

# A Tarskian Solution to the Problem of the Iterated Prisoners Dilemma

-Christopher Foster

Submission for Utah Philosophical Society, 10/9/2006

The familiar game known as the *prisoners dilemma* presents a situation in which mutual defection is rational, despite the fact that mutual cooperation would be Pareto optimal. This situation is ironic, but not quite paradoxical. Worse things occur, however, when we entertain the problem of the *iterated prisoners dilemma* – or a repeated, n-game case. In this case the standard reasoning goes: No matter what occurs in the first n-1 games, we will both defect in the last game (both knowing that there will be no more games we treat it as a one-game case). But then in the n-1<sup>st</sup> game, we know that our result will affect no future games, so we both defect again (since it again reduces to a one-game case). But the same reasoning applies to the n-2<sup>nd</sup> game case, etc. The result is that two rational players must defect the entire time, resulting in a terrible mutually negative score. This answer is the most commonly accepted view – that two rational players must defect the whole time, despite the startlingly negative consequences.

Many have tried to present ‘solutions’ (reasons a rational agent might cooperate) to this problem. I will argue that these proposals fail: David Lewis’ (and others’) *symmetry argument* is invalid, while Pettit and Sugden’s solution of convincing the other of one’s irrationality is self-defeating<sup>1</sup> and therefore unsound. I will, however, show that Pettit and Sugden’s reasoning, while unsound, proves that the problem of the iterated prisoners dilemma is a genuine paradox (since one can be rational by being irrational) and not merely a tough bullet to bite.

Faced with paradox, I turn to Tarski’s work on the concept of truth and find that *rationality*, like knowledge and truth, is a meta-predicate – it is a predicate that can properly only be applied to a situation that does not include itself. To prove this I will show that if one’s own rationality is allowed to be part of the situation upon which one is supposed to act then there are situations in which one is rational if and only if one isn’t. The conclusion must be drawn, with Tarski, that rationality can only be decided when acting upon a situation that does not include one’s own rationality. In the iterated prisoners dilemma the necessary assumption of mutually-known rationality violates this condition.

Therefore, I may choose rationally under the assumption that the *other* agent is rational, but not that he or she knows that I am. I may assume his or her rationality (but not my own) and may then *choose rationally* based on that assumption. Under such reasoning, the meta-strategy that I adopt is as follows: choose a strategy such that the other person’s best reply leaves me as well off as possible. The application of this meta-strategy in the iterated n-game prisoners dilemma results in a strategy of nice conditional cooperation (cooperation that never defects first but will defect if the other does so first). The simplest example of such a nice strategy is Tit-For-Tat. The Tit-For-Tat strategy has been shown to be optimal in the experimental results of Robert Axelrod, and has many interesting applications to politics and evolutionary biology.

According to my solution, the other player cheats me on the last game, but my results are *far* better than the standard *always defect* strategy. My paper does not find an *ad hoc* reason to justify a mutually beneficial outcome but demonstrates *why* the standard argument (based on the dominance principle) is wrong and why mutual conditional cooperation is the most rational possible strategy. By treating rationality as a metapredicate I open up a new way of solving problems and paradoxes of rationality and possibly pave the way for greater cooperation in economics, politics, and other situations that demand mutual cooperation.

---

<sup>1</sup> Meaning that its conclusion contradicts one of its premises.