The Puzzle of the Self-Torturer

Ryan Doody

Self-Torture and Intransitive Preferences

Warren Quinn's (1990) 'Puzzle of the Self-Torturer' can be described as follows:

> Suppose someone—who, for reasons that will become apparent, Quinn calls the self-torturer—has a special electric device attached to him. The device has 1001 settings: $0, 1, 2, 3, \ldots, 1000$ and works as follows: moving up a setting raises, by a tiny increment, the amount of electric current applied to the self-torturer's body. The increments in current are so small that the self-torturer cannot tell the difference between adjacent settings. He can, however, tell the difference between settings that are far apart. And, in fact, there are settings at which the self-torturer would experience excruciating pain.
>
> Once a week, the self-torturer can compare all the different settings. He must then go back to the setting he was at and decide if he wants to move up a setting. If he does so, he gets $\$10,000$, but he can never permanently return to a lower setting. Like most of us, the self-torturer would like to increase his fortune but also cares about feeling well. Since the self-torturer cannot feel any difference in comfort between adjacent settings but gets $\$10,000$ at each advance, he prefers, for any two consecutive settings s and $s+1$, stopping at $s+1$ to stopping at s. But, since he does not want to live in excruciating pain, even for a great fortune, he also prefers stopping at a low setting, such as 0, over stopping at a high setting, such as 1000.

Chrisoula Andreou, "Dynamic Choice," *Stanford Encyclopedia of Philosophy*, 2020.

The self-torturer appears to have *rational* but *intransitive* preferences.

Let Z_s be the result of advancing to, and stopping at, setting s. Then, the following two assumptions seem plausible:

(1) *Adjacent*: For any setting s, the self-torturer prefers Z_{s+1} to Z_s.

(2) *Endpoint:* The self-torturer prefers Z_0 to $Z_{1,000}$.

The self-torturer prefers Z_{s+1} to Z_s because (i) in terms of pain, $s+1$ and s are *phenomenologically indistinguishable*, and (ii) Z_{s+1} comes with an additional $\$10,000$.

And the following seems like a fairly plausible principle:

> *Indiscernibility Principle:* If there's no discernible difference in pain between X and Y, but X offers a discernibly greater amount of wealth than Y (and those are the only two values that matter to you), then it's rational to prefer X to Y.

And, the self-torturer prefers Z_0 to $Z_{1,000}$ because $Z_{1,000}$ involves *so much* pain—it's excruciating!—that no amount of money is worth it.

Quinn makes the following assumptions as well:

(3) *Instrumental Rationality*: What you, rationally, ought to do is solely a function of how your relevant options compare in terms of satisfying your *actual* preferences.

"I am thinking of rationality (as I have been throughout) as instrumental—as something that is and ought to be the slave of the agent's preferences" (p. 90).

(4) *Feasibility:* For any setting s, the self-torturer, if at s, has the option to advance to and stop at $s+1$.

"No inability stands in his way. It isn't that he lacks the will-power to stop at some reasonable initial goal" (p. 88)

(5) *Fully Rational, Fully Aware:* The self-torturer will, on all occasions, act rationally—and the self-torturer, at all times, knows this.

Problems and Possible Solutions

Quinn first addresses some potential objections to his interpretation of his example.

It seems, then, like the self-torturer has the following preferences:

$$Z_0 \prec Z_1 \prec Z_2 \prec \cdots \prec Z_{999} \prec Z_{1,000} \prec Z_0$$

But aren't these the *reasonable* preferences to have in this situation?

And, given that these are the preferences the self-torturer has, what is it rational for them to do?

1. *The self-torturer's preferences are changing.* No they aren't. At each time (including at stage 0), the self-torturer prefers, e.g., $Z_{1,000}$ to Z_{999}. And, at each time, the self-torturer prefers a plan that results in Z_{s+1} to one that results in Z_s.
2. *We are neglecting behavioral evidence.* Suppose that, in addition to there being no phenomenological difference between s and $s+1$, there are not discernible *behavioral* differences either.
3. *We are ignoring the measures of his discomfort.* The measure of the self-torturer's discomfort is *indeterminate*.
4. *We are ignoring the effects of "triangulation".* Suppose that this cannot be done either.
5. *We are ignoring a preference reversal.* It needn't be true that, for any positive setting s, the self-torturer either prefers Z_0 to Z_s or prefers Z_s to Z_0. Instead, the self-torturers preferences are *indeterminate*: there is no *first* setting that he disprefers to Z_0 (and, in fact, there's no *first* setting at which it becomes *indeterminate* whether he disprefers to it Z_0).
6. *The self-torturer's preferences are paradoxical, and so raise no interesting issues for rational choice.* These preferences seems reasonable given the agent's predicament.

"There is no fact of the matter about *exactly* how bad he feels at any setting" (1990, p. 81).

Quinn considers a potential solution: "Why not pick a reasonable looking stopping point, proceed to it, and then *really stop*?" Because it appears to violate,

p. 85

> *The Principle of Strategic Readjustment:* Strategies continue to have authority only if following the strategy continues to offer you what you prefer overall.

The self-torturer knows that, for any stopping point they might plan to adopt, when the time comes to stop, following the plan won't be what they prefer overall.

It would be better to formulate some new stopping point that's further along—and so on and so forth until they reach $Z_{1,000}$!

Discussion Questions

1. The Puzzle of the Self-Torturer concerns what it's *rational* to do—that is, given one's preferences and (subjective) degrees of belief. But what, if anything, is *best* for the self-torturer to do *objectively*? What would you *advise* them to do?
2. Is the *Principle of Strategic Readjustment* plausible? Is it true?
3. What, if anything, can the Toxin Puzzle teach us about the Puzzle of the Self-Torturer?
4. Is the problem that the self-torturer has "vague" goals?
5. Are there any "real life" examples that resemble the predicament of the self-torturer?