Reason and Pareto-Optimization

Introduction

The one-shot Prisoner's Dilemma has come to be the paradigmatic representation of the problem of cooperation. The Prisoner's Dilemma reveals an apparent tension between the conception of rationality prevalent in economics, according to which rational action aims at maximizing individual expected utility, and cooperation, which requires a constraint on individual utility-maximization. Thus, one can act rationally, or one can cooperate, but one cannot do both. For twenty-five years now and counting, David Gauthier has defended the rationality of conditional cooperation in one-shot Prisoner's Dilemmas. His best-known defense of this is found in his *Morals by Agreement* (1986), where he defends the rationality of adopting the disposition of 'constrained maximization'. In two recent papers (Gauthier 2013a, 2013b), Gauthier has recast his view. According to his most recent view, rationality sometimes requires that agents adopt a Pareto-optimizing perspective on rational choice instead of best reply reasoning characteristic of the classic maximizing conception of reason.

In this paper I argue that Gauthier's recommendation to Pareto-optimize can be interpreted in one of two ways. In the first, the directive to Pareto-optimize is issued to group agents. This interpretation collapses his view to a variant of team reasoning, a theory of rational choice, notably endorsed by Michael Bacharach (1999, 2006) and Robert Sugden (1993, 2003, 2014), according to which individuals reason as members of teams (or groups) and act on the basis of what is best for the team to which they belong. This alternative leaves us without a reason for an individual *qua* individual agent to cooperate in Prisoner's Dilemmas, and leaves Gauthier very far away from the original ambitions driving his project.

In the second interpretation, the injunction to Pareto-optimize is issued to individual agents, but cannot explain why defecting in Prisoner's Dilemmas is not rational.

The structure of this paper is as follows. I begin by introducing Gauthier's project of reconciling reason with cooperation in the Prisoner's Dilemma, and outline his argument for Pareto-optimization. I then (sections 2 and 3) turn to examine the two ways of understanding the recommendation to Pareto-optimize. On the first understanding (section 2), Pareto-optimization collapses into a variant of team reasoning, and the recommendation to Pareto-optimize is a prescription that is issued to a group of individuals. On the second understanding (section 3) the prescription to Pareto-optimize is issued to individuals *qua* individuals to arrive at Pareto-optimal outcomes. I argue that both alternatives are problematic for Gauthier. I argue further (section 4) that Gauthier's account faces a more fundamental problem that I call (following Michael Bratman) 'the fragmentation of practical reason' (Bratman 2013: 664). I conclude that Gauthier does not succeed in reconciling reason with cooperation in the Prisoner's Dilemma.

1. The Argument for Pareto-Optimization

1.1. The Prisoner's Dilemma

Our starting point is the Prisoner's Dilemma. The choreography of this dilemma is so familiar that I shall not rehearse it at length here. The matrix is as follows.

Table 1: ***The Prisoner's Dilemma***

		Player B	
		Cooperate	Defect
Player A	Cooperate	3,3	1,4
	Defect	4,1	2,2

On the orthodox conception of rationality—the one prevalent in economics and game theory and that I shall call the 'maximizing conception of reason'—a rational player chooses the act that will maximize her expected utility, given her expectation about how her opponent will choose. Her actions are best replies to the actions she expects her opponent to choose. In the Prisoner's Dilemma, the best reply to any expected strategy employed by one's opponent is always to defect. In the language of game theory, 'defect' is each player's *dominant* strategy and 'cooperate' her *dominated* strategy. A strategy is dominant if it is a player's best reply to whatever strategy her opponent chooses, and is dominated if there is always a strategy that yields higher expected utility. When both players act rationally, the resulting outcome—in this case, the *equilibrium* of the game—is one of mutual defection. A set of strategies is in equilibrium if no player has an incentive to change strategies, given the strategy employed by the other player.

Mutual defection may be the equilibrium of the game, but it is sub-optimal. Mutual cooperation leaves both players better off than they are by mutually defecting. In the Prisoner's Dilemma, mutual cooperation is what is known as a *Pareto-optimal outcome*, which describes a set of strategies where no player can do better without worsening the lot of the other player. In other words, the mutual defect outcome—the outcome that the maximizing conception of rationality requires that individuals arrive at—is sub-optimal. Gauthier finds this result problematic and seeks to show that cooperation in the one-shot Prisoner's Dilemma is indeed sometimes rational.

1.2. Gauthier's Reconciliation Project

Gauthier's best-known defense of the rationality of cooperation in the one-shot Prisoner's Dilemma is in his *Morals by Agreement* (Gauthier 1986). There he sets as his aim to show that compliance with the dictates of morality (modeled as the cooperative outcome in the Prisoner's Dilemma) is consistent with rational choice, and thus to provide an answer to the question, most famously raised by Hobbes's Foole, 'Why should I be moral?' According to Gauthier, cooperators are sometimes able to achieve benefits unavailable to non-cooperators and, thus, reason can recommend in those cases that individuals cooperate. He argues that, on the assumption that dispositions can be reliably detected by others (a condition he calls 'translucency'), individuals who adopt the disposition to conditionally cooperate—those he calls 'constrained maximizers'—do better than those he refers to as 'straightforward maximizers', who act in accordance with the maximizing conception of reason and who defect in Prisoner's Dilemmas when they can do so with impunity. In short, Gauthier argues that it is advantageous (and thus rational) for individuals to form a disposition to cooperate, and that actions that flow from a disposition that it is rational to have are themselves rational.

This view has encountered two serious criticisms. The first focuses on the translucency assumption and the extent to which individual dispositions are detectable by others (Smith 1991).[1] Without a reliable detection mechanism, correlation between strategies is diminished, and the benefits of cooperation no longer clearly exceed those of defection, since the possibility for exploitation of cooperators by defectors rises. If so, adopting a cooperative disposition is not rational because it is not advantageous. The second questions the rationality of actions that flow from rational dispositions (Parfit 2001, Thompson 2001). The basic complaint is that while it might be rational for an individual (because it is advantageous) to form a disposition to act in certain ways, it is not always rational for an individual (because it is not always advantageous) to perform actions that are recommended by that disposition. Thus, even if Gauthier could establish that individuals are sufficiently translucent to make adopting the disposition to cooperate advantageous, it does not follow that actions recommended by that disposition will be rational. These criticisms do much to undermine the success his project of showing the rationality of cooperation in Prisoner's Dilemmas, and in two recent papers (Gauthier 2013a, 2013b), Gauthier presents us with another way of doing so. This is to appeal to Pareto-optimality, and my concern in what follows will be with Gauthier's 2013 position. In these papers, Gauthier drops explicit reference to translucency and updates some of his earlier terminology. 'Constrained maximization' becomes 'Pareto-optimization', and those who follow its prescriptions, 'agreed Pareto-optimizers' (Gauthier 2013a: 609). 'Straightforward maximizers' are referred to as 'maximizers' or 'those who employ best reply strategies'. He also explicitly dissociates 'rational cooperation' from 'maximization' (Gauthier 2013a: 608), claiming that preserving that connection is misleading. But it is not obvious how much of substance has changed since

[1] For an extended discussion of Gauthier's *Morals by Agreement*, see Vallentyne 1991.

1986 on the issue of rational cooperation or whether Gauthier is presenting us with a mere revision in nomenclature. Ultimately, this will not matter, for I will argue that whether we interpret Gauthier's 2013 work as introducing a change in terminology or one of substance, his view is problematic.

1.3. Two Perspectives on Rationality

On Gauthier's 2013 view, rationality should sometimes require that agents adopt a Pareto-optimizing perspective on rational choice in lieu of employing best reply reasoning characteristic of the maximizing perspective. According to Gauthier, (2013a: 606), the Prisoner's Dilemma reveals the presence of two distinct conceptions of rationality: one that leads to equilibrium outcomes; the other, to Pareto-optimal outcomes. Gauthier's aim is to reject the former in favour of the latter. Indeed, he wants to remove what he refers to as 'the stranglehold' that the maximizing perspective has on rational choice. He says, 'I want to free our minds from the dogma that individual actions are rational only if maximizing, while keeping hold of the deeper idea that rational agents seek to bring about what they most value' (Gauthier 2013b: 195-196).

For Gauthier, the fact that a maximizing perspective leaves an agent worse off in choice situations like the Prisoner's Dilemma than she would be by following the prescriptions of another (non-maximizing) perspective is reason to reject that perspective. Since the maximizing perspective on rationality bars agents from reaching cooperative outcomes in Prisoner's Dilemmas, and since mutual cooperation yields a higher payoff to each individual than does mutual defection, Gauthier thinks that we should give up the view that all rational action should be maximizing. The basic idea is that best reply strategies do not always generate the most utility-maximizing outcomes for agents. This raises doubts about the normative adequacy of a conception of reason whose prescriptions do not guarantee the sought outcome.[2]

According to Gauthier, there is another perspective available in the Prisoner's Dilemma. This is the Pareto-optimizing perspective, which aims at an outcome where no player can do better without worsening the lot of the other player. In the Prisoner's Dilemma,

[2] Gauthier (2013b: 201-203) uses the Centipede puzzle to make the same point. In that game, two players play a game with 100 possible moves for a share of a pot that increases with each move. They take turns and have the option of either terminating the game at each move *n* or passing the move along to the other player. If a player terminates instead of passing the move along, she wins a greater share of the pot—in Gauthier's example, $10(*n* + 1)—than the player to whom the move is passed, who receives $10(n – 1). At each move *n*, a player does better by terminating than by passing the move along to the other player. The result is that the maximizing conception of reason prescribes to agents to terminate the game on the first round, leaving the first actor with $10 and the other player with $0 (versus $1100 and $1000 respectively had they played to the end). However, as Gauthier says, 'A theory of rational choice that has this as a result is plainly unsound' (2013b: 201). Gauthier's general strategy here and in the context of the Prisoner's Dilemma is to provide a *reductio ad adsurdum* argument against the maximizing conception of reason by showing that it leads to results that no theory of rationality should lead to.

‘cooperate/cooperate’ is a Pareto-optimal outcome.[3] That outcome also affords to agents a higher utility than does the equilibrium outcome (viz., ‘defect/defect’). Gauthier takes this fact as sufficient justification to adopt a Pareto-optimizing perspective on rational choice. According to him, given that in Prisoner’s Dilemmas Pareto-optimal outcomes afford to each agent higher utility than do equilibrium outcomes, rational action should aim at securing Pareto-optimal outcomes rather than equilibrium outcomes where there is a conflict between the two. If agents were able to break free from best reply reasoning, they could attain the cooperative benefits when they interact with like-minded individuals. Thus, Gauthier’s recommendation is to Pareto-optimize rather than maximize. The underlying presumption is that reasoning aimed at Pareto-optimality rather than best reply reasoning is rational because it contributes to bringing about what an agent most values—or, to borrow from Gauthier

[3] Notice that the ‘cooperate/defect’ and ‘defect/cooperate’ outcomes are also Pareto-optimal: given that the one who chooses ‘defect’ gets her best possible payoff, any improvement to the lot of the cooperator will make the defector worse off. There can therefore be no Pareto-improvement in those cases. Pareto-optimality thus isn’t a sufficient condition for the rationality of cooperation. Gauthier recognizes this. His central complaint is rather that the ‘defect/defect’ outcome that we are led to by maximization is sub-optimal, and suggests resolving this by placing the Pareto-condition on rational choice. He says: ‘a set of actions, one for each agent, is fully rational only if it yields a Pareto-optimal outcome’ (2013a: 607). So that rules out ‘defect/defect’ but doesn’t distinguish ‘cooperate/defect’, ‘defect/cooperate’, or ‘cooperate/cooperate’.

(1994), make one's life go as well as possible. Gauthier's argument can be reconstructed as follows:

> (P1) Rational action aims at making one's life go best.
>
> (P2) Pareto-optimization makes agents' lives go better in Prisoner's Dilemmas than maximization does.
>
> (C) Therefore, rational agents should Pareto-optimize rather than maximize.

We thus arrive at an argument for a Pareto-optimizing theory of rational choice, and a justification to cooperate in Prisoner's Dilemmas. In Prisoner's Dilemmas where best reply reasoning leads to sub-optimal equilibrium outcomes, individuals should aim to achieve Pareto-optimality. And since in the Prisoner's Dilemma achieving Pareto-optimality requires cooperation, rational individuals should cooperate.

In what follows I will show that 'agents' in Premise 2 of the above argument can be understood as referring either to group or to individual agents. I will argue that, understood as group agents, the conclusion—that rational agents should Pareto-optimize rather than maximize in Prisoner's Dilemmas—follows, but at the cost of abandoning what we might think are essential elements of Gauthier's project. I will argue further that if 'agents' is understood as individual agents the conclusion does not follow, since Pareto-optimization does not always make an individual agent's life go best. Thus, whether Pareto-optimization is interpreted as an injunction to group or individual agents, it does not follow that cooperation in Prisoner's Dilemmas is rational.

2. David Gauthier, Team Reasoner

2.1. Me as One of 'Us': Group Agency and Team Reasoning

One way of understanding the prescription to Pareto-optimize is as a prescription that is issued to group agents. Gauthier specifies that, 'Instead of supposing that an action is rational only if it maximizes the agent's payoff given the action of others, I am proposing that a set of actions, one for each agent, is fully rational only if it yields a Pareto-optimal outcome' (Gauthier 2013a: 606-607). According to him, the Pareto-optimizing theory prescribes 'a single set of directives to all interacting agents, with the directive to each premised on the acceptance by the others of the directives to them' (Gauthier 2013a: 607). Gauthier says that

> …on a Pareto-optimizing account a single directive is issued to all those interacting—to us, as it were, and to me only as one of us. Considerations that count as reasons for me as one of us would not count as reasons for me if I may not assume the 'us' (Gauthier 2013a: 608-609).

Gauthier seems here to be distinguishing reasons that I have as an individual agent from reasons that I have when there is an 'us'. This can be understood as employing features of team reasoning. Team reasoning is a fundamentally different form of reasoning from the individualistic and maximizing conception of reason. The central feature of team reasoning is that it shifts the unit of agency from *individuals* to *groups*. Those who are engaged in team reasoning ask 'What should we do?' rather than 'What should I do?' Team reasoning thus presupposes the existence of a team (or group). An individual who has identified with the team will consider which combination of actions done by the team's members will best achieve the goal of the team, and then will carry out her component part. Members of the team act rationally insofar as they perform their component parts to achieving the outcome that is best for the team with which they identify. Thus the basic structure of team reasoning, from the perspective of team members, is:

(1) We identify as members of Team A.

(2) Outcome BB is the outcome that is best for Team A.

(3) We should each, as members of Team A, perform our component parts B in bringing about BB.

(4) Therefore, I, as a member of Team A, should B.[4]

[4] This schema is adapted from Gold and Sugden (2007: 289), and simplified to make the basic structure of the reasoning clear. Left out of the above schema are assumptions about the common knowledge of team identification and the uniqueness of outcome BB as the outcome that is best for the team. Theories of team reasoning will differ in how to answer what makes individuals team identify, how to achieve common knowledge about team identity, and whether it is rational to reason as a team member under uncertainty that others have team identified. A discussion of these divergences is beyond the scope of my aims here. For an

We can see how cooperation can be rational once individuals shift from asking 'What should I do?' to 'What should we do?'. When my aim is to maximize what is best for me, no matter what the other player does, my best response is always to defect. But when my aim is to bring about an outcome that is best for the team, I should cooperate. We might spell out team reasoning in the Prisoner's Dilemma as follows:

(1) We identify as a members of Team A.

(2) Outcome 'cooperate/cooperate' (CC) is best for the team.

(3) We should each, as members of A, perform our component parts C to bring about CC.

(4) Therefore, I, as a member of A, should cooperate.[5]

overview of the central difference between theories on these points, see Gold and Sugden (2007: 294-308). On the differences in Bacharach and Sugden's views on team reasoning under uncertainty, see Gold (2012). Let us assume that in this simplified case there is common knowledge among members of A that all members have team identified. Let us also assume that outcome BB is indeed the outcome that is best for the team. With these assumptions in place, on the basis of what is best for the team with which she identifies, an individual team member reasons to perform her component part to achieving that aim.

[5] The above is a sample schema only showing how a team reasoner might arrive at the cooperative outcome in the Prisoner's Dilemma situation, with payoffs as described in the matrix outlined above. It should be noted that these payoffs can be altered in such a way that a team might turn out to be indifferent between various outcomes. That is not important for our

Thus, according to team reasoning, cooperation is rational for team members in virtue of it being a component in the outcome that is best for the team with which the individual identifies.

2.2. Team Reasoning and Pareto-Optimization

There are remarkable similarities between team reasoning and Gauthier's Pareto-optimizing theory of rational choice. Recall that Gauthier characterizes the directive issued by Pareto-optimization as one issued to 'us…and to me only as one of us' (Gauthier 2013a: 608-609). On this characterization, an agent's reason for acting derives from what she as an individual who is part of an 'us' can do to bring about what is best for 'us'. Thus, insofar as there is an 'us' and there is a profile of actions that is best for 'us', an agent can derive her reasons for action on the basis of that action being a component in bringing about the outcome that is best for us. Just as an individual has a reason to act on the basis of it being a component part in achieving what is prescribed to the team in team reasoning, so too is an individual's reason for action derived from a prescription made to the 'us' to which she belongs.

We can thus reconstruct the reasoning process required by the Pareto-optimizing theory in the Prisoner's Dilemma as:

(1) We are both members of 'us'.

purposes here, which is to show how we might (at least sometimes) render cooperation compatible with rationality in the Prisoner's Dilemma.

(2) The Pareto-optimal outcome is best for ‘us’.

(3) We should each, as members of ‘us’, do our component parts to bring about the Pareto-optimal outcome.

(4) Therefore, I, as a member of ‘us’, should cooperate.

The above schema reveals the structural identity between the reasoning performed by Gauthier’s Pareto-optimizers and that performed by team reasoners. Just as the team reasoner arrives at the conclusion that she should cooperate on the basis of what is good for the team with which she identifies, here we arrive at the conclusion that we should each cooperate on the basis of what is good for ‘us’. It is better for us as a team to cooperate than it is for us as a team to defect. Our reasons as individual members of ‘us’ to cooperate are dependent on the directive to arrive at the Pareto-optimal outcome—or, the C/C outcome—that is issued to the team with which we identify.

2.3. Problems for Gauthier

We might at this point ask why this does not solve Gauthier’s problem of reconciling rationality with cooperation. And here our answer will depend on what we take Gauthier’s aim to be. On the one hand, we may take Gauthier’s aim to be to advance an argument along the continuum of that advanced in *Morals by Agreement* that answers ‘Why should I be moral?’ In that project, Gauthier sought to show that it was rational for an individual to cooperate in Prisoner’s Dilemmas, where demonstrating that required showing that it was individually advantageous to cooperate (or, more accurately, to form a disposition to cooperate). However, what we have thus far considered is Gauthier’s argument for why it is rational for individual members of a team to cooperate. As we have seen, the claim is that it is rational for them to do so because doing so is a component part in generating an outcome that is better for the team than outcomes generated by any alternative profile of actions. This presupposes that an individual has already identified as a member of the team. Thus, while we may have an account of why it would be rational for a team member to cooperate in a Prisoner's Dilemma, we do not yet have an account of whether that agent's initial identification with the team is rational. Without such an account, Gauthier is unable to provide an answer to the necessary question, 'Why should *I* cooperate?'. This will leave anyone who was initially struck by the ambitions of Gauthier's *Morals by Agreement* project unsatisfied.

On the other hand, we might take Gauthier's latest ambition to simply be to defend why it would be rational for a member of a team to cooperate in the Prisoner's Dilemma. After all, Gauthier explicitly recognizes the 'inadequacy of deliberating as solitary persons' (2014b: 203) and advances reasons for, in his words, 'us' to cooperate. And his answer then is: because it's better for you, a member of the team that forms an 'us'. But if so, there is more work to be done. As Sugden points out, 'Any theory of team reasoning needs to explain which sets of individuals, under which circumstances, come to perceive themselves as teams'. Importantly, Sugden thinks the unit of agency must be specified before evaluations of rationality can be made. In his words, 'an action is rational for an agent to the extent that it can be expected to achieve that agent's objectives. Thus, the question 'Who am I?' (or 'Who are we?') is logically prior to rational choice' (Sugden 2015: 153).

How individuals come to conceive of themselves as members of a team remains unsettled in the literature.[6] Neither Sugden nor Bacharach thinks that the unit of agency is something that can be chosen. According to Bacharach, whether an individual identifies with the team is a matter of psychological framing and not something that is subject to choice and, thus, not something that can be evaluated as rational. Sugden also denies that we can speak of the rationality of team identification; he thinks that the scope of rational choice is limited to the particular unit of agency adopted, and does not apply in the case of team reasoning until after team identification has taken place. Susan Hurley (1989), on the other hand, thinks that the unit of agency can be chosen. However, she grounds the rationality of choosing between alternative units of agency in impartial considerations and agent-neutral goals, which are at odds with the rationality assumptions that drive Gauthier's project.

In contrast to all of these, Gauthier seems to think that the unit of agency is something that can be chosen, and that rational individuals will make that choice on the basis of the benefits that are accrued to them. He says, 'If it is beneficial for them [individuals facing a Prisoner's Dilemma] to join together and cooperate, then this is what, insofar as they are rational, they will do' (2013: 607). And, indeed, perhaps Gauthier could satisfy both the moral skeptic and the conditions set out by Sugden by showing that it would be advantageous (and thus rational) for an individual to become a member of the team. On this alternative, one could give a reason why it would be rational for an individual team member to cooperate in the Prisoner's Dilemma (because doing so constituted her component part in bringing about the outcome that is best for the team) and why it would be rational for an individual to

[6] For a comparison of the different ways team identification comes about, see Gold and Sugden (2007: 294-304).

become a team member. Presumably that answer would appeal to the benefits that could only be achieved by participating as a member of the team. But an argument of that sort would have to also explain why it would not be rational to feign becoming a team member and defecting while all others cooperate. And it is not obvious that such an answer can be found, since it seems that wherever an agent has reason to believe that others are going to team reason (and thus has a reason to believe that she can benefit from the cooperative endeavour), she too will have reason to defect.[7]

3. Individual Agency and Pareto-Optimization

3.1. What's In It For Me?

We have seen that it is better for *team members* to Pareto-optimize instead of maximize because doing so permits the team to achieve mutual cooperation, which is clearly better for

[7] There may be situations where it would be impossible to get the benefits of cooperation unless one becomes a team member. Those situations would be ones where it would be either impossible to feign team membership or impossible revert back to individual reasoning once one has joined the team. For either of these scenarios to obtain, we would need to assume that intentions are detectable so that an individual who lacked an intention to remain a team member would be excluded from a cooperative endeavour, or that individuals can lock themselves into team reasoning when they become members of the team. Not only are these assumptions questionable, but such a situation would severely restrict the conditions under which cooperation would be rational.

the team than mutual defection. I have argued that as an injunction issued to groups, Pareto-optimization gives a reason for team members to cooperate in Prisoner's Dilemmas but not necessarily a reason for individuals to become team members. I have argued further that it is not clear that such a reason can be found, since an individual will do best by defecting whenever she can do so with impunity.

I now turn to the second way of interpreting Gauthier's argument for Pareto-optimization. This is as a prescription that is issued to individual agents and not to the team to which they belong or to them as team members. As Gauthier says, 'if cooperation on agreed terms is to be had, then a rational agent will optimize; only if cooperation is not to be had will he maximize' (Gauthier 2013a: 609). This suggests that Gauthier has in mind that an agent can derive a reason to cooperate on the basis of what is best for *her as an individual agent.*

However, how this follows is not obvious. There are two main concerns here. First, it is not clear how an agent can derive a reason to cooperate from the mere injunction to Pareto-optimize. This is because the injunction to Pareto-optimize is different from the injunction to maximize. As Gauthier recognizes (2013a: 602), maximization happens at the level of individual actions (such that an individual must choose an action that will maximize her expected utility), whereas Pareto-optimization happens at the level of outcomes, which are only achievable if others likewise seek the Pareto-optimal outcome. To illustrate the difficulty, consider the following Hi-Lo game.

Table 2*: The Hi-Lo Game*

		Player B	
		Hi	Lo
Player A	Hi	3,3	0,0
	Lo	0,0	1,1

In the Hi-Lo game, each player does better by choosing Hi when she expects the other player to choose Hi and Lo when she expects the other player to choose Lo. But determining whether to choose Hi or Lo is not straightforward given the standard assumptions of classical game theory that each player is rational, and that there is common knowledge of each player's rationality. Consider the following line of reasoning, from the perspective of Player A:

(1) Hi/Hi is better for me than Lo/Lo.

(2) I should pick Hi if I can expect Player B to choose Hi.

(3) I can expect Player B to choose Hi if I expect that she expects I will choose Hi.

(4) Player B can expect that I will choose Hi if she expects that I will expect her to choose Hi.

In the Hi-Lo game, it is better for both of us if we both choose Hi. But our each choosing Hi depends on our expectations about what the other will choose, and these expectations are contingent on the expectations the other player has of our choices, which are in turn contingent on our expectations about the choices of the other player. It seems we are trapped at an infinite regress.[8]

A similar problem faces a player seeking the Pareto-optimal outcome in a Prisoner's Dilemma. A player's reasoning might run as follows:

(1) Pareto-optimal outcomes are better for me than equilibrium outcomes in Prisoner's Dilemma situations.
(2) In the Prisoner's Dilemma, I can bring about Pareto-optimal outcomes if I cooperate while the other player cooperates.
(3) I can expect the other player to cooperate when I can expect that she will expect that I will cooperate.
(4) The other player can expect that I will cooperate when she can expect that I can expect that she will cooperate.

Again, as in the Hi-Lo game, we are left without an answer to whether to cooperate. All that can be established in the above is that I should cooperate if I can expect the other player to cooperate. And the other player should cooperate when I can be expected to cooperate. But without knowing whether the other can be expected to cooperate, neither of us can know

[8] For a fuller discussion of the problematic nature of Hi-Lo games, see Bacharach (2006: 35-68).

whether to cooperate. Again, it seems we are trapped in an infinite regress.[9] Thus, appealing to the Pareto-optimality of the mutually cooperative outcome is insufficient to give individual agents a reason to cooperate. Unless agents know that the other is employing Pareto-optimization, individuals will have no reason to employ it themselves.

3.2. Making *My* Life Go Best

This brings us to the second concern about interpreting Pareto-optimization as an injunction to individuals. If it was a self-evident feature of rationality that individual agents should Pareto-optimize, then this might stop the regress and give an individual a reason to cooperate. However, to make the claim that it is rational for an individual to Pareto-optimize, it must be shown that Pareto-optimization makes an individual agent's life go best. And, while it is the case that Pareto-optimal outcomes make my life go better than do equilibrium outcomes, what gives me the *highest* utility—that is, makes my life go *best*—is unilateral defection. Thus, since what is *best* for individual agents is to defect on a cooperator, it does not follow from the mere fact that Pareto-optimal outcomes are better than equilibrium outcomes that rational individuals ought to aim at Pareto-optimal outcomes.

4. The Fragmentation of Practical Reason

4.1. Team and Individual Reasons

There is a more general problem here. According to Gauthier, there are two distinct modes of

[9] This might change if players could communicate. But then this moves us away from the original problem as presented in the PD.

reasoning: one that leads to equilibrium outcomes (what I have been calling the maximizing conception) and another that leads to Pareto-optimal outcomes (what Gauthier calls the Pareto-optimizing perspective, and that I have equated with team reasoning). We have seen that Gauthier recommends the adoption of the Pareto-optimizing perspective because doing so makes one's life go better than does the adoption of the maximizing conception. I have argued that Pareto-optimization gives a reason for team members to cooperate in Prisoner's Dilemmas but not necessarily a reason for individuals to cooperate. I have argued further that it is not clear that such a reason can be found, since an individual will do best by defecting whenever she can do so with impunity. Thus, while Pareto-optimization makes the lives of team members go better than does maximization, it does not do the same for the lives of individual agents, since unilateral defection makes an individual's life go *best*.

The more general problem is what Bratman (2013) diagnoses in a different context as 'the fragmentation of practical reason'. Bratman's criticism is directed at Gauthier's 1994 argument in "Assure and Threaten" concerning the rationality of making and carrying through with temporally extended plans. The basic problem is that while it might be rational (because it is a part of what make's one's life go best) for an agent to make a plan at a certain time, it might not be rational (because it is not part of what makes one's life go best) for an agent to carry through with the actions required by that plan when the time comes. Gauthier's concern in "Assure and Threaten" is with assurances and threats: he wants to be able to say that it is rational to make and then carry through with mutually advantageous assurances but not with mutually destructive threats. To deliver the desired result, Gauthier adopts what Bratman refers to as a 'pragmatic two-tier metatheory' (659), which determines which deliberative procedure will permit an agent to act in ways that are conducive to making her life go best. In

"Assure and Threaten," Gauthier defends a deliberative procedure that licenses some plans to be made and then carried through (e.g., assurances) and others (e.g., devastating threats) not.

The trouble is that the pragmatic test can be applied at the general or the particular level, which opens the door to a fragmentation of practical reason. At the general level, we arrive at a justification to employ a particular standard of deliberation that enjoins one to make assurances that one will cooperate with one's fellows. But at the particular level—the level of action—we arrive at a pressure to defect when one can do so with impunity. We are, as Bratman says, 'buffeted by two forms of rational pressure' (664).

To return to the context at hand, Gauthier seems to be applying this pragmatic test to the choice between Pareto-optimization and maximization as a standard of rationality. Gauthier's justification for giving up on maximization is that maximization does not contribute to making one's life go best. But this underlying justification does not always point to Pareto-optimization. The problem emerges from the fact that an individual may have a reason to Pareto-optimize (and become a team member) because doing so is conducive to making her life go well. But she may also have reason to maximize (and revert to an individual reasoner) when she can gain at the expense of other cooperators. This invites the question why one should privilege Pareto-optimization rather than maximization if they both pass the pragmatic test, albeit at different levels of application.

4.2. Why Privilege Teams Over Individuals?

The fragmentation of practical reason in the Bratman's context emerges because of an ambiguity about the level at which the pragmatic test should be applied. There the worry was whether the pragmatic test should be applied at the general or particular level, and what argument could be advanced in order to privilege the one over the other. In the context at hand, there is a related concern about which unit of agency should be privileged: that of the team or the individual?

We have seen that maximization and Pareto-optimization are modes of reasoning that correspond to two distinct units of agency: the former to individuals, and the latter to groups. We have also seen that Gauthier thinks that maximization should be replaced by Pareto-optimization in situations where cooperative benefits can be had. As he says, 'to the maximizer's charge that it cannot be rational for a person to take less than he can get, the Pareto-optimizer replies that it cannot be rational for each of a group of persons to take less than, acting together, each can get' (Gauthier 2013a: 607). Gauthier, as I understand him, is here saying that, while *you* can do better by maximizing, *you as part of the team* do better as a Pareto-optimizer. As we have seen, however, showing that a group agent does better by Pareto-optimizing than maximizing does not entail that an individual agent does better by Pareto-optimizing than by maximizing. Thus, appealing to the benefits conferred to the group agent does not speak to the challenge raised by the individual agent asking why she should not maximize.

This invites the question why we should care more about the benefits conferred to the group agent rather than those to the individual, and Gauthier does not provide us with an answer. It may be that he is appealing to the benefits of everyone, impartially considered, or to some principle associated with the social contract that may rely on notions beyond those that are derivable from game theory or considerations of individual advantage alone. But in that case, we might further ask why Gauthier would insist on a game theoretical framework at all, if he is ultimately to reject its very foundations in order to reconcile reason with cooperation.[10]

Conclusion

In this paper I have argued that Gauthier faces two problematic alternatives. In the first, Gauthier's recommendation to Pareto-optimize amounts to a recommendation to individuals on the basis of what is best for the team with which they identify. On this reading, Gauthier's view collapses into a variant of team reasoning. However, this leaves us without an answer to why I as an individual should cooperate. The second alternative is to understand the injunction to Pareto-optimize as a prescription to individual agents. However, without a common goal characteristic of team reasoning, agents will be either trapped in an infinite regress in trying to achieve cooperative outcomes, or unable to deem cooperation rational. The source of these difficulties can be found in the fragmentation of practical reason, which reveals that wherever an agent might have reason to employ a Pareto-optimizing perspective on rationality, she might equally have reason to employ a maximizing perspective. And as long as an agent links making one's life go best to rationality, it is hard to see how we can get around this difficulty.

Acknowledgments

Earlier versions of this paper were presented at the Centre for the Study of Mind in Nature's PPPE Workshop at the University of Oslo in 2014, and the Department of Philosophy Colloquium Series at Dalhousie University in 2015. Thanks to Alister Browne, Torfinn Huvenes, Duncan MacIntosh, Will Small, Sebastian Watzl, and two anonymous referees for

[10] I am indebted to an anonymous referee for this journal for this idea.

helpful comments and discussion on earlier drafts of the paper. This work was supported in part by the Research Council of Norway through its Centres of Excellence funding scheme, project number 179566/V20, and by Françoise Baylis' Canada Research Chair in Bioethics and Philosophy on "Impact Ethics: Making a Difference in Bioethics."

References

Bacharach, Michael 1999. Interactive Team Reasoning: a Contribution to the Theory of Co-operation, *Research in Economics* 53/2: 117-147.

Bacharach, Michael 2006. *Beyond Individual Choice: Teams and Frames in Game Theory*, Princeton: Princeton University Press.

Binmore, Ken 1994. *Game Theory and the Social Contract: Playing Fair* (Vol. 1), Cambridge, MA: MIT Press.

Bratman, Michael E. 2013. The Interplay of Intention and Reason, *Ethics* 123/4: 657-672.

Gauthier, David 1986. *Morals by Agreement*, Oxford: Oxford University Press.

Gauthier, David 1994. Assure and Threaten, *Ethics* 104/4: 690-721.

Gauthier, David 2013a. Twenty-Five On, *Ethics* 123/4: 601-624.

Gauthier, David 2013b. Achieving Pareto-Optimality: Invisible Hands, Social Contracts, and Rational Deliberation, *Rationality, Markets and Morals* 4/78: 191-204.

Gintis, Herbert 2000. Beyond *Homo Economicus*: Evidence From Experimental Economics. *Ecological Economics* 35/3: 311-322.

Gold, Natalie 2012. Team Reasoning, Framing and Cooperation, in *Evolution and Rationality: Decisions, Co-operation and Strategic Behaviour,* ed. Samir Okasha and Ken Binmore, Cambridge: Cambridge University Press: 185-212.

Gold, Natalie and Robert Sugden 2007. Theories of Team Agency, in *Rationality and Commitment*, ed. Fabienne Peter and Hans Bernhard Schmid, Oxford: Oxford University Press: 280-312.

Hakli, Raul, Kaarlo Miller, and Raimo Tuomela 2010. Two Kinds of We-Reasoning, *Economics and Philosophy* 26/3: 291-320.

Hurley, Susan L. 1989. *Natural Reasons: Personality and Polity*, Oxford: Oxford University Press.

Parfit, Derek 2001. Bombs and Coconuts, or Rational Irrationality, in *Practical Rationality and Preference: Essays for David Gauthier,* ed. Christopher W. Morris and Arthur Ripstein, Cambridge: Cambridge University Press: 81-97.

Setiya, Kieran 2014. Intentions, Plans, and Ethical Rationalism, in *Rational and Social Agency: The Philosophy of Michael Bratman*, ed. Manuel Vargas and Gideon Yaffe, Oxford: Oxford University Press: 56-82.

Smith, Holly 1991. Deriving Morality from Rationality, in *Contractarianism and Rational Choice: Essays on David Gauthier's Morals by Agreement*, ed. Peter Vallentyne, Cambridge: Cambridge University Press: 229-253.

Sugden, Robert 1993. Thinking as a Team: Towards an Explanation of Nonselfish Behavior, *Social Philosophy and Policy* 10(1): 69-89.

Sugden, Robert 2003. The Logic of Team Reasoning, *Philosophical Explorations* 6/3: 165-181.

Sugden, Robert 2011. Mutual Advantage, Conventions and Team Reasoning. *International Review of Economics* 58/1: 9-20.

Sugden, Robert 2015. Team Reasoning and Intentional Cooperation for Mutual Benefit, *Journal of Social Ontology* 1/1: 143-166

Thompson, Michael 2001. Two Forms of Practical Generality, in *Practical Rationality and Preference: Essays for David Gauthier,* ed. Christopher W. Morris and Arthur Ripstein, Cambridge: Cambridge University Press: 121-152.

Tuomela, Raimo 2007. Cooperation and the We-Perspective, in *Rationality and Commitment*, ed. Fabienne Peter and Hans Bernhard Schmid, Oxford: Oxford University Press: 227-257.

Vallentyne, Peter (Ed.) 1991. *Contractarianism and Rational choice: Essays on David Gauthier's Morals by Agreement*. Cambridge: Cambridge University Press, 1991.