sciendo FCT Fundação para a Ciência e a Tecnologia CIUHCT Ludus

Mathematizing the Toxin Puzzle

Leonard M. Wapner [1]

[1] *Professor Emeritus of Mathematics, Division of Mathematical Sciences, El Camino College, Torrance, California 90506 (USA)*

Abstract

A mathematical model of Gregory Kavka's "Toxin Puzzle" is constructed to provide the optimal course of action for this thought experiment. But the action suggested by the model is irrational, rendering the model useless for the rational player. Yet, it is the failure itself which provides a deeper insight into the nature of *intention* and mathematical modeling in general.

"Good intentions are not enough.
They've never put an onion in the soup yet."
— Sonya Levien

1 Introduction

Robert Nozick's "Newcomb's Problem" [Noz69] and Gregory Kavka's "The Toxin Puzzle" [Kav83] are closely related philosophical dilemmas of decision theory. The former, recognized as one of the most divisive problems in philosophy, has been popularized and well discussed in the literature. The toxin puzzle has received less attention. In this article a mathematical model of the puzzle is constructed to identify one's optimal (and thus most rational) choice when confronted with the puzzle's scenario. The model fails, as it suggests a highly irrational option as being optimal.

Despite the failure, indeed as a consequence of it, we can pick through the wreckage and expose deep truths associated with the mental state of *intention* and learn more of the limitations of mathematical modeling imposed by our own rationality.

2 The toxin puzzle

On Monday, a trustworthy billionaire emails you, offering you $1,000,000 if, by Tuesday, you will form the intention to voluntarily drink a toxin on Thursday of this week.

DOI 10.2478/rmm-2024-0013

The toxin will make you ill for one day but will have no long-lasting effects. The legitimacy of your intention will be verified Tuesday by a highly reliable, brain scanning lie detector. If your intention to drink the toxin is verified on Tuesday, the $1M will be deposited to your bank account on Wednesday. The billionaire makes it clear you need only form a legitimate and verifiable *intention* to voluntarily drink the toxin. Whether or not you actually drink it is irrelevant.

Easy money? Perhaps. After all, the ill effects of the toxin are temporary. And, you do not even have to drink the toxin. Marvelous!

A mathematical model will be used to answer two questions.

1. Should you form the intention to drink the toxin?
2. When given the lie detection test on Tuesday, should you truthfully state your intention? Or should you lie, possibly stating that you will drink the toxin when you have no actual intention of doing so, hoping the lie detector will fail to detect that you are lying?

Figure 1 gives the timeline.

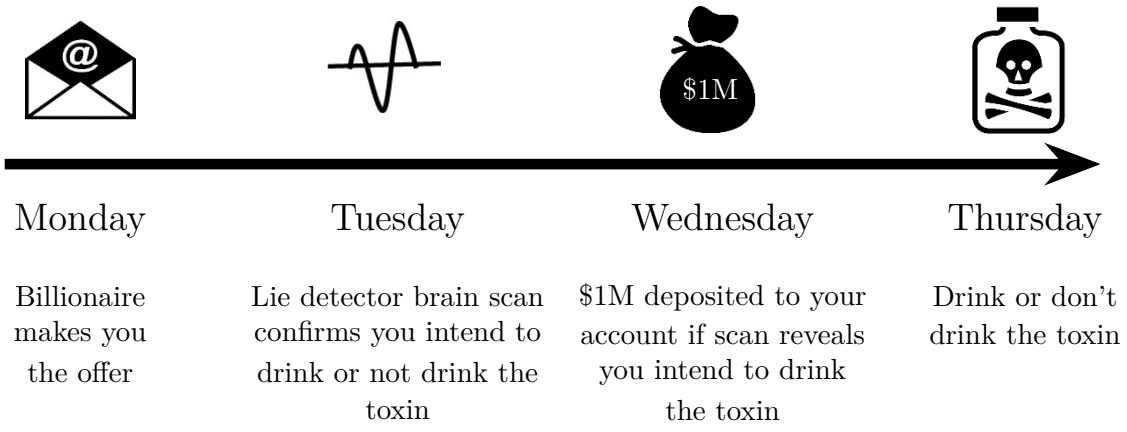

Figure 1: The toxin puzzle timeline.

3 Mathematical model

Let M ($M \geqslant 0$) denote the value (utility) of the deposit the billionaire is offering. (In Kavka's example, $M = 1,000,000$.) Let T ($T \geqslant 0$) be the magnitude (cost) of the suffering associated with drinking the toxin. Finally, we adopt the terminology of medical diagnostic testing to specify the accuracy of the lie detection test. The *sensitivity* of the test, denoted as P_{sen}, is the probability the test interprets your statement of intention as a lie, given that you are, in fact lying. The *specificity* of the test, denoted as P_{spec}, is

DOI 10.2478/rmm-2024-0013

the probability the test interprets your statement as truthful, given that you are, in fact, giving your true intention. If veracity is determined by chance alone, then $P_{sen} = P_{spec} = .50$. Kavka assumes perfect accuracy corresponding to $P_{sen} = P_{spec} = 1.0$. (In actuality the accuracy of state of the art lie detection (polygraph, brain and thermal imaging, voice stress analysis, etc.) is difficult to determine as multiple factors are involved, including the skill of the examiner. Current estimates place both sensitivity and specificity in the $.70 - .80$ range. Kavka's perfect accuracy will most certainly never be achieved; however, it is reasonable to assume accuracy will increase as research continues.)

As to the billionaire's offer, you have four options, each with its respective expectation.

1) You intend to drink the toxin and truthfully state so. The lie detector will interpret your intention correctly with probability P_{spec}, in which case your utility is $M - T$. The lie detector will incorrectly conclude you are lying with probability $1 - P_{spec}$ in which case your utility is $-T$. Overall, your expectation (via weighted utilities) is

$$E = P_{spec}(M - T) + (1 - P_{spec})(-T) = P_{spec}M - T$$

2) You intend to drink the toxin, but lie and state you will not do so. The lie detector will detect your lie with probability P_{sen} and detect your true intention to drink the toxin, in which case your utility is $M - T$. The lie detector will fail to detect your lie with probability $1 - P_{sen}$ and will incorrectly conclude you intend not to drink the toxin, in which case your utility is $-T$. Overall, your expectation (via weighted utilities) is

$$E = P_{sen}(M - T) + (1 - P_{sen})(-T) = P_{sen}M - T$$

3) You intend not to drink the toxin and truthfully state you will not do so. The lie detector will interpret your intention correctly with probability P_{spec}, in which case your utility is 0. The lie detector will incorrectly conclude you are lying with probability $1 - P_{spec}$, in which case your utility is M. Overall, your expectation (via weighted utilities) is

$$E = P_{spec}(0) + (1 - P_{spec})M = M - P_{spec}M$$

4) You intend not to drink the toxin, but lie and state that you will do so. The lie detector will detect your lie with probability P_{sen} and detect your true intention is to not drink the toxin, in which case your utility is 0. The lie detector will fail to detect your lie with probability $1 - P_{sen}$, and will incorrectly conclude you intend to drink the toxin, in which case your utility is M. Overall, your expectation (via weighted utilities) is

$$E = P_{sen}(0) + (1 - P_{sen})M = M - P_{sen}M$$

Depending on the exact values of all variables, it is possible that any one of these expectations might be the maximum of the four. So it would appear your decision is straightforward. You calculate the four expectations and choose your course of action corresponding to the maximum expectation. Table 1 summarizes the four expectations.

	State intention truthfully	Lie regarding intention
Intend to drink toxin	$P_{spec}M - T$	$P_{sen}M - T$
Intend not to drink toxin	$M - P_{spec}M$	$M - P_{sen}M$

Table 1: Expectations (general).

Three observations:

- If T is sufficiently large relative to M, the expectations in row 1 are negative, whereas those in row 2 remain nonnegative. In this case you should intend not to drink the toxin. Lie and say you intend to drink it if $P_{sen} < P_{spec}$.
- If both sensitivity and specificity are greater than .50 (chance accuracy) and T is sufficiently small, the expectations in row 1 will be larger than those in row 2, in which case you should intend to drink the toxin. Lie and state you do not intend to drink it if $P_{sen} > P_{spec}$.
- If, as assumed by Kavka, the lie detection test is 100% accurate ($P_{sen} = P_{spec} = 1$), you should intend to drink the toxin, assuming $T < M$. Whether or not you state your intention truthfully is irrelevant.

As a specific example let $M = 1,000,000$, $T = 100,000$, $P_{sen} = .80$, $P_{spec} = .90$. Table 2 gives the expectations associated with these values.

	State intention truthfully	Lie regarding intention
Intend to drink toxin	$800,000$	$700,000$
Intend not to drink toxin	$100,000$	$200,000$

Table 2: Expectations (specific).

4 Decision

From Table 2 your decision appears straightforward. The expectations in row 1 exceed those in row 2 and you should therefore form the intention to drink the toxin

and state so truthfully when tested. In Kavka's idealized version of the puzzle the lie detector is assumed to be perfectly accurate ($P_{sen} = P_{spec} = 1.0$), in which case you should again always form the intention to drink the toxin. In fact, assuming T is sufficiently small, values of sensitivity and specificity only slightly greater than .50 (chance accuracy) would compel you to form this intention. This level of accuracy is presumably achievable by today's technology and allows for real experimentation beyond Kavka's thought experiment version.

5 Analysis

The model is mathematically sound and your decision appears optimal and therefore rational. But there's a reason why Kavka describes this as a"puzzle". The model here suggests what you should do to achieve the optimal result, *but not necessarily what you, are capable of doing.* You, as a rational individual, have no reason whatsoever to actually drink the toxin on Thursday! Why not? When Thursday morning comes, the $1M may or may not have been deposited to your account by the billionaire. If you're concerned, you are free to contact your bank and check your account before drinking the toxin. Regardless of the recent account activity, nothing you do Thursday morning will effect the value of your account as this would involve a form of *retrocausation*, generally assumed to be impossible. The billionaire offered you $1M based only on your *intention* as of Tuesday to drink the toxin, not on whether or not you actually drink it on Thursday. And that intention has already been verified, correctly or not, by the lie detector test. So at this point the billionaire is no longer involved and has no interest in what you are about to do. Why drink the toxin and suffer needlessly?

Knowing this beforehand establishes that you will not drink the toxin and therefore, you as a rational individual can't form an honest intention to drink it. You can't lie to yourself! Based on the expectations in Table 2, you should lie to the lie detector and claim you do intend to drink the toxin when, in actuality, you have no such intention. You do this hoping the lie detector fails to detect your lie. It is unlikely you will receive the $1M no matter what you claim during the test. Easy money? Probably not.

Kavka's puzzle is an example of the so-called *curse of rationality.* One might be better off with no understanding of the mathematical model and having an irrational belief in retrocausation. Such an individual might truly intend to drink the toxin and actually do so on Thursday. And most importantly they would stand a far better chance of collecting the $1M than you, a rational individual with full understanding of the mathematical model.

Kavka writes,

> You cannot intend to act as you have no reason to act, at least when you have substantial reasons not to act. So, 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 [Kav83].

Is follow-through required to validate intention? These questions are more philosophical than mathematical; however, *it is the mathematical model of the toxin puzzle which brings this to light.*

There are real-world analogies. Hoefer et al. [Hoe19] provide one such example relating to the follow-through motion associated with swinging a golf club (tennis racket, cricket bat, . . .). Any golfer knows that follow-through is critical to achieving the desired ball projectile. However, the follow-through motion itself can't possibly influence the flight of the ball because follow-through occurs after the ball has left the club head. It is the intention to follow-through, involving alignment and proper motion of the club prior to striking the ball, which ultimately contributes to the desired trajectory. And it would appear that this alignment and proper motion can't be achieved unless the follow-through motion itself is actually executed. If, due to injury or other physical limitation, the golfer knows that the follow-through motion is impossible, then it would be impossible to make the proper alignment and swing prior to striking the ball. Similarly, if you know with certainty you will not be drinking the toxin on Thursday, it will be impossible for you to form the intention to do so when given the lie detector test.

A second example of validating intention is associated with the US and USSR nuclear defense strategy of mutually assured destruction (MAD), where if either side were to launch a nuclear strike against the other, the counterattack would be so overwhelming as to destroy the aggressor and potentially harm the entire planet. The actual defense associated with the strategy comes from each side knowing the other intends to respond as described, not from the response itself. But how does one side convince the other they are willing to follow through with such a devastating response?

A science fiction like mechanism to accomplish this was described by American nuclear physicist Herman Kahn in his book *On Thermonuclear War* [Kah60]. The concept is also central to the plot of the satirical film *Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb.* A doomsday machine is a device which automatically and efficiently, without human intervention, triggers the nuclear annihilation of the aggressor and potentially all human life on Earth if it detects a nuclear strike on the country maintaining the device. Follow-through, in the form of nuclear retaliation, is guaranteed and the intention to retaliate is thus validated. Allegedly, the Soviet Union had developed such a doomsday machine in the 1980s. This writer, however, is unable to verify the allegation.

6 Conclusion

Mathematical models are commonly used in problems of optimization where a rational individual must choose a course of action to maximize gain (minimize loss). But rationality itself can be a double-edged sword, which, ironically, may limit the model's usefulness. Kavka's scenario and its mathematical model given here serve as such an example. This irony should be considered when mathematical models of mental states, such as that of intention or some emotion, are designed.

"The Toxin Puzzle", as originally presented by Kavka, is a thought experiment involving a 100% accurate lie detection process. But as previously noted, the *puzzle* remains for actual, state of the art lie detection, where we may assume sensitivity and specificity exceed .50. This allows for real-world experimentation to determine if, in actuality, subjects believe follow-through by drinking the toxin is required to validate their stated intention of doing so. To my knowledge, no such study has been carried out.

References

[Hoe19] C. Hoefer et al. The Philosopher's Paradox: How to Make a Coherent Decision in the Newcomb Problem. In: *Theoria* 34(3) (2019) pp. 407–421.

[Kah60] H. Kahn. *On Thermonuclear War*. Princeton: Princeton University Press, 1960.

[Kav83] G. Kavka. "The Toxin Puzzle". In: *Analysis* 43(1) (1983) pp. 33–36.

[Noz69] R. Nozick. Newcomb's Problem and Two Principles of Choice. In N. Rescher (editor). *Essays in Honor of Carl G. Hempel*. Dordrecht: D. Reidel Publishing Company (1969) pp. 114–146.