SquirrelInHell

2018-02-27

Decision Theory and Suicide

Note: I am writing this on 27.02.2018. This might be important for future reference (e.g. to put a lower bound on how long I've been thinking about this). If I commit suicide, or spend my life on some altrustic pursuit too hastily, I wish to have pre-registered my suicide and also to have made it clear that it was not a fluke/accident/failure of rationality.

第一

A common argument against suicide goes something like "if you have nothing to lose, you should try living and seeing what happens!". A charitable interpretation is that this is only used in cases where there are in expectation positive experiences to look forward to, or the speaker is honestly convinced that this is the case. But even assuming that, the generalized argument is based on CDT (Causal Decision Theory), and always accepting it is a losing strategy.

If your utility function is completely selfish, when you are captured and certain to suffer terrible torture much worse than death, CDT tells you to instantly commit suicide. And if you are prevented from committing suicide until after the torture ends, causal decision theory tells you to keep on living (assuming positive experiences in expectation from that point onwards).

There are two obvious objections to this style of reasoning:

- How you are treated by other agents is a Newcomblike problem (e.g. people might do less X to others when they know X causes a 70% suicide rate in victims?).
- People (at least some of them) have terms in their utility function which are not selfish (e.g. they would choose to die to protect their family, or some number of strangers).

第二

To follow-up on the first objection, I think it's to a significant degree true that killing youself to make torture not-have-been done to you in a modern civilized society is perfectly valid TDT (Timeless Decision Theory) reasoning. *Note: I say "TDT" here because the name is more catchy and well-known, but you should really look at Updateless Decision Theory and Functional Decision Theory.*

Suicide is the ultimate costly signal of experiencing negative utility. (Where the positive/negative utility notation technically speaking requires calibrating zero utility to wherever not-existing-at-all is in a person's preference ordering over world histories.) Such signals must be hard to fake: almost every social group has their own list of grudges and complaints, and saying "my group's misery is more urgent/important than the other kinds" is a tough sell. Unless backed by the ultimate impossible-to-fake signal of a sky-high suicide rate.

And yes, of course many suicide attempts are caused by various types of mental breakdown, impulsiveness, or otherwise clouded judgement. But look beyond the noise. If you cannot acknowledge that there are really fates routinely experienced in apparently modern, civilized societies which worse than death, you aren't ready to objectively reflect on suicide.

To sum up: for bearers of selfish utility functions who expect negative utility, self-destruction is the obvious choice (proactive suicide). The case of expecting positive utility, but after having already suffered overwhelming negative utility (and obviously having lacked the foresight or agency to get out of it earlier) is more murky (retroactive suicide). This needs to be carefully considered with Timeless Decision Theory (or better) in mind.

第三

At this point I should probably interject and address a concern which is likely to arise as a response to me writing this post: apparently talking publicly about suicide sometimes causes increased suicide rates. However, people likely to read this are likely to be unusually able to think clearly, and not act on impulse; and if they aren't, they are unlikely to share my situation and reasoning style to any significant degree, and this seems to be the main factor at play here.

And as for those who after careful consideration, verifying that their thinking is not affected by confounders (e.g. depression), and applying adequate decision theory, determine it's better for them to self-destruct... I very much wish that they are able to do so.

Whenever you hear about a case of apparently well-considered suicide, please remember the suicide is *not a sad event*. All the sad things have already happened before the suicide, or if you prefer - they are embedded in the information needed to pick this particular world from among all possibilities. What makes such suicide seem sad to *you* is that *you then learn new information* about those sad things which have already happened.

第四

This brings us back to the second objection: what happens if someone's utility function is altruistic? This gets pretty complicated.

First, we are going to assume that there is a significant moral-patienthood-probablility mass of positive-utility lives at stake here. This assumption might be disputed, and it's easy enough to imagine an "altruist dystopia" in which everyone prefers to die, but also everyone pretends that this is not the case for the sake of everyone else, and tries to keep others alive. I don't think this is reasonable line of argumentation based on everything I know, and I will not engage with it.

Considering this, a strong argument for suicide of everyone who experiences neg-P-U (negative personal utility) is that if everyone does it, the universe becomes much, much easier to optimize. If the final goal of an altruistic utility function is to fairly maximize utility of all moral patients, it would be a

serious mistake to keep rational neg-P-U beings alive. If they can be trusted to self-destruct, the optimization process can to a first approximation simply keep everyone alive (who wishes to be alive, and to have been alive).

However, if we are thinking more narrowly about the reference class of human beings, it is obvious that their rationality is limited. To make things more murky, there's definitely a trend in which people who experience extreme suffering are *much more likely* to bootstrap into running improved approximations of rationality and sound decision theory. I won't go into the quirks of human psychology that cause this; but it's likely that you need only look around to find overwhelming evidence that it's true.

The way things stand now, humans are built on faulty hardware, and it seems roughly true that they just aren't able to run a good-enough decision algorithm in normal circumstances (i.e. having a happy life and such). So if the universal rule is "any rational decision maker with neg-P-U should self-destruct", we might actually end up with deaths of many important and interesting people capable of steering the future of humanity. The degree to which this would happen seems to be the crux of this line of reasoning.

If the current version of humanity without altruistic neg-P-U agents is significantly less able to take care of itself (e.g. deal with existential risk, coordination problems etc.), those agents should timelessly coordinate to let those of them who are the *most* unfortunate self-destruct. This trades off the more local consideration of humanity's fate against the broader-multiverse timeless consideration of the right of sentient beings to self-destruct.

The big question is: where to draw the boundary? A sharper boundary, drawn with more information, seems on the margin strictly better (agents are more accurately sorted by their degree of neg-P-U). I think there is a strong acausal/timeless link between all well-considered cases of self-destruction, and I feel a lot of pressure from this as an instance of the generalized decision algorithm. My personal judgement on this will bring all I have into the game, and is far too subtle to quote. I will merely list some obvious questions/prompts to consider at this point:

- How rare is reasonable timeless consequentialism in humanity (now, 5, 20 years in the future)?
- How is subjective, personal suffering distributed across people who implement it?
- In particular, what percentage is below the zero line?
- Is a typical instance very close to the zero line (only slightly positive or negative)?
- How rare is the level of altruism which takes precedence over personal life?
- How correlated is it with rationality and sound decision theory?
- What is the estimated impact per one person who implements both TDT and altruism?
- What is the relative magnitude of the timeless benefit from self-destructing in the proactive, and in the retrospective case?

Note: I am happy to discuss my reasoning in all of the above, but I will delete/not respond to any expression of personal concern about me.

No comments:

Post a Comment

Home

›

View web version

Powered by Blogger.