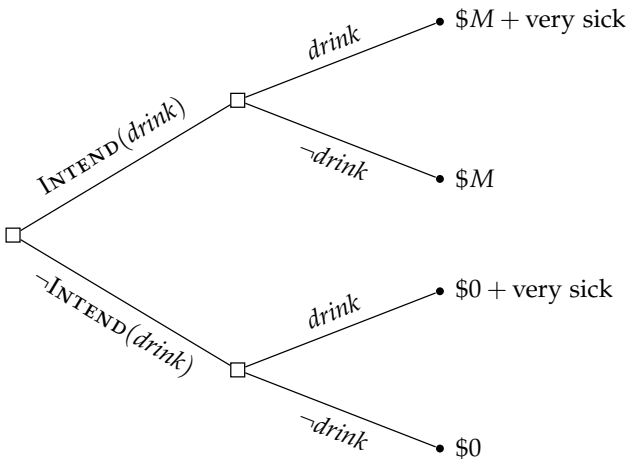
The Toxin Puzzle

Ryan Doody

To (Intend to) Drink or Not to (Intend to) Drink

Kavka's Toxin Puzzle: "An eccentric billionaire . . . places before you a vial of toxin . . . [and provides you with the following information:] If you drink [the toxin], [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. . . . 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. . . . [The] arrangement of. . . external incentives is ruled out, as are such alternative gimmicks as hiring a hypnotist to implant the intention. . . " (Kavka 1983, 33–4)

Kavka's Toxin Puzzle



The example raises three interesting questions:

1. Are you able to intend to drink the toxin?
2. Even if you are able, is it *rational* to intend to drink the toxin?
3. Is it rational to drink the toxin?

What do you think is the right answers to these questions? What can the Toxin Puzzle teach us about rationality and intentions?

Lessons for Intentions

Kavka thinks we should draw two lessons about intentions:

Your preference ranking (from best to worst):

$M

$M + very sick

$0

$0 + very sick

The Toxin Puzzle illustrates a tension between three plausible claims:

(1) If you're rational, you cannot form the *intention to* ϕ if you know it's irrational to ϕ.

- If you're rational and you know it's irrational to ϕ, you know you won't ϕ.
- You cannot intend to do something you know you won't do.

(2) It is rational to *intend* to drink the toxin.

- You'll win $M if, and only if, you intend to drink the toxin.
- You prefer ($M + very sick) to $0.

(3) You know it's irrational to drink the toxin.

- You know that you prefer $M to ($M + very sick) and $0 to ($0 + very sick).
- You know that drinking the toxin doesn't affect whether or not you get the $M.

What are some ways to resolve the tension?

(a) *Intentions are non-volitional.* Intentions are not "inner performances" or "self-directed commands." Instead, as our beliefs are constrained by our evidence, our intentions are constrained by our *reasons for action.*

. . . if they were, you would have no trouble forming the intention to drink the toxin. But, instead, it appears impossible—or, at least, very, very difficult—to do so.

(b) *An autonomous benefit case.* You have no reason to drink the toxin, but you do have reason to *intend* to drink the toxin. But, if you're rational, you can't intend to do something if you think there's no good reason to do it.

In autonomous benefit cases, you benefit from forming a certain intention but not from carrying out the associated action.

Consider, again, the claim:

(1) If you're rational, you cannot form the *intention to* ϕ if you know it's irrational to ϕ.

Example: Imagine that you know that it will be irrational for you to ϕ tomorrow, but suppose you also anticipate being irrational tomorrow! Can you, in this example, intend to ϕ? Could it be *rational* to?

Is it true? If so, is this a *psychological* fact? A *conceptual* fact? Or a *normative* fact? Can forming the intention to do something *change* what you think it might be rational to do?

Applications

1. *The Newcomb Problem.* Suppose you know that you'll face the Newcomb Problem tomorrow. Tonight, the predictor will scan your brain, which she'll use to predict whether you'll take one box or two. She'll either put the $\$M$ in the box or not. Tomorrow, you'll have to choose. Because in this case what you do tonight can *causally influence* the prediction, even CDT will advise you yo *intend* to *One box*. But, if you're rational, can you?

   If CDT is right, it's irrational for you to *One box* tomorrow. If you are reasonably self-aware, you—recognizing this fact—know that, tomorrow, you won't *One box* (even if you intend to). But if you know you won't *One box* tomorrow, you cannot form the *intention* to. And so, you won't get the $\$M$—even though, unlike in the original version of the Newcomb Problem, you have the ability to causally influence the prediction.

2. *Parfit's Hitchhiker.* "Suppose that I am driving at midnight through some desert. My car breaks down. You are a stranger and the only other driver near. I manage to stop you, and I offer you a great reward if you rescue me. I cannot reward you now, but I promise to do so when we reach my home. Suppose next that I am transparent, unable to deceive others. I cannot lie convincingly. Either a blush, or my tone of voice, always gives me away." (1984, 7)

   If the stranger rescues you and brings you to town, you will no longer have any reason to reward them. So, if you're rational and reasonably self-aware, you know that you won't reward the stranger if they rescue you. So, you cannot *intend* to reward them. And, because your intentions are transparent to the stranger, they won't rescue you.

3. *The Paradox of Deterrence.* Launching a retaliatory attack against the enemy—in addition to being immoral—might not, on balance, serve our national interests. Nevertheless, credibly *intending* to retaliate can be massively beneficial insofar as doing so acts as a deterrent against attack. Can it be rational to intend to do something you know, if the time comes, it won't make sense for you to do? Can it be moral to intend to do what you know is morally wrong?

   One real-life solution to this problem is to surrender control over to a *retaliation-agent*, of which there are three kinds: (i) people who are (for whatever reason) highly motivated to punish the offense in question, (ii) machines programmed to automatically retaliate if the offense occurs, (iii) a self-corrupted future self.