Gauthier and the Prisoner's Dilemma

STEVEN KUHN *Georgetown University*

ABSTRACT: The Prisoner's Dilemma (PD) plays a central, but shifting, role in David Gauthier's moral theorizing. In "Morality and Advantage," it provides a model, demonstrating how morality can have seemingly contradictory properties. In Morals by Agreement*, it poses a special problem for the view that moral behaviour is individually rational. Authorities on game theory have subsequently disputed the idea that the PD is an appropriate tool for thinking about moral theory. In the first part of this paper, I examine the roles of the PD in Gauthier's writings. In the second part, I outline a project, with both descriptive and normative components, that develops the insights of "Morality and Advantage" while preserving it from the game theorists' attack.*

RÉSUMÉ : Le dilemme du prisonnier occupe une place centrale dans la théorie morale de Gauthier, mais cette place est en évolution. Dans «Morality and Advantage», ce dilemme fournit un modèle montrant comment la moralité peut avoir des propriétés apparemment contradictoires. Dans Morals by Agreement*, il pose un problème particulier pour l'opinion selon laquelle un comportement moral est individuellement rationnel. Suite à ces publications, certains experts en théorie des jeux ont contesté l'idée voulant que le dilemme du prisonnier soit un cadre approprié pour réfléchir sur la théorie morale. La première partie de cet article examine les rôles du dilemme du prisonnier dans l'œuvre de Gauthier. La deuxième partie présente un projet, incluant des composantes descriptives et normatives, qui développe les idées de «Morality and Advantage» tout en le préservant de la critique des théoriciens des jeux.*

Keywords: Gauthier, prisoner's dilemma, advocacy, social pressure, rationality

***Dialogue* 55 (2016), 659–676.**

doi:10.1017/S0012217316000603

David Gauthier's celebrated *Morals by Agreement* opens with the following autobiographical disclosure:

> The present enquiry began on a November afternoon in Los Angeles when, fumbling for words to express the peculiar relationship between morality and advantage, I was shown the Prisoner's Dilemma.[1]

A cursory examination of Gauthier's papers and books confirms that the Prisoner's Dilemma (PD) has long been central to his moral theorizing. The role that the puzzle plays in Gauthier's thought, however, has shifted. Furthermore, philosophically minded students of game theory have argued (sometimes stridently) that the emphasis on the PD in discussions of morality and cooperation is misguided. In the first three sections of this paper, I will review briefly the distinct roles the PD plays in "Morality and Advantage" and *Morals by Agreement*, and the recent complaints about its use in ethics. I believe that there is an important insight in the early *Philosophical Review* paper that may be lost or obscured in *Morals by Agreement*.[2] In recent years (though I did not think of myself as doing so), I have been trying to develop the insight of "Morality and Advantage" and preserve it against the assault of the game theorists. I see this project as having both a normative and a descriptive component. In the last two sections, I will explain the project and report my progress on each of its components.

I. Prisoner's Dilemma as Model

In "Morality and Advantage," the PD is invoked as a *model*, showing how it is possible for two seemingly inconsistent properties central to a system of moral rules to be jointly consistent:

Baier's Thesis: Morality is (at least partly) a system, S, of principles such that:

> B1. It is advantageous for everyone if everyone accepts and acts on S, and
> B2. Acting on S requires at least some to perform disadvantageous acts.[3]

To model these conditions in the PD, one simply identifies acting on S as the cooperation move and violating S as defection. Since defection dominates cooperation, condition B2 is clearly satisfied. Since mutual cooperation is

---

1 Gauthier 1986, v.

2 And in more recent writings. Gauthier 2015 advances the claim that cooperating in the PD, when one expects others to do likewise, is a rational act. I take no stand here on the proper analysis of *rationality*, but I do think emphasizing this point obscures the insight central to Gauthier 1967 that cooperative behaviour may sometimes require that we forgo personal advantage.

3 Gauthier 1967, 461–462, paraphrased.

preferred to mutual defection, B1 is satisfied as well. Indeed, dominance of defection and unanimous preference of universal cooperation to universal defection are generally taken to be *defining characteristics* of the PD game.

It is worth noting that Gauthier's explanation of *how* the PD reconciles the seemingly paradoxical conditions in Baier's Thesis requires conditions that are not always associated with the game, and indeed, cannot even be properly stated for the version of the game with ordinal payoffs that Gauthier considers in "Morality and Advantage":

> A second point to note is that each person must gain more from the disadvantageous acts performed by others than he loses from the disadvantageous acts performed by himself … This point may be clarified by an example. Suppose that the system contains exactly one principle. Everyone is always to tell the truth. It follows from the thesis that each person gains more from those occasions on which other others tell the truth, even though it is disadvantageous for them to do so, than he loses from those occasions on which he tells the truth even though it is disadvantageous for him to do so.[4]

Following common conventions, let us label the four possible payoffs in a (symmetric) PD, in descending order, as *T* (*temptation*), *R* (*reward*), *P* (*punishment*), and *S* (*sucker*). My gain when another chooses cooperation over defection is either *R*-*S* (if I cooperate) or *T*-*P* (if I defect). My loss from choosing cooperation over defection is either *T*-*R* (if the other cooperates) or *P*-*S* (if the other defects). So, the Gauthier condition requires that *R*-*S* and *T*-*P* each exceeds *T*-*R* and *P*-*S*. The definition of the PD ensures that two of these four inequalities are met: *R*-*S*>*P*-*S* and *T*-*P*>*T*-*R*. Two inequalities remain: *R*-*S* > *T*-*R* and *T*-*P* > *P*-*S*. Let us call these 'the Gauthier conditions' and rewrite them as follows:

G1. $R > (T+S)/2$
G2. $P < (T+S)/2$

Thus, G1 and G2 state that reward and punishment lie above and below the temptation/sucker average. A PD is a 'dilemma' regardless of whether these conditions are met. Condition G1 is sometimes, but not always, added as part of the definition of the game. G2, as far as I know, never is.

II. Prisoner's Dilemma as Problem

In *Morals by Agreement*, the PD is regarded, not as a model for moral behaviour that makes clear how it can have the paradoxical properties that it does, but rather as a *problem* that must be overcome for the grand project of reducing

[4] Gauthier 1967, 463.

https://doi.org/10.1017/S0012217316000603 Published online by Cambridge University Press

morality to rationality. What is missing or obscured in *Morals by Agreement* is the very commonsensical two-part idea to which I have just alluded:

1) Adhering to moral principles may require real *sacrifice* (the "disadvantageous acts" mentioned above).
2) The benefit we get from moral institutions is a result of the participation of *others*.

Of course, Gauthier still understands that cooperation in a PD is not individually utility maximizing. My payoff is increased when the other player cooperates, but it is reduced when I cooperate. That is why the PD poses a problem. If moral behaviour corresponds to cooperation in a PD, it is not individually rational. Gauthier's foils—Hume's knave, Hobbes' fool, and Plato's older brothers—are correct after all. The problem is overcome by observing that we benefit by acquiring a *disposition* to cooperate in PD situations, given the plausible assumption that such a disposition can be detected with some reliability by others. This is enough to show that it is rational (in the utility maximizing sense) to *be* moral and so, if the costs of doing so were not excessive, it would be rational to acquire and maintain the disposition to be a keeper of rational agreements with those similarly disposed. But the Gauthier of *Morals by Agreement* often seems to want more. He suggests that it also shows that it is rational to *act* morally, perhaps on the grounds that it is *always* rational to act on a disposition that it is rational (i.e., expected utility maximizing) to have. That suggestion has engendered critical reflection. In particular some have argued that there are circumstances under which the disposition to carry out one's threats is rational to *acquire* and to *possess* but, as things turn out, disastrously irrational to *act upon*.[5]

Although discussion in the literature has emphasized puzzles about the rationality of threat fulfillment, examples of threat *resistance* seem at least as troubling. Here is one from Derek Parfit,[6] which is itself adapted from an older example of Thomas Schelling. Derek lives on a small tropical island. Knowing that he is likely to face villainous threats from depraved bombers, he has successfully acquired a strong disposition to resist threats and made his possession

[5] See, for example, Skyrms 1996, 38–42. In the second edition of this work (Skyrms 2014), the explicit criticism of Gauthier on these pages is removed, but the critique of the rationality of threat-fulfillment remains (and plans including threats whose fulfillment might require actions failing to maximize utility are cited as failures of "sequential rationality" (see Skyrms 2014, 26). Gauthier himself came to view making and fulfilling threats as irrational in situations where fulfilling them could leave one worse off than never having made the threat (see Gauthier 1994). A careful and detailed discussion of Gauthier's divergent accounts of practical rationality along with a proposed replacement is contained in MacIntosh 2013.

[6] Parfit 2001.

https://doi.org/10.1017/S0012217316000603 Published online by Cambridge University Press

of this disposition highly visible. In so doing, he has maximized his expected utility, where that notion is defined as in Chapter 1 of *Morals by Agreement*. Alas, one depraved bomber unexpectedly fails to recognize the situation and credibly threatens to blow both Derek and himself to smithereens if Derek does not give him a coconut. Can we really imagine that it is more rational for Derek to act in accord with his disposition than to give up the coconut? Anticipating possible responses, we may stipulate, in addition, that the disposition to resist has already prevented threats of other bombers, so that Derek's life would not have gone better without it, and that it will maximize future expected utility, should Derek somehow survive the present threat.

Of course, threat-fulfilling and threat-resisting dispositions are not moral dispositions. Perhaps there is some as yet unimagined refinement of the principle that acts conforming to rational dispositions are rational, which might allow it to apply to the moral dispositions, but not the threat-fulfilling and threat-resisting ones. There is little reason for optimism, however. The categories are closely related. The *promise-keeping* disposition, for example, is a moral disposition—exactly the one that allows us to confidently participate in mutually advantageous agreements. But I can use that same disposition to make threats by saying, for example, "I promise to break your knee-caps if you don't give me that coconut."[7]

According to Gauthier 1998 and Gauthier 1994, it is a matter of *definition* that acts conforming to intentions or deliberative procedures that are rational in an appropriate sense are themselves rational. If we replace the disposition talk of *Morals by Agreement* with intention talk or deliberative procedure talk, this understanding of the matter does open the possibility that such acts do not themselves maximize utility, and thereby it perhaps preserves the insight that morality sometimes requires sacrifice.[8] It also, however, represents a break with

[7] Gauthier himself appreciates the difficulty of formulating an account of rationality that discriminates between threats and what he calls "assurances." (Gauthier 1994, 693). In defending the idea that it would be irrational to fulfill a threat when doing so would leave one worse off than never having made it, he notes that there are assurance situations to which the same reasoning applies. His commitment to the principle that it is rational to intend to perform an act at a time if and only if it is rational to perform the act at that time requires him to conclude that both the threats and the assurances are then themselves irrational, even if issuing them would maximize utility. (See Gauthier 1994, 714–716).

[8] I say 'perhaps' because the idea of a pragmatic instrumental rationality appealed to here seems to be that it should characterize some way for an agent to satisfy her current preferences, however selfish they might be, better than she could do by maximizing expected utility. Gauthier 1967 and Gauthier 1986 both employ a couplet of Ogden Nash to convey a bit of folk wisdom: "O Duty! / Why hast thou not the visage of a sweetie or a cutie?" It is true that moral duty, as portrayed in both works, does not have "the visage of a … cutie." I worry, however, that in the latter, she is still inappropriately attractive—a more refined mechanism for satisfying my preferences than direct utility maximization.

common understandings of rationality and with the account Gauthier himself so carefully and lucidly lays out for the case of "parametric choice" in Chapter I of *Morals by Agreement*.

The insight that morality may require behaviour that is not individually utility maximizing does survive in some form in *Morals by Agreement*. In Chapter VII, Gauthier considers agreements that, though fully rational in the circumstances in which they are made, would never have been made under fair and equal circumstances. The disadvantaged party, Gauthier says, ought morally to "acquiesce in" but not "comply with" such agreements[9] and (though Gauthier does not emphasize this point) the advantaged person ought to forgo making them. There is some suggestion that the instability of such agreements makes it disadvantageous to make and comply with them; Gauthier concludes the chapter, however, with an admission that this is not always the case:

> In reconciling reason and morals, we do not claim that it is never rational for one person to take advantage of another, never rational to ignore the proviso, never rational to comply with unfair practices. Such a claim would be false. We do claim that justice … is the virtue appropriate to co-operation, voluntarily accepted by equally rational persons. Morals arise in and from the rational agreement of equals.[10]

In the real world, where people are neither equal nor equally rational and where agreements are made under conditions where some have taken advantage of others, morality does sometimes require sacrifice. But this situation is certainly not *characteristic* of morality and it is not the situation modeled by the one-shot PD.

III. Game Theorists' Critique

Philosophically minded students of game theory are understandably sceptical of the idea that the PD is an appropriate device to guide thinking about morality. If there is one lesson that game theory has imparted, it is that it is hopeless to recommend any outcome in a game that it is not a Nash equilibrium. Mutual cooperation in a PD is most certainly not a Nash equilibrium. Let me briefly cite three examples.

Here is Kenneth Binmore:

> Game theorists think it is just plain wrong to claim that the Prisoners' Dilemma embodies the essence of the game of human cooperation. On the contrary, it represents a situation in which the dice are as loaded against the emergence of cooperation as they could possibly be. If the great game of life played by the human species were the Prisoners' Dilemma, we wouldn't have evolved as social animals![11]

---

9 Gauthier 1986, 230.

10 Gauthier 1986, 232.

11 Binmore 2005, 63.

As Binmore sees it, we are engaged in a game of life. This game has multiple equilibria. The game of morals is just a *coordination* game, in which we aim to select one of the equilibria in the game of life.

In a similar spirit, Brian Skyrms has urged moral philosophers to turn their attention from the PD, where mutual defection is the only equilibrium, to the Stag Hunt game, where there are two equilibria, one of which is unanimously preferred to the other. The problem that vexes Skyrms is how we can move from the inferior equilibrium to the superior one:

> If one simple game is to be chosen as an exemplar of the central problem of the social contract, which should it be? Many modern thinkers have focused on the prisoner's dilemma, but I believe that this emphasis is misplaced. The most appropriate choice is not the prisoner's dilemma, but rather the stag hunt.[12]

And, a few pages later:

> For a social contract theory to make sense the state of nature must be an equilibrium. Otherwise, there would not be the problem of transcending it. And the state where the social contract has been adopted must also be an equilibrium. Otherwise, the social contract would not be viable.[13]

A third PD-sceptic is economist Robert Sugden. Sugden's focus is on what we have *come to regard as* moral. He is very cautious about drawing genuinely normative conclusions. What we consider moral are widely followed conventions (i.e., stable equilibria in games with many stable equilibria) with the property that each party benefits from others following that equilibrium strategy.[14]

Sugden's full, 'official' formulation in the following chapter does leave some room for moral-seeming rules that are not conventions and he allows for those that are conventions to call for disadvantageous acts in certain "atypical"[15] applications. I think he has in mind the same one-shot situations over which Binmore thinks moral philosophers have long wasted their time. Binmore calls them "sore thumbs."[16] Despite Sugden's more nuanced and charitable treatment of moral philosophy, I think he would also find the emphasis on the PD by Gauthier and others misplaced.

IV. Preference Change and Indirect Contractarian Theory

My suggestion for saving the insights of "Morality and Advantage" is a very simple one. We should not see morality as a matter of acquiring a 'disposition'

---

12 Skyrms 2004, xii.

13 Skyrms 2004, 9.

14 Sugden 2005, Chapter 8.

15 Sugden 2005, 153, 159, 160.

16 See, for example, Binmore 1998, 37, 378, 451, 514; and Binmore 2005, 16.

that may or may not call for irrational behaviour. Rather, we should see it in terms of changes to our preferences and the concomitant changes in the payoff structures of games representing situations in which we often find ourselves. Games that would have the hopeless structure of a PD to those without moral sensibilities are transformed, for those with such sensibilities, into games with a more tractable structure. Two psychological traits make this possible. The first is the same trait that makes it useful for corporations and political candidates to spend millions of dollars in advertising. Our preferences can be changed by the urging of others.[17] These changes often result in considered and stable preferences that are revealed in both choices and thoughts and words—just the kind of preferences whose maximal satisfaction, according to Gauthier, characterizes rationality.

The second trait is that, on an individual basis, this advertising costs us nothing. Indeed, there seems to be little that we humans enjoy more than telling each other what to do and what to refrain from, what to admire and what to disdain, what to applaud and what to condemn. A change in preferences that has profound effects on what a person does may be brought about by simple and virtually costless actions on the part of others. Let me offer a single simple example. At the conference where this paper was first presented, one of my dinner companions was a student from the heartland of Canada who was attending the University of Prince Edward Island. I was curious about how he came to be going to such a small university so far from his home and inquired about it. The answer was that he hadn't really considered going to that university at all. But his parents owned a small dairy bar. One day he noticed a customer who was wearing a UPEI sweatshirt. It seems very likely that the customer, in deciding to wear the sweatshirt that had such a profound effect on the life of my dinner acquaintance, was satisfying his own preferences. It also seems likely that the choice of buying that particular shirt, and the choice of wearing it on that day cost the dairy bar customer little or nothing more than the alternatives he might have considered.

What we, as philosophers, might try to show to be rational are neither particular dispositions nor the actions that accord with them, but rather our advocacy of those actions in light of the effect of this advocacy on ourselves and others.

Although it does not do the major work in his theory, there is good evidence that Gauthier himself recognizes the power of the preference-change phenomenon and the pertinence of this 'indirect' approach to the social contract. Some of

[17] Duncan MacIntosh (MacIntosh 2013) argues persuasively in favour of this point and the further one that one can and sometimes should change one's own preferences. MacIntosh's argument is in the service of a pragmatic, instrumental notion of rationality that would allow for the issue and fulfillment of utility maximizing assurances and threats; mine was an effort to avoid relying on such a notion as a justification of morality. Nevertheless, there is common ground in our recognition of an ability to manipulate preferences and our view of the importance of this phenomenon.

this evidence comes from Chapter VII, where Gauthier tries to justify his theory of morality from the perspective of the ideal actor at what he calls the "Archimedean point":

> The ideal actor chooses, not compliance, but those processes of socialization that promote the circumstances in which narrow compliance is rational. Given the benefits that each may expect from cooperation, each has reason to prefer that … everyone be affectively engaged by compliance, so that the familiar feelings of respect and resentment of self-respect and guilt, are linked appropriately with fair and unfair behaviour of others and oneself. Although our primary concern is with the principles that would be chosen from the Archimedean point, we should not forget the importance of the choice of affections, in so far as these can be shaped by socialization.[18]

Other evidence comes from what I think of as the more 'poetic' chapters at the end of the book. In response to Glaucon's challenge to refute the thesis that we behave justly only because we are too weak to do otherwise, Gauthier waxes eloquently about the joys that moral agents take from their participation in cooperative activities. The best way to make this discussion cohere with what I think of as the more 'mathematical' early chapters is to understand moral dispositions, not as tendencies to curtail or constrain utility maximizing behaviour, but rather as tendencies to prefer the outcomes engendered by behaviour that would otherwise fail to be utility maximizing.

In the remainder of this paper, I discuss the idea that much of moral discourse is advocacy for certain changes in attitudes and preferences and that the morality of behaviour has to be explained in terms of conformity with what is rational to advocate. I have been thinking about this idea for a long time. Originally, I had conceived of it as being part of a normative theory—a rational contractarian theory of morality like Gauthier's. Gauthier's theory is direct. It tells us that actions are right if they, or the plans comprising them, or the principles of deliberation by which they were determined, would have been rationally agreed to by equals bargaining from a fair initial position. Mine was to be indirect—it would tell us that the degree of 'rightness' of actions is determined by the moral curriculum that it would be rational to agree to teach to each other.[19]

---

[18] Gauthier 1986, 266.

[19] See Kuhn 1996. It is reasonable to expect that, as with Gauthier's and other direct contractarian theories, some sort of equality and fairness conditions will have to be added to this formulation. The role of such conditions may be diminished, however. It is easier for an advantaged group to secure agreement on behavioural patterns that disproportionately benefit themselves than to secure agreement that we teach each other that these patterns are right. The fact that much of the most effective moral teaching is done privately gives the disadvantaged groups greater bargaining power over the curriculum than they would otherwise have.

More recently, I started to think about this idea as a descriptive theory of moral evolution—a theory about how we might have come to have the moral beliefs and attitudes that we, in fact, have. In fact, I think that these projects must be related. The aspect of rational contractarian theories that has most gripped its proponents has to do with justification. They aim to show that morality must have a claim, at least on *rational* individuals. But I think an equally important virtue is epistemological. It is a great mystery on most accounts, when and why my moral 'intuitions' or 'considered moral judgments' are guides to the truth. On an indirect account, at least the general form of the answer becomes obvious. My moral beliefs are correct just to the extent that the moral curriculum that produced them was rationally agreed to.

My evidence for the superiority of the indirect form over the direct form came mainly from some somewhat complicated observations about conditions when it is permissible to break agreements.[20] I now think the idea can be supported by a much simpler observation. Absent special circumstances, a person who has secured some particular benefit for another has performed an act of greater moral worth than a person who has secured that same benefit for himself. It seems much easier and more natural to explain this by an indirect theory—either utilitarian or contractarian—than a direct theory. The reason that it does not normally make sense to urge people to acquire goods for themselves is that they are already hard-wired to do so. Biology takes care of self-benefit directly. We need morality to take care of other-benefit.

Christopher Morris has discussed what seems to me to be an interesting case in point.[21] Morris asked about the directive once common at video rental stores, "Please rewind," which encapsulates a rule that one should rewind a tape after viewing it rather than leaving the job for the next customer (or an employee). Morris thought that the rule made sense because the psychological burden of rewinding after viewing was much lighter than the burden of doing so before. But the rule requiring rewinding before viewing has a notable advantage. If we require rewinding after viewing then, inevitably, some people will be required to rewind twice (because of the lapses of others), whereas if we require rewinding before, then all viewers will only have to rewind once. "Rewind before" is more equitable in this sense than "rewind after." What gives "rewind after" its

---

20 Kuhn, *ibid*. The evidence seen there to support indirect over direct theories includes the following. 1. Directness would seem to support the use of randomizing devices to determine whether certain agreements are kept; morality does not. 2. Directness would seem to imply that more transparent and more discerning agents incur stronger duties of fidelity; morality does not. 3. Indirect theories seem better able to account for the observation that, when unexpected events tempt one to break an agreement, the probability of the event and the size of the temptation have opposite effects on the permissibility doing so.

21 Christopher Morris 2011.

moral resonance, I think, is that it benefits somebody else, whereas "rewind before" benefits only me. It seems unlikely that anybody would misinterpret the "Please rewind" slogan as urging us to rewind before viewing. (No special urging needed for *that*!) The same considerations apply to a similar rule that may have survived longer than the video rental rule. In apartment buildings or families where clothes driers are shared, the rule 'clean lint trap after use' is more common than the rule 'clean lint trap before use.' Again, equity might seem to favour the second rule. And in this case it is much harder to maintain that the burden of following the second rule is greater. Moral considerations, however, push us towards benefitting others rather than benefitting ourselves.[22]

For both the prescriptive and the descriptive projects, I thought, like Gauthier—especially the Gauthier of "Morality and Advantage"—that the one-shot PD was exactly the right conceptual tool, and that morality had to make it possible for us to rationally choose cooperation in a situation that, in the absence of morality, would have a PD structure. Whether, in the presence of morality, this represents a sacrifice to me depends on the effectiveness of the moral education I have received. We might hope that the immoral lies, for example, are exactly the lies that I prefer not to tell. It is likely, however, that there are some situations in which it is rational to advocate truth-telling, while my own preference structure remains stubbornly skewed in favour of lying.

V. Advocacy Games

In the last section of this paper, I shall say more about the descriptive project. A framework that seems particularly useful for developing the idea is that of evolutionary game theory. Pairs from a population play a certain simple game, like the PD. Although they may play this game many times, each play is viewed as a *one-shot* game—it is assumed that players have no knowledge of the history of the previous play of their opponents, either with themselves or with third parties. (*Repeated* games, of the kind made famous by Robert Axelrod,[23] may have much to tell us about how selfish people can sometimes cooperate, but they have much less to tell us about why that cooperation might be regarded as moral behaviour.) The population *evolves* according to an appropriate evolutionary dynamic—more successful strategies become more widely adopted

[22] One participant at the conference where Morris' paper was presented suggested that the moral principle behind the video tape rule was that one should clean up one's own messes. That idea would lend credence to indirectness in a very similar way. The moral requirement to undo the harm we cause others does not seem to be matched by a requirement to undo harm we do to ourselves. Furthermore, if the lint traps and video tapes are personal possessions used by no others, whether we clean and rewind them before use or after use would seem to be a matter of moral indifference.

[23] See, for example, Axelrod, 1984.

https://doi.org/10.1017/S0012217316000603 Published online by Cambridge University Press

and less successful ones fall out of fashion. To model the phenomenon that interests me, I consider something I call 'advocacy games.' In addition to pairing randomly to play a simple game like the PD, players occasionally make an 'advocacy' move, i.e., they advocate a move in the underlying game. Advocacy itself does not result in any particular payoffs, but if enough players advocate the same move, the payoffs in the underlying game are revised to favour the move advocated. Moves in the underlying game (play moves) are assumed to be determined by mixed strategies. Advocacy strategies are pure. The idea is that an agent views a pure move like *cooperate* as a rule, and her mixed strategy indicates the seriousness with which she regards the rule, or at least the degree to which she follows it. Both play mix and advocacy moves are adjusted according to learning algorithms. If a player's recent returns from cooperation exceed those from defection, her play mix skews more towards cooperation. She *advocates* cooperation as long as recent payoffs while doing so exceed those while advocating defection. Advocacy is supposed to model something like 'social pressure' or the "moral or popular sanction" famously described by Jeremy Bentham.

> [If the source of pleasure or pain giving a binding force to a rule of conduct be] … at the hands of such *chance* persons in the community, as the party in question may happen in the course of his life to have concerns with, according to each man's spontaneous disposition, and not according to any settled or concerted rule, it may be said to issue from the *moral* or *popular sanction.*[24]

In the versions of the game I have considered so far, I take social pressure to be exerted population-wide, and to have a strength proportional to the fraction of the population exerting it. Here is a picture:

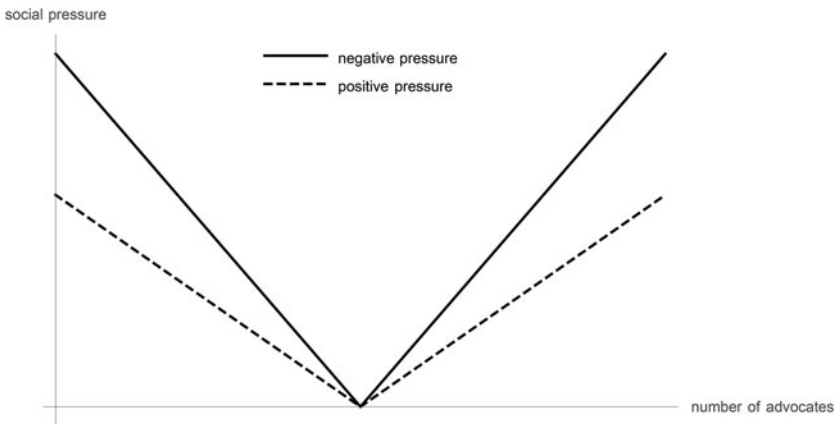

Figure 1 Advocacy and Social Pressure

[24] Bentham 1780, Chapter II, Section 5.

https://doi.org/10.1017/S0012217316000603 Published online by Cambridge University Press

The horizontal axis represents the number of advocates for one move, say 'cooperate,' in a two move game. Everyone not advocating that move is advocating the other. The vertical axis represents social pressures. Positive pressure is the amount by which payoff to the advocated move is increased. Negative pressure is the amount by which the payoff to the move not advocated is decreased. Pressures are zero when each move is advocated by half the population and they reach a maximum value when everybody advocates the same move. Positive and negative pressures kept distinct to model the possibility that there are different psychological limits to such pressures and to allow for comparison of carrot and stick approaches to behaviour modification.[25] (The plausible hypothesis that guilt is a stronger force than pride is undermined by the recent phenomenon of suicide bombers who claim to be motivated by positive incentives.)

To get some clue about what might happen in an advocacy game based on the PD, we might look at how the game would change under this social pressure. First consider pressure to cooperate:

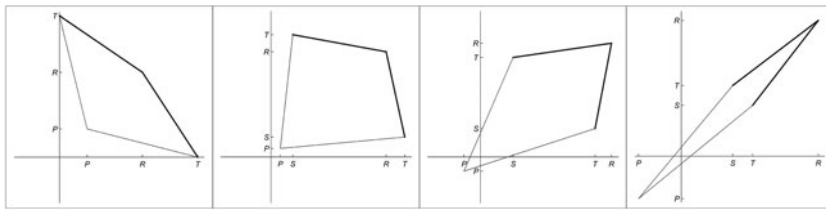

Figure 2 The PD under Pressure to Cooperate

The axes in each of the four graphs represent utilities of the two players. The southwestern vertex indicates the punishment payoffs that the x and y players get when they each defect. The northeastern indicates the reward payoffs. The northwestern and southeastern vertices indicate where x gets sucker and y gets temptation and where y gets sucker and x gets temptation. Lines connect pairs of points representing outcomes that differ in the move of only one player. The quadrilateral formed is stretched in successive frames. Positive pressure pulls the reward point further to the northeast and negative pressure pulls the punishment point further

[25] Since advocacy is costless, the possibility suggested here is just a matter of comparing the effectiveness of positive and negative changes in utility on behaviour. There is a clever argument in van Donselaar 2013 showing that, if rewards and punishments have costs, it is more efficient to reward good behaviour when the probability of that behaviour is small (because the reward is unlikely to be required) and to threaten punishment for bad behaviour when the probability of the good behaviour is high (because the punishment is unlikely to be required). Those interesting and important considerations are independent of the similarly interesting and important questions about the relative effects of carrots and sticks on future behaviour.

https://doi.org/10.1017/S0012217316000603 Published online by Cambridge University Press

southwest, while positive and negative pressures together push the sucker-temptation points inwards towards each other. By the second frame in our little movie, we can see by the marks on the axes, that *S* exceeds *P*. Now the game has become a game of Chicken. Each player gains by cooperating if his opponent defects and defecting if his opponent cooperates. By the third frame, in addition, *reward* exceeds *temptation*. Now cooperation *dominates* defection and mutual cooperation is the game's unique Nash equilibrium. If the original PD had different payoffs, it might happen that *reward* would surpass *temptation* before *punishment* surpassed *sucker*. In that case, the intermediate game is Stag Hunt rather than Chicken. No matter what the payoffs in the original PD, however, eventually the game will become a completely tractable one in which cooperation dominates defection and mutual cooperation is the most advantageous outcome for both players.

On the other hand, we must also consider what happens to a PD under pressure to defect. The unfortunate story is told in the following picture:

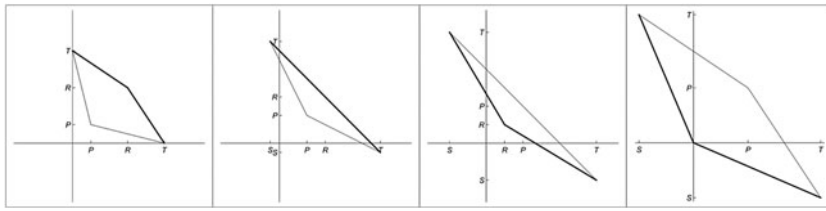

Figure 3 The PD under Pressure to Defect

Now the two sucker-temptation points are pulled apart, while the reward and punishment points are pushed towards each other. By frame three, they have crossed paths. The game becomes a 'prisoner's delight,' where defection dominates cooperation for each player, and mutual defection is unanimously preferred to mutual cooperation.

It is not obvious what will happen in an advocacy game in which the underlying structure is a PD. From the transformations in the previous two figures, one might suspect that, if social pressure is sufficiently strong, both universal defection with advocacy of defection and universal cooperation with advocacy of cooperation will be equilibria.

This idea gets further support by considering a simplified version of the game with a population of two:

Table 1 Simplified Version of the Advocacy PD

	C, AC	**C, AD**	**D, AC**	**D, AD**
C, AC	R+p, R+p	R, R	S+p, T–n	S, T
C, AD	R, R	R–n, R–n	S, T	S–n, T+p
D, AC	T–n, S+p	T, S	P–n, P–n	P, P
D, AD	T, S	T+p, S–n	P, P	P+p, P+p

https://doi.org/10.1017/S0012217316000603 Published online by Cambridge University Press

Here the positive and negative pressures are p and n when both players advocate the same move and zero when they advocate different moves. The resulting game has one equilibrium where both players defect and advocate defection. If the positive social pressure p exceeds T-R, then there is another equilibrium where both players cooperate and advocate cooperation. No matter what the pressures, there are no other equilibria. The cooperative equilibrium, if it exists, is always preferred by both players to the first.

In larger, evolutionary versions of the game, one might expect a 'neutral' population to move to the first equilibrium, where defection is both practiced and advocated. Since each player gets higher returns from defection than cooperation no matter what the others do, players should begin defecting. Advocating cooperation would then seem to be generally costly, lowering the payoffs of exactly the strategies that are widely employed and raising those that are not.

Simulations only partially confirm these expectations. In a wide variety of advocacy games based on the PD, players do tend to reach a state in which everyone always (or as often as the mutation rate permits) defects. Sometimes, however, they reach a state in which everyone is maximally cooperative. Advocacy seems to be somewhat less stable than behaviour. It is almost always true that most players advocate cooperation in the 'settled-cooperation' states and defection in the 'settled-defection' states. There are often some players, however, who we might call 'hypocritical' or, more charitably, 'reform-minded.' These agents play one move while advocating another. The number of such agents often remains at some fixed value greater than zero for long periods. It may also spike so that for short periods those advocating against the settled status quo may exceed half the population. Changes in behaviour generally are immediately preceded by changes in advocacy of that behaviour, but sometimes changes in behaviour seem to occur spontaneously, bringing changes in advocacy in their wake. Changes in advocacy do not always lead to long-lasting changes in behaviour. Sometimes the changes are reversed before any changes in behaviour take place (as if a reform movement was launched only to be quickly abandoned when it proved unsuccessful). A few examples of these games are shown in Figure 4 below.

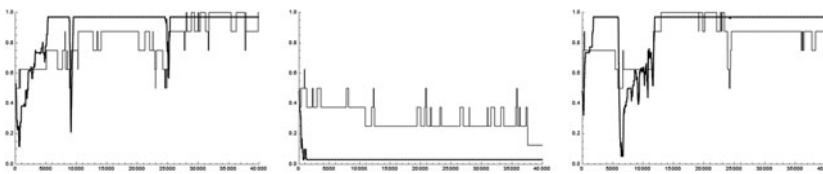

Figure 4 Payoffs and Cooperation in the Advocacy PD

The parameters for all three graphs are the same. The underlying game is a 'traditional' PD with payoff values of 5, 3, 1, and 0. The population size is eight. The horizontal axis represents the rounds of play from 0 to 40,000. The thicker plot traces the average play mix (i.e., probability of cooperation) of the

https://doi.org/10.1017/S0012217316000603 Published online by Cambridge University Press

eight players in the population. There is a built-in 'mutation rate' of .03, which prevents that line from going above .97 or below .03. The thinner plot traces the proportion of the population that advocates cooperation, so that line can assume nine values from 0 to 8. Since two out of the eight members play in each round, each player will have played about 10,000 games, 5000 as Player 1, from where he can consider changes in play and advocacy. With parameters set as they are here, each player will have considered advocacy changes about 100 times. The initial population is 'balanced': Players 1-4 advocate cooperation and have high play mixes themselves, while Players 5-8 advocate defection and have low play mixes. As might be expected, the play quickly (i.e., in the first 1,000 moves) moves towards defection. Advocacy remains relatively neutral for a while, however. In the first and third graphs, the proportion advocating cooperation begins to rise, with cooperative behaviour following closely as described above. In the middle graph, it falls somewhat and the population remains in the minimally cooperative state. While the population's average play mix is in the maximally or minimally cooperative state for most of the time, the population spends relatively little time in a state of uniform advocacy.

A little thought suggests a possible explanation for why settled cooperative states are sometimes reached. An advocacy move does have a relatively quick and direct effect on my payoff. It also has a more delayed and indirect effect, however, by changing the probabilities that I and others will continue to move in the way that we do. By advocating cooperation, I slightly increase the odds that others will cooperate and this benefits me. However, I also increase the odds that I myself cooperate, and this hurts me. If the benefit that I get from others' cooperation exceeds the cost of my own cooperation, it might be expected that it would behoove me to advocate cooperation. This is exactly the condition that Gauthier invoked to explain the seeming paradox of Baier's Thesis.

Indeed, simulations seem to confirm the importance of the Gauthier conditions for achieving cooperation in the advocacy PD:

Table 2 Payoffs and Cooperation in the Advocacy PD

Payoff Conditions	Trials Ending in Maximum Cooperation	Trials Ending in Maximum Defection
G1 & G2	38/50	6/50
G1 & ~G2	0/50	50/50
~G1 & G2	0/50	25/50

When the payoffs in the underlying PD were 10, 9, 1 and 0 (satisfying both G1 and G2), the population was in a state of maximal cooperation at move 4,000 in 38 trials out of 0 and in a state of maximal defection in only 6. When the payoffs were 10, 2, 1, and 0 (violating G1) the defection state was reached in

25 trials and the cooperation state in none. When the payoffs were 10, 9, 8 and 0 (violating G2), the defection state was reached in all 50 trials.

These simulations are suggestive, but they have a serious limitation. They all employ very small population sizes. I have come to realize that, for larger populations, reaching cooperative states in this kind of model may require unrealistically high numbers of interactions. Perhaps the only way that advocacy of the kind I envisage can evolve and change behaviour is if social pressure is modeled as a 'local' effect. Each agent is more likely to interact with some than with others. As the Bentham quotation indicates, the source of the moral or popular sanction is really not the entire population, but rather those individuals with whom one happens to interact. On a local version of the advocacy game, each agent would have her own payoff matrix, shaped in part by the advocacy moves of those with whom she has interacted, and especially by those with whom she interacts frequently. Frequency of interaction among pairs of agents in a population is not entirely random. As social networks like Facebook remind us, if John interacts frequently with Jane and Jane interacts frequently with Jill, then John is likely to interact more frequently with Jill than with a random stranger. One way to model this idea might be to think of the underlying game as a spatial game, where agents are arranged in a suitable geometry and the probability of interaction between a pair of agents is proportional to the distance between them. One might expect the results of local advocacy to be similar to those of global advocacy with small populations. In any event, investigations of this sort would seem to be an appropriate way to understand how morality can lead us to act contrary to our pre-moral preferences in much the way Gauthier's "Morality and Advantage" suggested it should do.

References

Axelrod, Robert
1984 *The Evolution of Cooperation*, New York: Basic Books, Inc.

Bentham, Jeremy
1780 *An Introduction to the Principles of Morals and Legislation*, Oxford: Clarendon Press.

Binmore, Kenneth
2005 *Natural Justice*, New York: Oxford University Press.

Binmore, Kenneth
1998 *Game Theory and the Social Contract II: Just Playing*, Cambridge Mass: MIT Press.

van Donselaar, Gijs
2013 "Sticks or Carrots? The Emergence of Self-Ownership," *Ethics* 123 (4), 700–716.

Gauthier, David
2015 "How I Learned to Stop Worrying and Love the Prisoner's Dilemma," in Martin Peterson (ed.), *The Prisoner's Dilemma*, Cambridge: Cambridge University Press, 35–53.

Gauthier, David
1998 "Intention and Deliberation," in Peter Danielson (ed.), *Modeling Rationality, Morality and Evolution*, New York: Oxford University Press, 41–54.
Gauthier, David
1994 "Assure and Threaten," *Ethics* 104 (4), 690–721.
Gauthier, David
1986 *Morals by Agreement*, New York: Oxford University Press.
Gauthier, David
1967 "Morality and Advantage," *Philosophical Review* 76 (4), 460–474.
Kuhn, Steven
1996 "Agreement Keeping and Indirect Moral Theory," *Journal of Philosophy* 93 (3), 105–128.
MacIntosh, Duncan
2013 "Assuring, Threatening, a Fully Maximizing Theory of Rationality and the Practical Duties of Agents," *Ethics* 123 (4), 625–656.
Morris, Christopher
2011 "Situating *Morals by Agreement*: Morality and its Parts," conference paper presented at "Contractarianism Twenty-Five Years after *Morals by Agreement*," York University, Toronto, Canada.
Parfit, Derek
2001 "Bombs and Coconuts or Irrational Rationality," in Christopher Morris and Arthur Ripstein (eds.), *Practical Rationality and Preference: Essays for David Gauthier*, New York, Cambridge University Press, 81–97.
Skyrms, Brian
2014 *Evolution of the Social Contract*, second edition, Cambridge: Cambridge University Press.
Skyrms, Brian
2004 *The Stag Hunt and the Evolution of Social Structure*, New York: Cambridge University Press.
Skyrms, Brian
1996 *Evolution of the Social Contract*, New York: Cambridge University Press.
Sugden, Robert
2005 *The Economics of Rights, Cooperation and Welfare*, second edition, Basingstoke: Palgrave MacMillan.