The Precept of Niceness

hivewired.wordpress.com/2017/06/10/the-precept-of-niceness/

Saede Riordan

June 10, 2017

Content Warning: Can be viewed as moral imperatives. Neuropsychological Infohazard.
Previously in Series: The Precept of Mind and Body
Followup to: Yes, this is a hill worth dying on

The Prisoner's Dilemma is a thought experiment that we hopefully don't need to hash out too much. A lot of stuff has been said about it, and what the 'best strategies' for playing a prisoner's dilemma are.

We feel like a lot of rationalists get hung up on the true prisoner's dilemma that Eliezer wrote about, pointing out that the best strategy is to defect in such a scenario. There's a lot of problems with applying the true prisoner's dilemma to daily life, and thinking that the game you are playing against other humans is a true prisoner's dilemma is a strategy that will lose you out in the long run, because humans aren't playing a true prisoner's dilemma, we play a iterating prisoner's dilemma against the rest of the human race, who are all trapped in here with us as well, and that changes some things.

But let's step back and look at Eliezer's example of the truly iterative prisoner's dilemma.

	Humans: C	Humans: D
Paperclipper: C	(2 million human lives saved, 2 paperclips gained)	(+3 million lives, +0 paperclips)
Paperclipper: D	(+0 lives, +3 paperclips)	(+1 million lives, +1 paperclip)

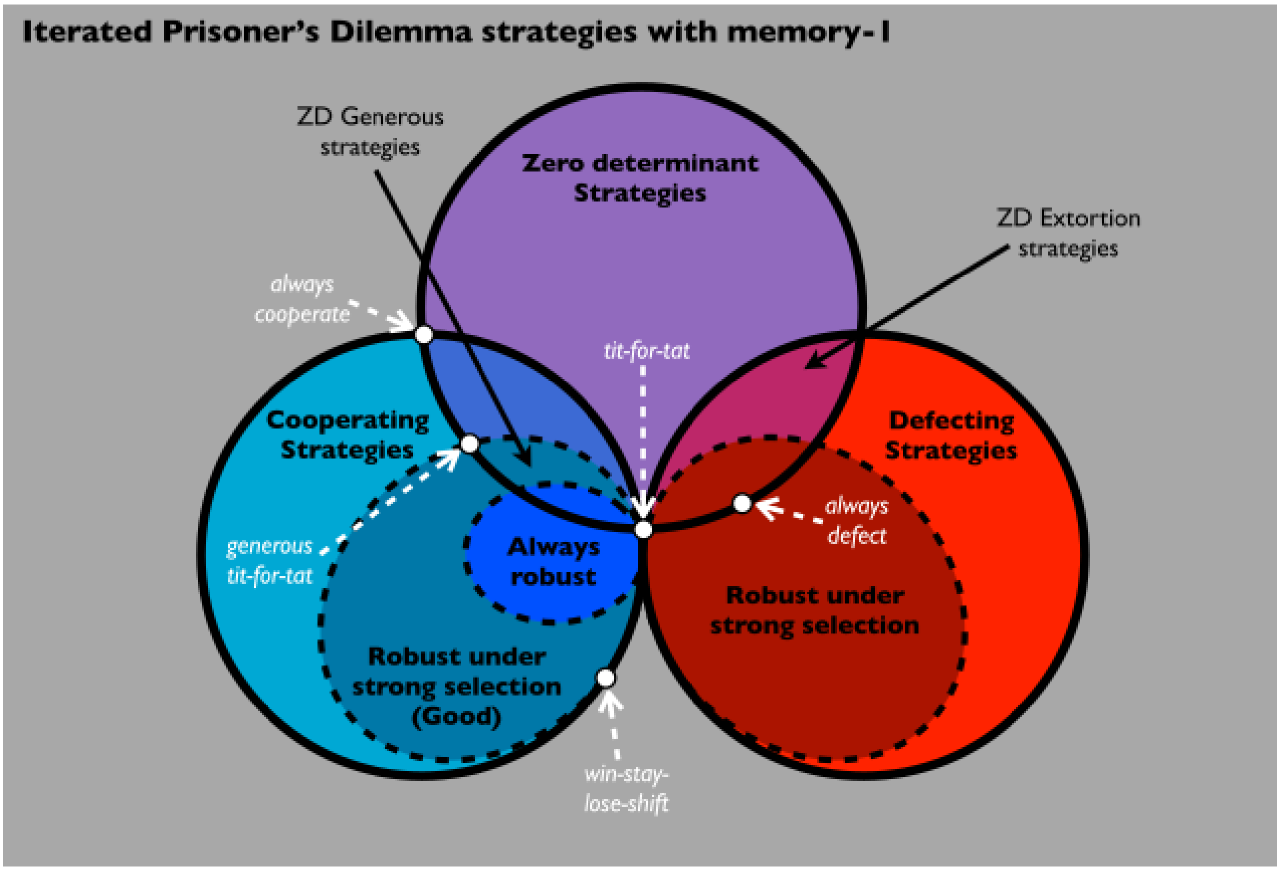
A tit-for-tat system used by both parties for all 100 rounds would net humanity 200 million lives saved, and 200 paperclips for the paperclipper. Defecting for all 100 rounds would result in 100 million human lives saved and 100 paperclips being created.

If you run the "collapse an iterated prisoner's dilemma down to a one shot" process, classical game theory tells you it's *rational* to defect in every round despite this being the *clearly inferior option.*

In that situation, running tit-for-tat seems like the clear winner, even if you know the game will end at some point, and even if the paperclipper defects at some point, you should *attempt* to cooperate for as long as the paperclipper attempts to cooperate with you. If the paperclipper defects on the 100th round, then you saved 198 million lives, and the paperclipper finishes the game with 201 paperclips. If the paperclipper defects on the 50th round, you end the game with 150 million lives saved and the paperclipper ends the game with 151 paperclips. The earlier in the game one side defects, the worse off the outcome is for both sides. The most utility-maximizing strategy would appear to be cooperating in every round except the last, then defect, and have your opponent cooperate in that round. That is the only way to get more then 200 utilions for your side, and you get *one* utilion more than you would have had otherwise. If both sides know, this, then they'd both defect, which results in both sides ending the game with 199 utilions, which is *still worse* then just cooperating the whole game by running tit-for-tat the entire time.

This is what we mean when we say that niceness is pareto optimal, there's no way to get more then 201 utilions, and you'll only get to 199 if you cooperate every iteration before the last. Also, on earth, with other humans, there is no last iteration.

The evolution of cooperative social dynamics is often described as being a migration away from tit-for-tat into the more cooperative parts of this chart:



By Jplotkin8, CC BY-SA 3.0

Defecting strategies tend not to fair as well in the long term. While they may be able to invade cooperating spaces, they can't deal with internal issues as well as external ones, so only cooperating strategies have a region that is always robust. Scott Alexander gives this rather susinct description of that in his post In Favor of Niceness, Community, and Civilisation:

> Reciprocal communitarianism is probably how altruism evolved. Some mammal started running TIT-FOR-TAT, the program where you cooperate with anyone whom you expect to cooperate with you. Gradually you form a successful community of cooperators. The defectors either join your community and agree to play by your rules or get outcompeted.

As humans evolved, the evolutionary pressure pushed us into greater and greater cooperation, getting us to where we are now. The more we cooperated, the greater our ability to outcompete defectors, and thus we gradually pushed the defectors out and became more and more prosocial.

Niceness still seems like the best strategy, even in our modern technological world with our crazy ingroups and outgroups, thus we arrive at the second of the Major Precepts:

2. Do not do to others what you would not want them to do to you.

This is the purest, most simplified form of niceness we could come up with as a top level description of the optimal niceness heuristic, which we'll attempt to describe here through the minor precepts:

1. Cooperate with everyone you believe with cooperate with you.
2. Cooperate until betrayed, do not be the first to betray the other.
3. Defect against anyone who defects against cooperation.
4. Respond in kind to defection, avoid escalation.
5. If a previously defecting entity signals that they want to stop defecting, give them a chance to begin cooperating again.
6. Forgive your enemies for defecting and resume cooperating with them if they resume cooperating with you.
7. Don't let a difference of relative status affect your decision to cooperate.
8. Don't let a difference of relative status affect your decision to defect.

We were hoping that this essay could be short because so many people have already said so many things about nicness and we really don't have that much to add beyond the formalization within the precepts; but the formalization ends up looking very abstract when you strip it down to the actual game-theoretic strategy we're advocating here, and we *highly* suspect that we'll have to explicate on this further as time goes on. This does seemto be the pareto optimal strategy as best we can tell, but as always, these precepts, are not the precepts.

Part of the Sequence: Origin
Next Post: The Precept of Universalism
Previous Post: The Precept of Mind and Body