

A paradox for supertask decision makers

Andrew Bacon

December 2, 2008

When faced with an infinite sequence of decisions, rational agents can do some very strange things. Barrett and Arntzenius [1] present a puzzle where the agent is offered an infinite sequence of choices. If the agent acts rationally at each point, then he is guaranteed to make a loss after the sequence is over and he may be certain of this. You might take this to be a counterexample to the principle that if one acts rationally with respect to an interval of time, then one has acted rationally at every subinterval.

The puzzles I am concerned with involve a violation of a related principle. An agent may undergo a sequence of decision problems such that it is possible to respond rationally to any given problem, yet it is impossible to respond rationally to all of them together, even though the choices are independent. The first puzzle involves a game between two players in which both players have a winning strategy. It is thus impossible for both players to follow their strategy, meaning that there must have been point at which at least one player was able to play according to his strategy, wanted to, but didn't. The second puzzle involves offering an agent a sequence of choices such that it is possible to respond rationally to each choice, making a guaranteed profit, yet it is impossible to respond rationally at every point, meaning there's a time when you're able to act so as to maximise utility, but don't.

1 Puzzle one

The first puzzle may be described as a game between two players: Alice and Bob.

For each $n \in \omega$, at $\frac{1}{n}$ hours past 12pm Alice and Bob will play a round of the game. A round involves two moves: firstly Alice chooses either 1 or 0, and then Bob makes a similar choice. The moves are made in that order, and both players hear each choice. Alice wins the round if Bob's choice was the same as hers, and Bob wins if his choice is different. The game finishes at 1pm, Alice wins the game if she wins at least one round, Bob wins the game if he wins every round.¹

¹It seems natural to suppose that backwards supertasks, such as this one, are possible if

On the face of it, Alice doesn't have a hope in hell of winning. For all Bob has to do, at each round, is to say exactly the opposite of what Alice says. Since there's nothing Alice can do to prevent him doing this at every round, it seems she's bound to lose.

It should, then, be quite surprising to find out that there *is* something Alice can do to ensure she wins. That is, Alice has a winning strategy. By a strategy for this game, I mean a function which takes any possible initial sequence of moves of the game to a move. The move represents what choice that player would make on the upcoming round given an initial sequence of play, if she were adopting that strategy. A winning strategy for a player is one such that, if at every point in the game the player makes the move that strategy suggests given the sequence of moves played so far, that player will win. I shall occasionally talk about the n th round, by which I mean the n th round from the end, i.e., the round that takes place at $\frac{1}{n}$ hours past 12.

As I said, Alice has a winning strategy for the game described above. There are various ways that Bob could play throughout a whole game, but any way he plays can be encoded as an ω -sequence of 1's and 0's, where the n th term in the sequence represents how he responds in the n th round. Before the game starts, Alice chooses her strategy as follows. Alice divides these sequences into equivalence classes according to whether they differ by finitely many moves at most. With the help of the Axiom of Choice, Alice then picks a representative from each equivalence class. At any point after the game has started, Alice will know what moves Bob has made at infinitely many of the rounds, and will only be ignorant of the moves Bob is yet to play, of which there are only finitely many. Thus, at any point after 12pm, Alice will know to which equivalence class the sequence of moves Bob will eventually make belongs. Her strategy at each round, then, is to play how the representative sequence for this equivalence class predicts Bob will play at that round. If the representative sequence is right about Bob's move at that round, Alice will win that round. However, the representative sequence and the sequence that represents how Bob actually played, must be in the same equivalence class: they must be the same at all but finitely many rounds. If Alice played according to the representative sequence at every round, then she will have won all but finitely many of the rounds, meaning that she has won the game.

This result has some very surprising consequences. For example, suppose Bob decided to flip a coin to decide his move at every round. If Alice follows her strategy, then she will be guaranteed to correctly guess infinitely many of the coin flips, and indeed, there will be a point at which she has correctly guessed infinitely many flips *in a row*. Intuitively, there should be no strategy that could guarantee that Alice guesses even one flip correctly, if the coin is fair. Secondly, Bob also appears to have a winning strategy: all Bob needs to do is say the opposite of what Alice says at every round.

the ordinary kind are. For example, presumably Zeno's performs such a supertask every time he moves, in much the same way as he performs a forwards supertask.

This last fact should be puzzling, since only one player can win a game. This means that for any given game, at least one player will not be able to successfully implement their strategy at every round despite, we may assume, being physically able to, and wanting to.

2 Puzzle two

One might have thought that the problem in the last section was due to the axiom of choice, or the possibility of beings that can grasp infinitely complex strategies. The following puzzle does not involve either of these elements.

For each $n \in \omega$, at $\frac{1}{n}$ hours past 12pm, Alice will be asked to choose either 1 or 0. If she answers according to the following rule, she will receive a chocolate.² The rule is: choose 1, if you have chosen 0 at every previous round, and chose 0 otherwise (i.e. if you have chosen 1 on at least one other round.)

Much like Bob's strategy in the first game, this is a relatively easy strategy to implement, and it does not require the axiom of choice to generate. However, it is not possible to follow the rule at every time between 12pm and 1pm. The reasoning is essentially that involved in Yablo's paradox: either Alice chooses 1 on some round, or she always chose 0. (i) if on some round, she correctly followed the rule and chose 1, then she must have chosen 0 on all the previous rounds. In particular, she must have chosen 0 on the immediately preceding round. In which case she has violated the rule on this round, since she chose 0, when all the previous rounds were 0. (ii) if she always chose 0, then she violated the rule at every round by not choosing 1.

3 Consequences

It is commonly thought that to be rational is to have certain dispositional properties. It is not enough to have always acted in the most rational way, otherwise one could be rational by having never needed to make any decisions at all. A perfectly rational agent must also be *disposed* to act rationally in the sense that, if he were offered a given decision problem, he would respond in a way that maximised his expected utility. The robustness of rational behaviour under such counterfactual suppositions is essential, for example, in game theory for motivating the various equilibrium concepts.

Just how one spells out these dispositions is a delicate matter. For example, we should not expect the agent to continue to behave rationally if he underwent some cognitive malfunction. But what seems clear is that the choices involved at each round of the two puzzles above have the quality of a decision problem, where the agents are free to act rationally. In the first puzzle, we may assume

²Or at least, something that can immediately be converted into hedons before the next choice.

that both players have a stake in winning the game. To avoid irrelevant details we may go one further and stipulate that both players have a stake in playing a given winning strategy at each round. Similar remarks apply to the second puzzle. So as not to get mixed up with accumulating infinite utilities, we may think of each round as a single decision problem, having a reward which gets spent before the next round.

The fact that, necessarily, there will be a player and a round in one of the two puzzles who does not act so as to maximise utility, suggests that no one can have the right counterfactual properties required of perfect rationality.

On closer scrutiny, however, this doesn't quite follow. Let us concentrate on the second puzzle. For each possible initial sequence of choices Alice could have made, t , let E_t be the proposition that Alice has received evidence that she has so far chosen according to the sequence t . Let R_t be the proposition that Alice acts rationally at the next round: she follows the rule and receives a reward. Then we may consistently assert the following schema:

$$E_t \Box \rightarrow R_t \tag{1}$$

Each instance is intuitively true if Alice is rational, since in the closest worlds where Alice has taken part of a game up to the point t , she responds by following the rule so that she may receive a reward.

So it seems like Alice can consistently have the dispositions required to be rational, by responding correctly in the merely possible situations where E_t obtains - even if she has acted irrationally earlier at that world. But what happens if Alice is in a world where she is actually going to be put through one of these supertask decision sequences? Since for each sequence that actually occurs, t , we have E_t and $E_t \Box \rightarrow R_t$ we may infer that R_t . This is impossible: Alice cannot follow the rule on every round. It is impossible for an ideally rational agent to find themselves in a situation where they will undergo such a procedure.³

The puzzles also seem to give rise to counterexamples to the deontic Barcan formula. For at each point in the game rationality requires Alice to follow her strategy, yet to require that Alice follow it at *every* point would be to require the impossible.⁴⁵

³One might put it paradoxically: suppose that Alice and Ecila are intrinsic duplicates at 11am. Between 12pm and 1pm Alice will undergo one of these decision sequences, while Ecila does not, and never will. Ecila has the right dispositional properties to be rational, yet Alice does not, since there will be some instance of (1) she does not satisfy.

⁴One might think the deontic Barcan formula fails for more mundane reasons. The interesting thing about these puzzles is that they provide counterexamples to the much weaker principle that a conjunction of requirements is a requirement - which remains valid even over the class of variable domain Kripke models. Similar remarks apply also to the Barrett-Arntzenius puzzle.

⁵[Acknowledgements]

References

- [1] Jeffrey Barrett and Frank Arntzenius. An infinite decision puzzle. *Theory and Decision*, 46(1):101–3, 1999.
- [2] Stephen Yablo. A reply to new zeno. *Analysis*, 60(2):pp. 148–151, 2000.