Penultimate draft

11. The Inner Struggle: Why You Should Cooperate With People You Will Never Meet Again

---

Martin Peterson

1. Introduction

In Book IV of the *Republic*, Plato argues that the soul consists of three separate entities: reason, spirit, and appetite. In a key passage, Socrates and Glaucon discuss the inner struggle of a thirsty man who desires to drink at the same time as he feels an urge not to do so:

> "The soul of the thirsty then, in so far as it thirsts, wishes nothing else than to drink, and yearns for this and its impulse is towards this."
> "Obviously."
> "Then if anything draws it back when thirsty it must be something different in it from that which thirsts and drives it like a beast to drink. For it cannot be, we say, that the same thing with the same part of itself at the same time acts in opposite ways about the same thing."
> "We must admit that it does not."
> …
> "Is it not that there is something in the soul that bids them drink and a something that forbids, a different something that masters that which bids?"
> "I think so." [1]

[1] 439A-C.

Plato was not a game theorist. And his theory of the soul is, of course, open for different interpretations. Despite this, game theorists have good reason to take seriously the idea, which is at least a *possible* interpretation of what Plato sought to say, that agents are complex entities made up of several distinct subagents. In this interpretation, the man described by Plato faces an inner struggle between a thirsty and a non-thirsty subagent, who have opposing preference orderings. The decision to drink is an outcome of the game played by these two subagents. What the external agent decides to do will depend on the structure of the game defined by the preference orderings of his internal subagents.

The aim of this essay is not to defend any view about what Plato himself thought about the structure of the soul. My modest goal is to show that *if* we accept what I shall call the Platonic Theory of Subagents (and I shall call it "Platonic" irrespective of whether Plato actually endorsed it or not), *then* this has important implications for how we should analyze the Prisoner's Dilemma. In essence, my claim is that given that we accept the Platonic Theory of Subagents, then there is a class of single-shot Prisoner's Dilemmas in which it is rational to cooperate for roughly the same reason as it is rational to cooperate in indefinitely repeated versions of the game. This casts doubt on the orthodox analysis of single-shot Prisoner's Dilemmas, according to which it is never rational to cooperate in such games.

The key idea of the argument I shall defend is that even if you and I will never meet again in future rounds of the game, my subagents play *another* Prisoner's Dilemma that determines what it is rational for me to do in the external game. My subagents play an indefinitely repeated game, and because each subagent can be punished by other subagents in future rounds if "he" does not cooperate, it becomes important for each subagent to adjust the strategy in the current round to the threat of future punishment. I will show that under realistic assumptions, each player's subagents in a two-person internal game benefit from cooperating with the other

subagent, meaning that it is rational to cooperate in many (but not all) external single-shot Prisoner's Dilemmas. A central premise of this argument is that the internal games played by subagents do not change the structure of the external game played by you and me. For my argument to go through, I must be able to show that the external game is a genuine single-shot Prisoner's Dilemma rather than a game in which interactions between subagents lead to shifts in the preference orderings of the external agents.

The idea that we sometimes face internal Prisoner's Dilemmas was first proposed in a modern game-theoretical context by Gregory Kavka.[2] Kavka limits the discussion of what I call the Platonic Theory of Subagents to internal single-shot Prisoner's Dilemmas. He therefore overlooks a key feature of the argument explored here -- namely, the claim that internal and *indefinitely repeated* Prisoner's Dilemmas played by two or more subagents can be part of an external single-shot Prisoner's Dilemma played by you and me, in a sense that makes it rational for us to cooperate. It is this feature of my argument that is novel.

It should be noted that advocates of the Platonic Theory of Subagents need not believe that subagents exist in the same literal sense as tables and chairs. All they have to concede to is that some external agents behave *as if* their behavior is governed by separate subagents. The key premise of the view defended here is thus less controversial than one might think. Moreover, even if we insist on reading Plato's Theory of Subagents literally, there would still be cases in which the theory is relativity uncontroversial. Consider, for instance, games played by collective agents, such as the board of a large organization. The board is an external agent playing games against other collective, external agents, and each board member is an internal subagent of this collective agent. It is not unreasonable to think that

[2] See Sections 1 and 2 in Kavka (1991) and Kavka (1993). Somewhat surprisingly, Kavka does not mention the connection with Plato's tripartite theory of the soul. For a discussing of the psychological aspects of Kavka's proposal, see Frank (1996), who also mentions the connection with connection with Plato's tripartite theory of the soul.

decisions taken by a collective agent can sometimes be determined by internal games played by the subagents of the collective agent.

2. Indefinitely Repeated Internal Prisoner's Dilemmas

Before proceeding, it is helpful to recapitulate some basic facts about the Prisoner's Dilemma. In the generic (external) single-shot two-person Prisoner's Dilemma illustrated in Figure 11.1 each player's outcomes rank as follows: A > B > C > D.[3]

		Col	
		Cooperate	Do not
Row	Cooperate	B, B	D, A
	Do not	A, D	C, C

Figure 11.1 The Prisoner's Dilemma

External single-shot two-person Prisoner's Dilemmas sometimes occur in situations where we might not expect to find them. Consider the following example:

> *The Headlamps Game*
>
> You are driving along an empty highway late at night. Your headlights are in the high beam mode. You suddenly see another car coming towards you. Its headlamps are also in the high beam mode. You and the other driver know that you will never meet again. (Even if you did, you would not recognize the headlights you see as a car you have met before.) Both drivers have to decide simultaneously whether to dip the lights. The four possible outcomes are as follows.

[3] We also assume that B > (A + D) / 2. This ensures that players cannot benefit more from alternating between cooperative and non-cooperative moves in iterated games, compared to mutual cooperation.

(A) You keep high beam and the other driver dips the lights.

(B) Both drivers dip the lights.

(C) Both drivers keep high beam.

(D) You dip the lights and the other driver keeps high beam.

The two drivers face an external single-shot Prisoner's Dilemma. Each driver will be better off by *not* dipping his lights, no matter what the other driver decides to do, so A > B > C > D. Despite this, many drivers around the world do actually dip their lights, even when they meet drivers they know they will never meet again (and even when they know that there is no risk of being caught by the police or of being punished in some other way).

Empirical studies indicate that about fifty percent of laypeople cooperate when confronted with a single-shot, two-player Prisoner's Dilemma.[4] A possible explanation is that various social norms influence our behavior.[5] Social norms summarize an expectation about how other people will behave in a situation, which a rational player will take into account when playing the game. In many cases the rational response to such an expectation is to reconstitute the norm by acting in accordance with it, which in turn makes it rational for others to also take this expectation into account and act in accordance with the norm.

However, even if we can explain why certain norms have emerged, and sometimes face situations in which it is rational to take them into account, a fundamental normative challenge still remains to be addressed: Why should the individual who faces a single-shot Prisoner's Dilemma care about these norms when the mechanisms that support the norm are not in place? In the Headlamps Game, the non-cooperative act dominates the cooperative one, so no matter what

[4] See Camerer (2003).

[5] The literature on this topic is very extensive. For an influential and stringent account, see Bicchieri (2006).

you expect the opponent to do, you will be better off by not cooperating. Why not just ignore whatever norms you normally care about and perform the dominant non-cooperative alternative?

A possible response could be that the norms in question *alter* the agent's preference ordering. But this would entail that the game is no longer a Prisoner's Dilemma, which makes the notion of norms somewhat uninteresting in the present context. If a game thought to be a Prisoner's Dilemma turns out not to have this structure, because the agents' all-things-considered preferences are different from what they appeared to be, this tells us nothing about whether it is rational to cooperate in the single-shot Prisoner's Dilemma.[6] The normative question that is the focus of this chapter remains unanswered: Is it ever rational to cooperate in a single-shot Prisoner's Dilemma?

Another and more sophisticated response is to introduce the concept of *indirect reciprocity*.[7] Even if A and B play the Prisoner's Dilemma only once, B has to take into account that B will play against other players in the future. Therefore, if B does not cooperate with A, and A informs C and D about this, B will quickly gain a bad reputation in the community. Whoever B meets in the future will know that B is not likely to cooperate and will therefore adjust her strategy to this. However, if B instead cooperates with A, and A passes on this information to others, B will gain a good reputation in the community and future opponents will know that B is likely to cooperate. The same applies to all other players: by not cooperating in a one-shot game they are likely to ruin their reputation, which will harm them in the long run. The upshot is that selfish players can benefit more in the long run by cooperating in one-shot games than by not cooperating, because this often makes it more likely that other players (who also seek to protect their reputation) will cooperate with them.

[6] Cf. Binmore's contribution to this volume.

[7] See Nowak and Sigmund (2005) for an overview of the literature on indirect reciprocity and community reinforcement.

So it may very well be rational to cooperate in the one-shot game if one takes into account how one's reputation in the community will be affected.[8]

Although the literature on indirect reciprocity is interesting and helps us to shed light on many social and biological processes, it is worth keeping in mind that indirect reciprocity is not a good reason for cooperating in the Headlamps Game. When you meet other cars in the dark you know that most of them will play the cooperative strategy, but because it is dark they will not be able to identify you and gossip about your non-cooperative behavior. All they will learn when you refuse to cooperate is that there is exactly one non-cooperative player on the roads. This will not affect their behavior. Moreover, if considerations of indirect reciprocity were to actually change the drivers' behavior they would no longer be playing a one-shot Prisoner's Dilemma. As explained above, a game in which other considerations and values are introduced might be interesting for many reasons, but it is no longer a one-shot Prisoner's Dilemma.

Compare the external single-shot version of the Headlamps game with the following repeated version of the same game.

The Headlamps Game on a Small Island

> You live on a small island, on which there is only one other car (which is always driven by the same person). Every time you meet a car in the dark, you know that you are playing against the same opponent. Although the actual number of rounds may well be finite, there is no pre-determined last round of the game, meaning that we cannot solve the game by reasoning backwards.[9] Both drivers must base their

[8] A substantial amount of research into theories of indirect reciprocity have been based on computer simulations. Researchers have tried to determine under exactly what conditions indirect reciprocity is the successful strategy for members of a community who play a series of one-shot games against new opponents in the community. The review article by Nowak and Sigmund (2005) gives a good overview.

[9] If the number of rounds to be played is known at the outset, the backwards induction argument tells us that if the players know that they will play the game *n* times, then they will have no reason to

> decisions about what to do on their expectations about the future. Under reasonable empirical assumptions, each player's expected future gain of cooperating will outweigh the short-term benefit of the non-cooperative move. The upshot is that the two drivers will quickly learn to cooperate, because it is in their own best interest to do so.

The repeated Headlamps game illustrates a general and important insight about the Prisoner's Dilemma: In *indefinitely* repeated versions of the game, in which the probability is sufficiently high in each new round that that round is not the last, each player's expected future gain of cooperating will outweigh the short-term benefits of the non-cooperative move. According to the Folk Theorems for indefinitely repeated games (this is a set of theorems that cover slightly different types of games that were known and discussed by game theorists long before they were formally proven), stable cooperation can arise if players play strategies that are sufficiently reciprocal, i.e. if each player cooperates with a high probability given that the other player did so in the previous round, but otherwise not.[10] Therefore, "in the shadow of the future", we can expect rational players to cooperate in indefinitely repeated Prisoner's Dilemmas.

Let us now put a few things together. First, we know that in an ordinary external single-shot Prisoner's Dilemma the only rationally permissible alternative is to play the non-cooperative strategy, because it dominates the cooperative strategy. Second, we know that in indefinitely repeated versions of the game it will, under some reasonable empirical assumptions, be rational to cooperate. This naturally leads to the question whether a set of alternative strategies can

---

cooperate in the *n*:th round. Therefore, since the players know that they will not cooperate in the *n*:th round, it is irrational to cooperate in round *n - 1*, and so on and so forth, for all rounds up to and including the first round. Some of the technical assumptions of the backwards induction argument are controversial.

[10] For a useful introduction and overview of the Folk Theorems, see Chapter 6 in Hargreaves Heap and Varoufakis (1995). Two of the most well-known examples of reciprocal strategies are grim trigger and tit-for-tat.

simultaneously figure as alternative strategies in an external single-shot Prisoner's Dilemma *and* in another indefinitely repeated internal version of the game?

Enter the Platonic Theory of Subagents. When Anne and Bob play an external single-shot Prisoner's Dilemma, Anne's subagents may at the same time be playing another internal game that determines her decision in the external game. The internal game played by Anne's subagents can be described as an indefinitely repeated game (at least if Anne believes that she is not about to die in the near future). The same is true of Bob and his subagents.

This very rough sketch indicates that there may exist a class of external single-shot Prisoner's Dilemmas in which agents cooperate for what appears to be the same reason as rational players cooperate in some indefinitely repeated versions of the game. The next section examines the details of this argument in greater detail.

3. My Argument

Given that we accept the Platonic Theory of Subagents, we could expect some internal games played by our subagents to be similar in structure to external Prisoner's Dilemmas played by ordinary agents. The Prisoner's Dilemma can thus arise in a place that has often been overlooked by game theorists: inside ourselves. For instance, if you are about to buy a new car several subagents, each with a separate preference ordering, may influence your choice. Each subagent may rank attributes such as safety, fuel economy, and elegance differently.[11] Which car you buy depends on the equilibrium reached in the game played by your subagents, and some of these internal games will be similar in structure to the indefinitely repeated Headlamps Game discussed above.

Although it is interesting in its own right to point out that a Prisoner's Dilemma can arise "within" the agent, this insight is in itself no reason for revising the orthodox analysis of single-shot external Prisoner's Dilemmas. However, as

[11] This example is discussed in detail in Kavka (1991).

explained above, I propose that if we combine this idea with the thought that some internal games are repeated indefinitely, we obtain a powerful argument for thinking that it is sometimes rational to cooperate in an external single-shot Prisoner's Dilemma. (Note that a game is repeated indefinitely many times does not mean that it is repeated infinitely many times. It just means that there is no pre-determined last round; in each round there is a non-zero probability that that round is not the last.[12])

The key premise of my argument is that there is a sense in which single-shot external Prisoner's Dilemmas can be thought of as being parts of other, indefinitely repeated internal Prisoner's Dilemmas. When such single-shot external Prisoner's Dilemmas figure in other indefinitely repeated internal Prisoner's Dilemmas, it is often rational to cooperate. In the shadow of the future, each subagent benefits more from cooperating than from defecting, because the subagents will meet each other again in future rounds of the internal game. In order to explain in greater detail how this argument works, it is helpful to consider an example.

Selfish vs. Altruistic Donations

You are about to decide whether to make a large donation to your Alma Mater. You also have to decide whether to brag to others about the donation. What you eventually decide to do will be determined by the preference orderings of your subagents. To keep the example simple, we assume that only two subagents influence your decision: an egocentric and an altruistic one. Depending on the strategies played by the subagents (who control different parts of the decision-making process) you will end up performing exactly one of the following four acts:

[12] This assumption blocks the backwards-induction argument, which presupposes that the player knows before the last round actually occurs that that round will be the last.

$d \wedge b$	Donate and brag about the donation.
$d \wedge \neg b$	Donate and do not brag about the donation.
$\neg d \wedge \neg b$	Do not donate and do not brag about the donation.
$\neg d \wedge b$	Do not donate and brag about the donation.

Your two subagents rank the four acts differently. In order to facilitate comparisons with the generic Prisoner's Dilemma in Figure 11.1 it is helpful to label the four alternatives with capital letters. In Table 11.2, the four alternatives are listed from the best to the worst, i.e. A > B > C > D.

	Altruistic Subagent	Egocentric Subagent
(A)	$d \wedge \neg b$	$\neg d \wedge b$
(B)	$d \wedge b$	$d \wedge b$
(C)	$\neg d \wedge \neg b$	$\neg d \wedge \neg b$
(D)	$\neg d \wedge b$	$d \wedge \neg b$

Table 11.2

Although it does not matter *why* the subagents have the preferences they have, it is not difficult to construct a story that could explain this. Imagine, for instance, that the reason why the altruistic subagent has the preference ordering listed in Table 11.2 is that she assigns a high weight to the fact that a donation is made but cares less about bragging. The egocentric subagent assigns a lot of weight to the fact that the agent brags, but cares less about the donation. The two subagents thus agree on the ranking of $d \wedge b$ and $\neg d \wedge \neg b$, but they do so for different reasons.

Note that the preference orderings of the two subagents fulfill the conditions of a Prisoner's Dilemma. We can, if we so wish, describe the altruistic subagent as a moral decision maker concerned with the wellbeing of others and the egocentric

subagent as a decision maker concerned primarily with her own prestige and reputation.

Let us imagine that your egocentric and altruistic subagents face the internal game described above many times and that they both believe in every round that there is some non-zero probability that that round is not the last. Both subagents then have to think ahead and take the expected future gain of cooperation into account, and then compare this to the short-term benefit of the non-cooperative strategy. Given some reasonable assumptions about each opponent's strategy in this indefinitely repeated game, it is rational for the two subagents to cooperate, meaning that they will reach a stable Nash equilibrium in which the external agent (you) make a donation and brag about it.

If we were to state the argument proposed here in its most general form, the technical details would inevitably become very complex. There is a risk that we would fail to see the wood for the trees. So in this essay, the best way forward might be to consider the simple case in which both subagents play the very simple *grim trigger strategy*. By definition, a player playing the grim trigger strategy cooperates in the initial round of the game, as well as in every future round, as long as the other player also cooperates. However, as soon as the other subagent defects, a player playing the grim trigger strategy will defect in *all* future rounds of the game, no matter what the other players does. That is, a non-cooperative move will never be forgiven.

Suppose that each subagent knows that the opponent plays the grim trigger strategy. Under what conditions will it then be rational for the subagents to cooperate? Without loss of generality, we can set the utility of the outcome in which both players do not cooperate to zero (because utility is measured on an interval scale). So in Figure 11.3 it holds that C=0 and A> B> 0 > D.

		Col	
		Cooperate	Do not
Row	Cooperate	B, B	D, A
	Do not	A, D	0, 0

Figure 11.3

We know that if both subagents cooperate in every round, and the probability is p in each round that that round is not the last, the expected utility of cooperating is $\mathrm{B} + p\mathrm{B} + p^2\mathrm{B} + p^3\mathrm{B} \ldots = \mathrm{B}/(1-p)$ for each subagent. Moreover, a subagent who does not cooperate in the current round will get A units of utility in this round and 0 in all future rounds (because the opponent plays the grim trigger strategy). By putting these two facts together, it can be easily verified that it is rational for each subagent to cooperate in the current round if and only if $\mathrm{B}/(1-p) \geq \mathrm{A}$, which entails that each subagent will cooperate as along as $p \geq 1 - (\mathrm{B}/\mathrm{A})$. So, for example, if the utility of $\mathrm{d} \wedge \neg\mathrm{b}$ is 2 units and that of $\mathrm{d} \wedge \mathrm{b}$ is 1 unit, it is rational for each subagent to cooperate as longs the probability is at least ½ that the current round is not the last.

If the subagents do not play the grim trigger strategy but play some other reciprocal strategy (such as tit-for-tat with some probability of forgiving non-cooperative moves) then the calculations will be somewhat different. But the general message still holds: rational subagents should cooperate because they run a risk of being punished in the future if they don't. The argument spelled out here is in fact rather insensitive to which particular strategy one thinks a rational agent would play in the indefinitely repeated Prisoner's Dilemma. As explained in Section 2, the Folk Theorems teach us that cooperation will always be rational, given that

the probability is sufficiently high that the current round is not the last and given that all subagents plays sufficiently reciprocal strategies.

The example sketched above is just one of many illustrations of single-shot Prisoner's Dilemmas in which it is rational to cooperate. Consider, for instance, the single-shot Headlamps game discussed earlier. Imagine that each external agent has two internal subagents and that the first subagent is an altruistic subagent whose primary desire is to dip the lights whenever she meets another car. Her second, less weighty desire is to not exceed the speed limit. The second subagent is more selfish. Her primary desire is to drive as fast as possible (that is, exceed the speed limit). Her second, less weighty desire is to not get distracted by dipping the lights. By combing the two choices, dipping the lights (d) or not (¬d), and exceeding the speed limit (e) or not (¬e), we see that the preferences of the two subagents determine which of the four possible alternatives the driver will eventually perform. Consider Table 11.4, which is analogous to Table 11.2.

	Subagent 1	Subagent 2
(A)	$d \wedge \neg e$	$\neg d \wedge e$
(B)	$d \wedge e$	$d \wedge e$
(C)	$\neg d \wedge \neg e$	$\neg d \wedge \neg e$
(D)	$\neg d \wedge e$	$d \wedge \neg e$

Table 11.4

If the game described in Table 11.4 is repeated indefinitely many times, with sufficiently high probability, then the subagents of the first driver will cooperate, i.e. play $d \wedge e$. If we believe that the behavior of the second driver is also determined by the preferences of two (or more) subagents, then those subagents will also cooperate. Hence, both drivers will dip their lights and exceed the speed limit, despite the fact that they play a single-shot Prisoner's Dilemma.

Clearly, the two examples outlined in this section have the same structure. The general recipe for generating further examples is simple: (i) Take a game in which the choice of each external agent is determined by a game played by two internal subagents. (ii) Ensure that for each external player of the single-shot game, there is one altruistic and one selfish subagent. (iii) Then, given reasonable assumptions about the preferences and beliefs of these subagents, the subagents face an indefinitely repeated Prisoner's Dilemma. (iv) Because the selfish and altruistic subagents have reason to believe that they will meet each other in future rounds of the game the two subagents both benefit from cooperating.

4. Not a Prisoner's Dilemma?

At this point it could be objected that the external game played by the drivers in the Headlamps Game is not a genuine Prisoner's Dilemma, because the internal games played by the subagents *change* the preferences of the external agents.[13] As Binmore puts it in his contribution to this volume,

> critics of the orthodox analysis focus not on the game itself, but on the various stories used to introduce the game. They then look for a way to retell the story that makes it rational to cooperate …. If successful, the new story necessarily leads to a new game in which it is indeed a Nash equilibrium for both players to [cooperate]. (Binmore, this volume, p. 34.)

The best response to Binmore's objection is to stress that the structure of the original single-shot game has not changed. The external agents still have the preferences they have, and those preferences constitute a single-shot Prisoner's Dilemma. However, as Gauthier argues out in his contribution to this volume, the preference of what I call the external agent is not always revealed in her choice behavior. The external agent *prefers* not to cooperate in the single-shot Prisoner's Dilemma, but the

[13] I would like to thank Paul Weirich for raising this objection to me.

rational choice is nevertheless to cooperate, because of the structure of the internal game played by the subagents.

In the revised version of the single-shot Headlamps Game discussed in Section 3, the two subagents dip their lights and exceed the speed limit in the internal game. But this does not entail that the agents' preferences in the external game have changed. The external agents still prefer to not dip their lights in the single-shot Headlamps Game. Under the assumption that our choices are outcomes of games played by internal subagents, preferences and choices sometimes come apart. Consider Plato's thirsty man mentioned in the introduction. The thirsty man prefers to drink. But it does not follow that the rational choice for the external agent is to drink, because the rational choice is an outcome of a game played by several internal subagents. What makes it rational for the external agent to refrain from *choosing* according to his preference is the structure of the game played by these internal subagents.

It is worth pointing out that the Platonic account of choice and preference is incompatible with revealed preference theory. According to the Platonic account, the fact that an option is chosen does not entail that it is preferred. There is an extensive and growing literature that questions the truth of revealed preference theory.[14] Revealed preference theory is no longer something that is universally accepted by all game theorists.

5. The Strong and the Weak Interpretation

Let us consider two further objections. First, critics could ask whether it really makes sense to claim that people are governed by altruistic and selfish subagents. What arguments, if any, can be offered for this psychological claim? The second objection has to do with the generalizability of the argument. Even if it is *possible* that situations of the type described here can arise, it remains to determine how

[14] See, for instance, Gauthier's and Hausman's contributions to this volume.

common they are. If it only happens under exceptional circumstances that rational agents governed by two or more subagents have reason to cooperate, it seems that my point is of little general interest.

Let us begin with the first objection. How plausible it is to claim that an external agent's decision to cooperate in the Prisoner's Dilemma is an outcome of a game played by two or more internal subagents? The answer of course depends on how we interpret the Platonic Theory of Subagents. At least two fundamentally different interpretations are possible. I shall refer to these as the *Strong* and the *Weak Interpretations*.

According to the Strong Interpretation, subagents exist in more or less the same sense as tables and chairs. Ordinary people literally have subagents in their heads. I am an agent, and somewhere within me there exist two or more subagents whose decisions jointly determine what I do. There is no fundamental difference between the internal game played by my subagents and, say, a game of chess played by two Grand Masters.

The obvious weakness of the Strong Interpretation is that there seems to be very few controlled scientific studies that support it.[15] The consensus view among contemporary psychologists is that Plato was wrong, at least if we interpret him along the lines suggested here. The Platonic Theory of Subagents is perhaps a useful metaphor that captures something we all tend to experience from time to time. But it is not a claim that can be interpreted literally and supported by scientific evidence.

Let us now consider the Weak Interpretation. According to this interpretation, the Platonic Theory of Subagents is just an analytic tool. We should not believe that people are *actually* made up of separate subagents. The point is rather that people behave *as if* their behavior is guided by internal subagents. Kavka explicitly defends the Weak Interpretation:

---

[15] It is worth pointing out that Kavka (1991) claims that there is at least one scientific study that supports what I call the Platonic Theory of Subagents.

> I suggest we explore some of the implications of treating individual suborderings, desires, criteria, or dimensions of evaluation *as though* they were represented by distinct subagents (within the individual) who seek to achieve satisfaction of the desire (criterion, etc.). I do *not* claim that we are in fact composed of multiple distinct selves, each of which forms an integrated unit over time and has separate dispositions or values from the other selves of the same individual. (Kavka 1991: 148)

The main attraction of the Weak Interpretation is that it is less demanding than the strong one from a metaphysical point of view. No matter whether we have reason to think that subagents really exist, we could nevertheless reason *as if* they exist. Anyone who is familiar with Bayesian decision theory will recognize this argumentative strategy:[16] Bayesian decision theory is a theory of rational choice for individuals confronted with a set of uncertain prospects. According to this theory, preferences over uncertain prospects must fulfill certain structural conditions in order to be rational. Given that these structural constrains are met, it can be shown that a rational agent behaves *as if* his or her choices were governed by the principle of maximizing subjective expected utility.

In order to assess the plausibility of the Weak Interpretation it is helpful to discuss the analogy with Bayesian decision theory a bit further. I believe there are important similarities, but also a crucial dissimilarity.

Note that Bayesian decision theorists do not claim that people actually have utilities and subjective probabilities in their heads. Utilities and subjective probabilities are abstract entities constructed for predicting and explaining people's behavior when confronted with a certain type of decisions. The as-if clause in the Weak Interpretation of the Platonic theory is meant to function in the same way as the corresponding clause in Bayesian decision theory. Advocates of the Weak Interpretation do not ask us to believe that people have subagents in their heads.

[16] See e.g. Ramsey (1926), Savage (1954) and Jeffrey (1983).

On the contrary, subagents are fictional entities we use for rationalizing people's behavior.

The main advantage of the Weak Interpretation is that it is less demanding from a metaphysical point of view than the Strong Interpretation. However, it is not clear, at least not at this stage, that the Weak Interpretation can really support the claim that it is rational to cooperate in single-shot Prisoner's Dilemmas. In order to see this, it is important to distinguish between descriptive and normative applications of game and decision theory.

First consider descriptive interpretations of Bayesian decision theory. Although one can of course question whether people actually behave in accordance with the axioms of the theory (there is plenty of empirical evidence to the contrary) the as-if clause is rather unproblematic in descriptive interpretations.[17] No matter whether people actually have subjective probabilities and utilities in their heads, it can be helpful from a predictive as well as from an explanatory point of view to ascribe such entities to agents. As long as the decision theorist is able to elicit accurate predictions and fruitful explanations there is little to worry about.

However, when we turn to normative interpretations of the Bayesian theory it is far from evident that the as-if clause gives the normative decision theorist what she wants. The problem is that normative versions of Bayesian decision theory put the cart before the horse, meaning that the theory is not able to offer the agent sufficient action guidance. F. P. Ramsey, the inventor of Bayesian decision theory, was aware of this problem. In a brief note written two years after "Truth and Probability", Ramsey remarks that:

> sometimes the [probability] number is used itself in making a practical decision. How? I want to say in accordance with the law of mathematical expectation; but I

[17] For a brief summary of the descriptive literature, see Peterson (2009:Ch. ??).

> cannot do this, for we could only use that rule if we had measured goods and bads. (Ramsey 1928/1931: 256)

Let us try to unpack Ramsey's point. The reason why he cannot advise a rational agent to apply the principle of maximizing expected utility for making decisions is that the agent's preferences over the set of available alternatives is used for *defining* the notion of utility (and subjective probability). The "goods" and "bads" are derived from the agent's preferences over all available prospects and this preference order is supposed to be complete. Hence, Ramsey's rational agent is supposed to know already from the beginning whether he prefers one alternative prospect (that is, one good or bad) to another, meaning that for the ideal agent no action guiding information can be elicited from the theory. The completeness axiom entails that the rational agent already knows what to do, since the set of entities the agent is supposed to have complete preferences over includes all the acts she can choose to perform. This is why Ramsey writes that he cannot say what he wants to say.

For non-ideal agents, whose preferences are either incomplete or violate at least one of the other preference axioms, it is of course true that Ramsey's theory (as well as the theories proposed by other Bayesians) can be action guiding in an indirect sense. If a non-ideal agent discovers that her current preferences are incomplete, or violate some of the other axioms, then the agent should revise her preferences such that all the axioms are fulfilled. However, the theory tells us nothing about *how* the preferences should be revised. Any revision that yields a set of preferences that obeys the axioms will do.

Let me try to relate these insights about as-if reasoning in Bayesian decision theory to the Prisoner's Dilemma. First note that as long as we use game theory for descriptive purposes, such as for explaining and predicting how people react when confronted with the Prisoner's Dilemma, it seems to be no more problematic to refer to hypothetical subagents than to hypothetical utility and probability functions. Given that the predictions and explanations are accurate, it is legitimate to claim

that people behave *as if* their decisions are governed by something that might in the end not exist.

However, when we turn to normative issues in game theory, it is yet unclear whether it is sufficient to rely on as-if reasoning. What is the normative relevance of the fact that we can reason as if our decisions were governed by subagents if we suspect that no such subagents exist? To put it briefly, the worry is that the normative relevance of as-if reasoning in game theory is no greater than the normative relevance of as-if reasoning in Bayesian decision theory.

In order to steer clear of this objection, it is important to observe that there is in fact a crucial difference between as-if reasoning in game theory and Bayesian decision theory. In the analysis of the single-shot Prisoner's Dilemma proposed here, the key idea is that ordinary agents like you and me should reason as if our decisions were governed by a set of subagents playing an indefinitely repeated Prisoner's Dilemma. The reason why we should reason as if such subagents determine our choices is that this is a good *descriptive* account of how humans function. It is not a normative claim about what makes a set of preferences rational, or about how non-ideal agents should revise their preferences. The point is merely that we should take into account how agents function when we formulate normative principles for how a rational agent should reason when confronted with the Prisoner's Dilemma.

As noted earlier, about fifty percent of ordinary people without training in game theory cooperate with their opponent when confronted with the single-shot Prisoner's Dilemma.[18] By reasoning *as if* our decisions were determined by games played by subagents, we can explain why this is so. Hence, the theory of subagents is not unreasonable if interpreted as a descriptive as-if claim.

Once we believe that the as-if account of ourselves is descriptively accurate, this makes it plausible to defend the following normative claim: If an agent who

[18] See Section 2 and Camerer (2003).

plays the Prisoner's Dilemma can be accurately described as someone who believes that her subagents will start to punish each other if she does not cooperate now (that is, if the agent can be described *as if* this was the case), then the agent should take this insight into account when deciding what to do in a single-shot Prisoner's Dilemma. Therefore, the Weak Interpretation is sufficiently strong for warranting the claim that it is rational to cooperate in single-shot Prisoner's Dilemmas.

The upshot of all this is that the Weak Interpretation is more attractive than the strong one. We should reason *as if* the Platonic Theory of Subagents is correct, without claiming that such subagents actually exist. A similar argumentative strategy is widely accepted in Bayesian decision theory, where it works well for descriptive purposes. (All I claim is that this argumentative strategy is coherent from a conceptual and philosophical point of view. The fact that it may very well be an inaccurate descriptive account is irrelevant in the present context.) The as-if approach to normative decision theory is, however, of limited value since it offer us no or little action guidance. The fully rational Bayesian agent already knows what to do. And non-ideal agents can merely use the theory for revising their preferences such that they become compatible with the structural axioms of rational preferences proposed by Bayesians.

6. Possible Generalizations

As noted in the introduction, there is at least one type of game in which the claim that external agents consists of internal subagents is uncontroversial irrespective of how it is interpreted, i.e. in which it is unproblematic to claim that internal subagents play games that determine what external agents do. The games I have in mind are games played by collective agents, such as firms and other organizations. For such games, even the Strong Interpretation discussed in Section 4 seems applicable. Imagine, for instance, that the board of a large company is thinking of doing business with a foreign company they have never worked with in the past. If the two companies manage to negotiate a deal, they will only do business for a short

period of time and they will never interact again in the future. Suppose that the game played by the two companies is (for reasons we need not worry about here) a single-shot Prisoner's Dilemma. Would it be rational for the two boards to cooperate?

If we generalize the Platonic Theory of Subagents from ordinary humans to collective agents (as Plato famously did himself in the *Republic*) we see that it may be rational for the external players to cooperate in this single-shot Prisoner's Dilemma. The board of each company can be viewed as a collective external agent comprising several internal subagents. In each board meeting the members of the board know that there is a sufficiently high probability that that board meeting is not the last. Therefore, it is rational for them to take into account how other board members will behave in the future. The upshot is that each board member has to consider how other board members would react if he or she were to play a non-cooperative strategy.

It is not crucial to the argument to determine precisely what range of real-life cases can be covered by the Strong Interpretation. As emphasized in Section 4, the Weak Interpretation is sufficiently strong for justifying the normative claim about single-shot Prisoner's Dilemmas that is central to this essay. So the strength of the argument defended here primarily depends on how large the range of cases is that is covered by the Weak Interpretation.

Unsurprisingly, it is not difficult to identify at least *some* single-shot Prisoner's Dilemmas to which the argument for cooperation does not apply. Imagine, for instance, that Alice and Bob play a single-shot Prisoner's Dilemma and that they know that this is the last decision they will make before they die. Once they have decided what to do in the current round, a pandemic disease will kill them, which they know. In this situation it would clearly not be rational for Alice and Bob to cooperate, simply because their subagents know that they will never get an opportunity to play against their opponents again in the future.

The insight that the argument for playing cooperative strategies merely applies to *some* single-shot Prisoner's Dilemmas might make us doubt the scope of the argument. Perhaps the fraction of single-shot Prisoner's Dilemmas to which the argument applies is very small or even minuscule? My response to this worry consists of two points. First, it is arguably interesting in its own right to point out that there is a least *some* single-shot Prisoner's Dilemmas in which it is rational to cooperate. The consensus view in the literature is that there is *no* such game in which it is rational to cooperate.

My second point is less defensive. Here is what I have to say: Although it is true that not every external single-shot Prisoner's Dilemma can be reconstructed as some game in which the agents play another set of internal Prisoner's Dilemmas guided by the preference orderings of their subagents, this does not entail that the argument outlined here fails. On the contrary, the key idea of the argument can be applied to a much broader set of games. In order to see this, note that it can be rational for the external agents playing the single-shot Prisoner's Dilemma to cooperate even if the internal games played by the subagents are not Prisoner's Dilemmas. All that matter for the argument to go through is that the internal games are ones in which it is rational for the subagents to cooperate. Needless to say, there are numerous other games, iterated as well as non-iterated ones, in which it is rational for subagents to cooperate. The key questions that determine the scope of the argument are thus the following: (i) Can every external single-shot Prisoner's Dilemma be reconstructed as a game played by two or more internal subagents? (ii) If so, is it rational for the subagents to cooperate in those internal games?

As we have seen above, the answer to the second question varies from case to case. There is a wide range of cases in which it is rational for two or more subagents to cooperate, but this is not always the case. There are exceptions.

But what about the first question, can every single-shot external Prisoner's Dilemma really be reconstructed as a game played by two or more subagents? The answer is yes. From a technical point of view, it is trivial that *every* game can be

reconstructed along the lines proposed here. A game is essentially a list of preferences orderings expressed by agents over a set of possible outcomes. It takes little reflection to see that it is always possible to construct *some* game, which we may refer to as an internal game, that has the alternative acts corresponding to those preference orderings as outcomes.

A final worry about the argument arises even if we think it is indeed true that some single-shot external Prisoner's Dilemmas can be represented as a game in which the external agent's behavior is governed by a set of internal games played by subagents. The worry is that nothing seems to guarantee that this representation is unique. There might very well be other ways of representing the game, according to which it would not be rational to cooperate. What should a rational agent do if it is rational to cooperate according to some representation of the game, but not according to others? The best response is to point out that although this is an issue that is interesting and worth discussing, this is also a very general objection that takes us far beyond the scope of the present essay. Very broadly speaking, nothing seems to exclude that in some situations one and the same set of facts about a situation can be described equally well by different games. This insight is not unique for the current discussion. In fact, game theorists have had surprisingly little to say about how we should determine which game someone is actually playing. It is one thing to view a game as a formal object and study its technical properties, but a completely different task to determine how well such a formalization represents the real-world phenomena we are ultimately interested in.

If the formalization I propose is plausible, this may have important implications for how we should address single-shot Prisoner's Dilemmas in society. The good news is that we would then actually benefit more from cooperating with each other than what has been previously recognized.