An Approach to Giving In to Threats Without Incentivizing Them

Mikhail Samin
AI Governance and Safety Institute
ms@aigsi.org

Abstract

We present how Logical Decision Theory (LDT) can effectively handle threats, ultimatums, and commitments in decision-making scenarios, incentivizing cooperation and fair outcomes. By employing a proposed strategy, LDT agents can often give in to threats (losing far less utility than if they never gave in), but without making themselves exploitable or incentivizing others to make threats. We illustrate these principles through well-known game theory examples and offer ideas on how these strategies can be applied to broader domains. These ideas contribute to understanding how rational agents could achieve cooperation even when their notions of fair splits of gains are different, offering valuable implications for problems related to commitment races and s-risks.

1 Introduction

We present a strategy that Logical Decision Theory (LDT)[1] [7, 10, 5, 3] can use to give in to threats, ultimatums, and commitments, while incentivizing cooperation and fair[2] splits instead.

Although the core of this result is not new, we believe that further popularization of it is beneficial for the study of rational agents, decision theory, and AI safety. When we've demonstrated this strategy at the ILIAD conference and elsewhere, it made it much more intuitive to many researchers that smart agents probably won't engage in behavior like threatening each other, participating in commitment races, or otherwise expending utility to penalize other agents.

2 The Ultimatum Game

This part is largely taken from [11][3].

You're in the Ultimatum game. You're offered 0-10 dollars. You can accept or reject the offer. If you accept, you get what's offered, and the offerer gets $(10 - \mathit{offer})$. If you reject, both you and the offerer get

[1] Logical Decision Theories are a family of formal decision theories that systematically win in traditional decision-theoretic problems and are an improvement in the formal description of rational decision-making, studied mainly for their relevance to the safety of advanced AI systems. See the references for a more detailed introduction. Other decision theories would, if given a choice, decide to self-modify into decision theories that follow the strategy described here for all future games, but this discussion is out of the scope of this paper.

[2] Humans might use the Shapley value[4], the ROSE value[1], or their intuitive feeling of fairness. Other agents might use very different notions of fairness.

[3] See ProjectLawful.com: Eliezer's latest story, past 1M words.

nothing.

A simplest strategy that incentivizes fair splits is to accept everything ≥ 5 and reject everything < 5. The offerer can't do better than by offering you 5. If you accepted offers of 1, the offerer that knows this would always offer you 1 and get 9, instead of being incentivized to give you 5. Being unexploitable in the sense of incentivizing fair splits is a very important property that your strategy might have.

With the simplest strategy, if you're offered 5..10, you get 5..10; if you're offered 0..4, you get 0 in expectation.

Can you do better than that? What is a strategy that you could use that would get more than 0 in expectation if you're offered 1..4, while still being unexploitable (i.e., still incentivizing splits of at least 5)?

We encourage the reader to stop here and try to come up with a strategy before continuing.

The solution, explained in fictional settings in [11][4] (children split 12 jellychips, so the offers are 0..12[5]):

> When the children return the next day, the older children tell them the correct solution to the original Ultimatum Game.
>
> It goes like this:
>
> When somebody offers you a 7:5 split, instead of the 6:6 split that would be fair, you should accept their offer with slightly less than 6/7 probability. Their expected value from offering you 7:5, in this case, is 7× slightly less than 6/7, or slightly less than 6. This ensures they can't do any better by offering you an unfair split; but neither do you try to destroy all their expected value in retaliation. It could be an honest mistake, especially if the real situation is any more complicated than the original Ultimatum Game.
>
> If they offer you 8:4, accept with probability slightly-more-less than 6/8, so they do even worse in their own expectation by offering you 8:4.
>
> It's not about retaliating harder, the harder they hit you with an unfair price - that point gets hammered in pretty hard to the kids, a Watcher[6] steps in to repeat it. This setup isn't about retaliation, it's about what both sides have to do, to turn the problem of dividing the gains, into a matter of *fairness*; to create the incentive setup whereby both sides don't expect to do any better by distorting their own estimate of what is 'fair'.
>
> *[The next stage involves a complicated dynamic-puzzle with two stations, that requires two players working simultaneously to solve. After it's been solved, one player locks in a number on a 0-12 dial, the other player may press a button, and the puzzle station spits out jellychips thus divided.*
>
> *The gotcha is, the 2-player puzzle-game isn't always of equal difficulty for both players. Sometimes, one of them needs to work a lot harder than the other.]*
>
> They play the 2-station video games again. There's less anger and shouting this time. Sometimes, somebody rolls a continuous-die and then rejects somebody's offer, but whoever gets rejected knows that they're not being *punished*. Everybody is just following the Algorithm.

[4]The idea of unexploitable cooperation with agents with different notions of fairness seems to have first been introduced in [6], with agents accepting unfair (according to them) bargains in which the other agent does worse than in the fair point on the Pareto frontier; but it didn't suggest accepting unfair bargains probabilistically, to create new points where the other agent does just slightly worse in expectation than it would've in the fair point. One of the comments almost achieved the result, but didn't suggest adding $-\epsilon$ to the probability of giving in, so the result was considered exploitable (as the other agent was indifferent between making a threat and accepting the fair bargain). See also [9].

[5]*dath ilani*, a fictional human civilization in [11], use a base 12 number system.

[6]A dath ilani teacher.

> Your notion of fairness didn't match their notion of fairness, and they did what the Algorithm says to do in that case, but they know you didn't mean anything by it, because they know you know they're following the Algorithm, so they know you know you don't have any incentive to distort your own estimate of what's fair, so they know you weren't trying to get away with anything, and you know they know that, and you know they're not trying to punish you. You can already foresee the part where you're going to be asked to play this game for longer, until fewer offers get rejected, as people learn to converge on a shared idea of what is fair.
>
> Sometimes you offer the other kid an extra jellychip, when you're not sure yourself, to make sure they don't reject you. Sometimes they accept your offer and then toss a jellychip back to you, because they think you offered more than was fair. It's not how the game would be played between dath ilan and true aliens, but it's often how the game is played in real life. In dath ilan, that is.

So, if in the game with \$0..10, you're offered \$4 instead of the fair \$5, you understand that if you accept, the other player will get \$6 - and so you accept with the probability of slightly less than $\frac{5}{6}$, making the offerer receive, in expectation, slightly less than the fair \$5: $\mathbb{E}[U] = 6 * (5/6 - \epsilon) = 5 - (6 * \epsilon)$, for some small ϵ. You still get \$4 most of the time when you're offered this unfair split, but you're incentivizing fair splits. Even if you're offered \$1, you accept slightly less than in $5/9$ cases - which is more than half of the time, but still incentivizes offering you the fair 5-5 split instead.

If the other player makes a commitment to offer you \$4 regardless of what you do, it simply doesn't change what you do when you're offered \$4. You want to accept \$4 with $p = 5/6 - \epsilon$ regardless of what led to this offer. Otherwise, you'll incentivize offers of \$4 instead of \$5. This means other players don't make bad commitments (and if they do, you usually give in).

(This is symmetrical. If you're the offerer, and the other player accepts only at least \$6 and always rejects \$5 or lower, you can offer \$6 with $p = 5/6 - \epsilon$ [8] or otherwise offer less and be rejected.)

2.1 Different notions of fairness

This allows even very different agents with very different notions of fairness to cooperate most of the time. Among humans, the Shapley value and the ROSE value have been presented in [4] and [1] as notions of fairness that satisfy some natural desiderata; non-human agents might potentially have very different notions of fairness. The strategy presented here allows agents to cooperate most of the time (almost always for small differences in their notions of fairness), while not incentivizing what would be unfair according to them. For example, if one agent believes the fair split to be half and half in some game and another agent inherently believes that it should get 51%, a simple strategy to incentivize fairness by rejecting everything unfair would mean the two agents never cooperate, while the strategy presented here allows them to cooperate in over 98% of games while still incentivizing fairness.

3 Threats, Commitments, and Ultimatums

You can follow the same procedure in all games. Figure out the fair split of gains, then try to coordinate on it; if the other agent is not willing to agree to the fair split and demands something else, agree to their ultimatum probabilistically, in a way that incentivizes the fair split instead.

Importantly, rational agents can mutually cooperate in a way dependent on each other's cooperation due to results such as [2], which enables them to robustly mutually cooperate in the prisoner's dilemma,

though specific mechanisms for coordination and robust cooperation are outside the scope of this paper.

3.1 Game of Chicken

Let's say the payoff matrix is:

	Player 2: Swerve	Player 2: Don't Swerve
Player 1: Swerve	0, 0	-1, 5
Player 1: Don't Swerve	5, -1	-100, -100

Table 1: Payoff matrix for the Game of Chicken

Let's assume we consider the fair split in this game to be 2, you can achieve it by coordinating on throwing a fair coin to determine who does what.

If the other player instead commits to not swerve, you calculate that if you give in, they get 5; the fair payoff is 2; so you simply give in and swerve with $p = 97\%$, making the other player get less than 2 in expectation; they would've done better by cooperating. Note that this decision procedure is much better than never giving in to threats - which would correspond to getting -100 every time instead of just 3% of the time - while still having the property that it's better for everyone to not threaten you at all.

3.2 Stones

If the other player is a stone[7] with "Threat" written on it, you should do the same thing, even if it looks like the stone's behavior doesn't depend on what you'll do in response. Responding to actions and ignoring the internals when threatened means you'll get a lot fewer stones thrown at you.

3.3 What if I don't know the other player's payoffs?

You want to make decisions that don't incentivize threatening you. If you receive a threat and know nothing about the other agent's payoffs, simply don't give in to the threat! (If you have some information, you can transparently give in with a probability low enough that you're certain transparently making decisions this way isn't incentivizing this threat.)

3.4 What if the other player makes a commitment before I make any decisions?

Even without the above strategy, why would this matter? You can just make the right decisions you want to make. You can use information when you want to be using it and not use it when it doesn't make sense to use it. The time at which you receive the information doesn't have to be an input into what you consider if you think it doesn't matter when you receive it.

With the above algorithm, if you receive a threat, you simply look at it and give in to it most of the time in many games, all while incentivizing not threatening you, because the other player can get more utility if they don't threaten you.

(In reality, making decisions this way means you'll rarely receive threats. In most games, you'll coordinate with the other player on extracting the most utility. Agents will look at you, understand that threatening you means less utility, and you won't have to spend time googling random number generators and

[7] A player with deterministic behavior in a game with known payoffs is sometimes called a "stone" after the idea of stones with "Cooperate" written on them in the Prisoner's Dilemma. It's desirable for rational agents to always defect against deterministic agents that don't change what they do based on any external circumstances.

probabilistically giving in. It doesn't make sense for the other agent to make threatening commitments; and if they do, it's slightly bad for them.

It's never a good idea to threaten an LDT agent.)

3.5 Implications

For humans, the proposed strategy is applicable to bargaining and threat-shaped situations.

In the study of the safety of general AI systems, we want to understand the space of possible minds, describe a part of that space that contains a safe target, and then engineer our way into achieving that target while understanding and avoiding failure modes. Decision theories that agents implement are an important feature of the space of possible minds. Moreover, if we find AI systems with deep learning or if AI systems experience other optimization pressure, they're likely to naturally implement the kinds of decision theories that allow them to systematically win and achieve goals, which makes properties and implications of logical decision theories plausibly describe properties of future AI systems.

We would expect future general AI systems and other agents, including, over long time horizons, even aliens, to usually succeed at coordinating with each other. We wouldn't expect commitment races. We find it inspiring that across a large universe, agents would be incentivized to deal fairly with each other, and would usually mutually cooperate. Unfortunately, humans aren't able to participate in coordination schemes of that type, and this paper should not be taken as implying that AI will be nice to humans.

We found the proposed strategy makes it more intuitive to many that LDT agents always have a positive value of information and don't try to avoid learning: new information can't coerce them into losing utility if they simply behave to incentivize fair splits regardless of the information they see. Agents can simply ignore the reasons for threatening behavior and refuse to give in often enough that it doesn't make sense to threaten them.

We think this implies that threats are very rare among rational agents. We further think it's unlikely s-risks would materialize, as rational agents respond to threats in ways that make threatening them (or making utility-penalizing commitments) negative in expectation, compared to cooperation.

References

[1] Alexander Appel. *Threat-Resistant Bargaining Megapost: Introducing the ROSE Value*. 2022. URL: https://lesswrong.com/posts/vJ7ggyjuP4u2yHNcP/threat-resistant-bargaining-megapost-introducing-the-rose.

[2] Mihaly Barasz et al. *Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic*. 2021. arXiv: 1401.5577 [cs.GT]. URL: https://arxiv.org/abs/1401.5577.

[3] Wei Dai. *Towards a New Decision Theory*. 2009. URL: https://lesswrong.com/posts/de3xjFaACCAk6imzv/towards-a-new-decision-theory.

[4] Lloyd S. Shapley. "A Value for n-Person Games". In: *Contributions to the Theory of Games (AM-28), Volume II*. Ed. by H. W. Kuhn and A. W. Tucker. Princeton University Press, 1953, pp. 307–317. URL: https://www.rand.org/content/dam/rand/pubs/papers/2021/P295.pdf.

[5] Nate Soares and Benja Fallenstein. *Toward Idealized Decision Theory*. 2015. arXiv: 1507.01986 [cs.AI]. URL: https://arxiv.org/abs/1507.01986.

[6] Eliezer Yudkowsky. *Cooperating with agents with different ideas of fairness, while resisting exploitation*. 2013. URL: https://lesswrong.com/posts/z2YwmzuT7nWx62Kfh/cooperating-with-agents-with-different-ideas-of-fairness.

[7] Eliezer Yudkowsky. *Introduction to Logical Decision Theory for Computer Scientists - Arbital*. 2016. URL: https://www.lesswrong.com/w/logical-decision-theories?l=5d6&lens=introduction-to-logical-decision-theory-for-computer.

[8] Eliezer Yudkowsky. *The Commitment Races Problem - Reply*. 2022. URL: https://lesswrong.com/posts/brXr7PJ2W4Na2EW2q/the-commitment-races-problem?commentId=tYBPjetgZW4iMqe4s.

[9] Eliezer Yudkowsky. *Ultimatum Game - Arbital*. 2016. URL: https://arbital.com/p/ultimatum_game?l=5tp.

[10] Eliezer Yudkowsky and Nate Soares. *Functional Decision Theory: A New Theory of Instrumental Rationality*. 2018. arXiv: 1710.05060 [cs.AI]. URL: https://arxiv.org/abs/1710.05060.

[11] Iarwain (Eliezer Yudkowsky) and lintamande. *planecrash (a fictional story) - Jellychip Division*. 2021. URL: https://www.glowfic.com/replies/1729110#reply-1729110.