method' may make no reference to the notion of 'self-warranting beliefs'. This is important since the notion of 'self-warranting beliefs' has been notoriously troublesome and a source of much epistemological mischief.

The idea that reference to reliable methods or processes plays a role in the analysis of knowledge and that it may replace reference to Cartesian-type justification, while not entirely new (the idea was apparently suggested by Ramsey in 1929),[4] seems to be an idea whose time has come. Richard Grandy, in a useful survey of some of the relevant literature, finds anticipations of the idea in work by Watling and by Unger, and a more developed form of it in work by Armstrong.[5] Grandy finds decisive objections to the analyses of each of these philosophers, but he argues persuasively that they are on the right track in stressing the epistemological importance of reliability over Cartesian-type justification. In the end, however, he confesses that 'we have thus far only a very poorly developed theory of reliability' (p. 209). That conclusion seems to be right. And until we develop a more adequate theory of reliability, epistemologists properly wary of promissory notes will remain sceptical whether the notion of reliability may be incorporated into an analysis of knowledge so as to provide the basis for a satisfactory response to scepticism.

University of Maryland,
College Park, Maryland 20742, U.S.A.

[4] F. P. Ramsey, *Foundations*, ed. by D. H. Mellor (London, 1978), pp. 126-7.

[5] Grandy, *op. cit.* Further developments of the idea may be found in Goldman, *op. cit.*, and in Fred Dretske, 'Conclusive Reasons', *The Australasian Journal of Philosophy*, 49 (1971) 1-22.

THE TOXIN PUZZLE

By GREGORY S. KAVKA

YOU are feeling extremely lucky. 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. (Your spouse, a crack biochemist, confirms the properties of the toxin.) 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. (This is confirmed by your daughter, a lawyer, after she examines the legal and financial documents that the billionaire has signed.) All you have to do is sign the agreement and then intend at midnight tonight to drink the stuff tomorrow afternoon. You are perfectly free to change your mind after receiving the money and not drink the toxin. (The presence or absence of the intention is to be determined by the latest 'mind-reading' brain scanner and computing device designed by the great Doctor X. As a cognitive scientist, materialist, and faithful former student of Doctor X, you have no doubt that the machine will correctly detect the presence or absence of the relevant intention.)

Confronted with this offer, you gleefully sign the contract, thinking 'what an easy way to become a millionaire'. Not long afterwards, however, you begin to worry. You had been thinking that you could avoid drinking the toxin and just pocket the million. But you realize that if you are thinking in those terms when midnight rolls around, you will not be intending to drink the toxin tomorrow. So maybe you will actually have to drink the stuff to collect the money. It will not be pleasant, but it is sure worth a day of suffering to become a millionaire.

However, as occurs to you immediately, it cannot really be necessary to drink the toxin to pocket the money. That money will either be or not be in your bank account by 10 a.m. tomorrow, you will know then whether it is there or not, and your drinking or not drinking the toxin hours later cannot affect the completed financial transaction. So instead of planning to drink the toxin, you decide to intend today to drink it and then change your mind after midnight. But if that is your plan, then it is obvious that you do not intend to drink the toxin. (At most you intend to intend to drink it.) For having such an intention is incompatible with planning to change your mind tomorrow morning.

At this point, your son, a strategist for the Pentagon, makes a useful suggestion. Why not bind yourself to drink the stuff tomorrow, by today making irreversible arrangements that will give you sufficient independent incentive to drink it? You might promise someone who would not later release you from the promise that you will drink the toxin tomorrow afternoon. Or you could sign a legal agreement obligating you to donate all your financial assets (including the million if you win it) to your least favourite political party, if you do not drink it. You might even hire a hitman to kill you if you do not swallow the toxin. This would assure you of a day of misery, but also of becoming rich.

Unfortunately, your daughter the lawyer, who has read the contract carefully, points out that arrangement of such external incentives is ruled out, as are such alternative gimmicks as hiring a hypnotist to implant the intention, forgetting the main relevant facts of the situation, and so forth. (Promising *yourself* that you

will drink the toxin could help if you were one of those strange people who take pride in never releasing oneself from a promise to oneself, no matter what the circumstances. Alas, you are not.)

Thrown back on your own resources, you desperately try to convince yourself that, despite the temporal sequence, drinking the toxin tomorrow afternoon is a necessary condition of pocketing the million that morning. Remembering Newcomb's Problem, you seek inductive evidence that this is so, hoping that previous recipients of the billionaire's offer won the million when and only when they drank the toxin. But, alas, your nephew, a private investigator, discovers that you are the first one to receive the offer (or that past winners drank less often than past losers). By now midnight is fast approaching and in a panic you try to summon up an act of will, gritting your teeth and muttering 'I will drink that toxin' over and over again.

We need not complete this tale of high hopes disappointed (or fulfilled) to make the point that there is a puzzle lurking here. You are asked to form a simple intention to perform an act that is well within your power. This is the kind of thing we all do many times every day. You are provided with an overwhelming incentive for doing so. Yet you cannot do so (or have extreme difficulty doing so) without resorting to exotic tricks involving hypnosis, hired killers, etc. Nor are your difficulties traceable to an uncontrollable fear of the negative consequences of the act in question – you would be perfectly willing to undergo the after-effects of the toxin to earn the million.

Two points underlie our puzzle. The first concerns the nature of intentions. If intentions were inner performances or self-directed commands, you would have no trouble earning your million. You would only need to keep your eye on the clock, and then perform or command to yourself at midnight. Similarly, if intentions were simply decisions, and decisions were volitions fully under the agent's control, there would be no problem. But intentions are better viewed as dispositions to act which are based on *reasons to act* – features of the act itself or its (possible) consequences that are valued by the agent. (Specifying the exact nature of the relationship between intentions and the reasons that they are based on is a difficult and worthy task, but one that need not detain us. For an account that is generally congenial to the views presented here, see Davidson's 'Intending', in his *Essays on Actions and Events*.) Thus, we can explain your difficulty in earning a fortune: you cannot intend to act as you have no reason to act, at least when you have substantial reasons not to act. And you have (or will have when the time comes) no reason to drink the toxin, and a very good reason not to, for it will make you quite sick for a day.

This brings us to our second point. While you have no reasons to drink the toxin, you have every reason (or at least a million reasons) to *intend* to drink it. Now when reasons for intending and reasons

for acting diverge, as they do here, confusion often reigns. For we are inclined to evaluate the rationality of the intention both in terms of its consequences and in terms of the rationality of the intended action. As a result, when we have good reasons to intend but not to act, conflicting standards of evaluation come into play and something has to give way: either rational action, rational intention, or aspects of the agent's own rationality (e.g., his correct belief that drinking the toxin is not necessary for winning the million).

I made some similar points in an earlier article ('Some Paradoxes of Deterrence', *Journal of Philosophy*, June 1978), but there I was discussing an example involving conditional intentions. The toxin puzzle broadens the application of that discussion, by showing that its conclusions may apply to cases involving unconditional intentions as well. It also reveals that intentions are only partly volitional. One cannot intend whatever one wants to intend any more than one can believe whatever one wants to believe. As our beliefs are constrained by our evidence, so our intentions are constrained by our reasons for action.[1]

University of California, Irvine,
California 92717, U.S.A.

[1] The puzzle discussed here emerged from a conversation, some years ago, with Tyler Burge about 'Some Paradoxes of Deterrence'. I have profitably discussed it with Paul Humphries, Rick O'Neil, and Virginia Warren, but am alone responsible for its present form and the conclusions derived from it. I am grateful to Doris Olin for suggesting a needed change in an earlier draft.

JIM AND THE INDIANS

By MARTIN HOLLIS

I WOULD not be writing these memoirs but for a nasty incident in the small South American town of _____ in the summer of 19___. Indeed I would have kept quiet altogether, if a fellow called Bernard Williams hadn't got wind of it and spread a version which I cannot endorse.[1]

[1] The direct references are to Bernard Williams' part of J. J. Smart and B. Williams, *Utilitarianism: For and Against*, Cambridge University Press, 1973, esp. p. 98ff., and the oblique ones to 'Internal and External Reasons' in Williams' collection of essays *Moral Luck*, Cambridge University Press, 1981. I would like to thank Peter Hobbis for his helpful comments.