May 18th, 9:00 AM - May 21st, 5:00 PM

Productive versus destructive cooperation

Sheldon Wein
Saint Mary's University

Radu Neculau

Productive *versus* destructive cooperation

SHELDON WEIN

Department of Philosophy
Saint Mary's University
Halifax, Nova Scotia, B3H 3C3
Canada
sheldon.wein@gmail.com

ABSTRACT: Many of the problems we face can usefully be modeled as prisoners' dilemmas. All the standard game-theoretic solutions to prisoners' dilemmas lead, in the real world, to assurance games. But too often some aspects of our social interaction are as much obscured by, as illuminated by, game theory. Removing some of the epistemic constraints often accepted by game theorists will enable us to distinguish between productive and destructive prisoners' dilemmas. Doing so is an important step in understanding the nature of some of our social problems.

KEYWORDS: Assurance games, decision theory, Hobbes, prisoners' dilemmas, public goods, Rousseau.

1. INTRODUCTION

Though the fact that strategic choice can lead to sub-optimal outcomes has been appreciated for some time—at least since the work of Hobbes, Hume, and Rousseau—it was only during the Twentieth Century that we finally developed a clear and precise game-theoretic way to model both the prisoners' dilemma and assurance games. These models were developed during a period when both minimalist assumptions and mathematical rigor were—properly, in my view—highly valued. Thus our accounts of both the prisoners' dilemma and assurance games are standardly represented using only ordinal utility scales, and a background assumption is that interpersonal utility comparisons are not possible. For most purposes this ontological economy serves us extremely well. It has enabled theorists to represent the most important points in a manner that is simple without being simplistic, and it has allowed for many more people to understand extremely complex features of various coordination problems faced by contemporary societies than otherwise would be possible. (In this regard, the development of game theory with only ordinal utility scales and no interpersonal comparisons is similar to moving from the roman numeral system to Arabic numerals: suddenly almost everyone (regardless of their competence in mathematics) could understand a lot more than they could prior to that development. And just as we are not required to use game-theoretic models to explain strategic choice (neither Hobbes nor Rousseau had the benefit of game theory to help them), there is nothing in mathematics that requires the use of the Arabic number system. But we all know that, in both cases, all of us—those of us with ordinary levels of talent and mathematical geniuses—are able *in fact* to see much more than we otherwise could.) Such parsimony has been beneficial both to theorists and to practitioners—those who seek to find practical solutions, or find ways of implementing proposals the theoreticians have developed—and has allowed them a better chance of understanding the true nature of some of the problems we as a species deal with when trying to coordinate our interactions. So, on

Zenker, F. (ed.). *Argumentation: Cognition and Community. Proceedings of the 9th International Conference of the Ontario Society for the Study of Argumentation* (OSSA), May 18-21, 2011. Windsor, ON (CD ROM), pp. 1-12.

the whole (even if we set aside the ontological and epistemic arguments against cardinal utility scales and interpersonal comparisons of utility), the fact that we have shunned them has been extremely beneficial. (For an opposing view by a leader in the field, see Sen (2009), especially Chapter 8.)

But, as with most things, these benefits come with costs. In this paper I will explore how slightly enriching the game-theoretic understanding of prisoners' dilemmas and assurance games is significant in planning to avoid assurance dampers and which sorts of assurance amplifiers we should seek in various situations. In this case, the enriched games amount to no more than what we all absorbed when we learned about prisoners' dilemmas and assurance games. In particular, I will argue that we need to distinguish between what I will call productive prisoners' dilemmas and destructive prisoners' dilemmas.[1]

2. PRISONERS' DILEMMAS AND ASSURANCE GAMES

Prisoners' dilemmas and assurance games are defined in terms of the preferences of the participants. A prisoners' dilemma is a game where the players' preferences are these (from most preferred to least preferred): lone defection, mutual cooperation, mutual defection, and (worst of all) lone cooperation. An assurance game is any game where, for each participant, mutual cooperation is the best outcome and lone cooperation is the worst outcome. [2] Given these definitions, only ordinal utility scales are needed, and there is no assumption (indeed, there is usually the denial) that interpersonal comparisons of utility are possible.

In one sense this parsimony is to be expected. Why rely on assumptions regarding more extensive information than is needed when such assumptions face both epistemic and ontological challenges? But, looked at another way, the parsimony is surprising. After all, the original stories that virtually everyone uses to learn about the prisoners' dilemma and the assurance game (implicitly, at least) assume both cardinal utility scales and that interpersonal comparisons over the outcomes are possible. The story from which the prisoners' dilemma gets its name always has payoffs put in terms of something quite objective, time in jail. Admittedly, we are told only that the players prefer less time in jail to longer periods in jail, so strictly speaking we are not told how much one prefers, say, 1 year in jail to 10 years. However, we all tend to make implicit assumptions. (And, in my quick unscientific look through the literature the defect/defect option was *always* the collectively worst outcome. Of course, on reflection we might think that there is diminishing marginal disutility to time in jail and that we cannot make exact comparisons of how much Row and Column value freedom. But, for the purposes of what we are learning—that individual rationality leads to a Pareto sub-optimal outcome—I suspect almost everyone assumes that defect/defect is the collectively worst outcome. But doing that is, strictly speaking, not allowed.) The same is true of assurance games—although, given

[1] For reasons that will become clear, I am going to refer to these as PPDs and DPDs. Since selecting strategies for iterated prisoners' dilemma games is not itself a prisoners' dilemma problem, I will confine myself in this paper to one-shot prisoners' dilemmas. Of course, one way to (dis)solve an apparent prisoners' dilemma is show the parties that it is part of an iterated series of such interactions.

[2] I will not be concerned with whether the participants are indifferent between the remaining two outcomes or if one is preferred to the other.

that few of us have either venison or rabbit as a regular part of our diets, Rousseau's story of the stag hunters does not carry for us, as it did for his contemporary readers, the same sense of just how special eating deer meat is. But Rousseau is, nonetheless, able to indicate that had his hunters been able to kill the deer their reward would have been—both collectively and individually—quite great. Indeed, his point seems to be that despite the enormous payoff of killing the deer the hunters will not be able to achieve this result.

3. ORIGINAL STORIES

3.1 The prisoners' dilemma

The prisoners' dilemma derives its name from the following story. Row and Column have been accused of a crime. They have agreed with each other not to confess to the crime. But the prosecuting attorney tells Row that if she confesses to the crime and Column remains silent, Row will get off. If both confess, both will go to jail for a medium length of time. If both remain silent, both will go to jail for a short time. Of course, since the prosecutor is offering the same deal to Column as she is offering to Row, if Row remains silent and Column confesses, then Row will go to jail for a long time and Column will get off. Row must decide whether she should cooperate with Column and remain silent, or whether she should defect and confess to the prosecutor. Column also faces this choice.[3]

It would seem that it is most rational for Row to defect from her arrangement with Column and confess to the prosecutor, for if Row defects, she is better off no matter what Column does. That is, if Column defects, Row is better off defecting (she'll get a medium-length sentence) than she is cooperating (she will get a long sentence). And if Column cooperates, Row is still better off defecting (she will get off with no time in jail) than she is cooperating (she will get a short time in jail). The same is true for Column. So if each wants to minimize her jail time, both should defect. But if both defect, both will get a medium-length sentence in jail. If, instead, both cooperate, both only have to spend a short time in jail. The dilemma is simply that by doing what appears to be the rational thing for each to do in order to spend as little time as possible in jail, both will spend more time in jail than if both had acted irrationally.

In the following matrix, the numbers represent the number of years in jail:

[3] Typically, it is assumed that Row and Column are non-tuistic, that they take no interest in the other's interests. But this assumption is not essential. Even if Row and Column do care for each other—indeed even if each loves the other to the exclusion of herself—they can find themselves in a prisoners' dilemma. On this see Wein (1985).

Column

		cooperate	defect
Row	cooperate	***1***,1	***7***,0
	defect	***0***,7	***5***,5

Fig. 1. The prisoners' dilemma in terms of years in jail

If Row wants to stay out of jail, she will defect. If Column wants to avoid jail, she will defect. But if both defect, each spends longer in jail than if they had both cooperated. Since the players are assumed to prefer less time to more time in jail we can display their situation as follows:

Column

		cooperate	defect
Row	cooperate	***2***,2	***4***,1
	defect	***1***,4	***3***,3

Fig. 2. The prisoners' dilemma in terms of the players' preferences

3.2 The Assurance Game (or The Stag Hunt)

In Part II of his *Discourse on Inequality*, Rousseau tells what has come to be known as the stag hunt story. In Rousseau's tale a group of hunters go out into the forest to hunt for game. The hunters can cooperate and together hunt for a stag, surround it, kill it, and then eat very well. Alternatively, each might hunt on his own and catch a few rabbits and merely survive. The best outcome for each is that they all cooperate and kill the deer. But if even one hunter abandons the cooperative stag hunt to catch rabbits, the stag will escape through the "hole" that the hunter who has gone after a rabbit has left in the "fence". It is rational for each to continue to cooperate in the stag hunt rather than to defect to hunt for rabbits if, and only if, each hunter has adequate assurance that all others will also continue to cooperate. If any hunter lacks the assurance that all the others will continue to cooperate in the stag hunt, then she should abandon the stage hunt and go chase rabbits. This assurance that the other hunters will hunt the stag rather than chase a rabbit is something every hunter needs and which every hunter knows that every other hunter needs.

The best outcome for each is for joint cooperation resulting in lots of venison for everyone. The next-best outcome is to hunt rabbits on one's own regardless of what others do. The worst outcome is to continue the stag hunt when even one other hunter has abandoned it to chase rabbits. Because it is rational to continue hunting for the stag only if one has adequate assurance that others will do so also, games with this structure have come to be called assurance games. The matrices below display the standard assurance game:

		Column	
		cooperate	defect
Row	cooperate	***venison***, venison	***nothing***, rabbit
	defect	***rabbit***, nothing	***rabbit***, rabbit

Fig. 3. An assurance game in gastronomic terms

Since venison is preferred to rabbit and rabbit is preferred to nothing, we can represent the outcomes as below:

		Column	
		cooperate	defect
Row	cooperate	***best***, best	***worst***, neither best nor worst
	defect	***neither best nor worst***, worst	***neither best nor worst***, neither best nor worst

Fig. 4. An assurance game in terms of the players' preferences

3.3 Understanding Epistemic Impoverishment

Since game theorists typically assume only ordinal utility rankings and no interpersonal comparisons of value, there are many things about the individuals we are discussing which we do not know—indeed, cannot know. We cannot, for instance, say that Row likes her best outcome compared with her second-best one a lot more than she prefers the second-best outcome to her third-best outcome.[4] Nor can we say that Row's preferences are stronger (or weaker) than Column's when it comes to jail time. And we cannot say of one hunter that she likes rabbit more than another does, nor that she only slightly prefers eating venison to eating rabbit but finds the difference between having some rabbit to eat and going hungry to be enormous. It is not just that there are such facts and we happen not to know them; it is assumed that no such facts exist, that there is nothing there to be known.[5]

4. "SOLVING" PRISONERS' DILEMMAS

As I have argued elsewhere (Wein 2007b), every solution we have for escaping the suboptimal outcomes to which rationality moves us in prisoners' dilemma games leads, in the real world, to an assurance game. (To take just one example, if we follow Hobbes's recommendation and adopt an authority solution we are, in effect, hiring someone or an institution to change the circumstances so that we are no longer in a prisoners' dilemma game. But whether the institution has sufficient support to warrant your support depends

[4] Though, an important upside to accepting these epistemic constraints is that game theorists have developed extremely clever ways of reducing the impact of such constraints.

[5] This has led some, influenced by an overly constrained understanding of positivist theories of language, to hold that such claims are not even meaningful.

on whether enough others support or recognize the institution as an authority. Authorities only have authority as long as enough people recognize them as having authority. But then the question of whether to recognize a would-be authority as an actual authority is for each person an assurance problem.) I have also argued that argumentation theorists need to pay more attention to how to solve assurance games (Wein 2011).

When one surveys the literature on prisoners' dilemmas, one finds numerous proposed solutions. Hobbes tells us to use an authority solution. Gauthier (1986) says that we should change our conception of rationality from being straightforward maximizers to being constrained maximizers. MacIntosh (1992) tells us we should change our preferences, while Cave (1998) holds that we should cultivate the virtue of cooperation. Mintoff (2000) and Danielson (1992) think we should program ourselves to cooperate when doing so is likely to be reciprocated. And there are many more.

Most theorists—certainly all those I have listed above—seem to believe that their solution is *the* solution, that one size fits all. (I suspect, though I cannot argue for it here, that this is at least partly because they have viewed prisoners' dilemma games in the impoverished way that game theory represents them.) But thinking that there is a single solution to the problems that confront us because individual utility-maximizing rationality sometimes leads us to interact in sub-optimal ways is a huge mistake.

For most of human existence we did not get along very well, in part because we had little in the way of tools to help us coordinate our interactions. But lately (between 10 to 20 millennia ago) we have developed a host of devices for creating and sustaining larger societies: superstitions, and traditions, and complex emotions, and etiquette, and religions, and moralities, and even legal systems. These enable us interact cooperatively and, consequently, to live in groups much larger than was previously possible. And, subsequently, we have found ingenious ways to add substantial complexity to our traditions, feelings, religions, morals, codes of etiquette, and legal systems so that many of us now live in mega-cities each containing more humans than once roamed the entire planet.[6]

It is unlikely that one theoretical solution has spawned so many different practical ways of dealing with the problem. This suggests that the coordination problems we face are more varied and complex than is sometimes assumed. I now turn to distinguishing between what I take to be the most important division among prisoners' dilemma games.

5. PRODUCTIVE AND DESTRUCTIVE PRISONERS' DILEMMAS

Since a prisoners' dilemma is defined solely in terms of preferences over outcomes and since I am going to describe situations where more is discussed than just preferences over outcomes, I will adopt the following convention: a "prisoners' dilemma" is any situation where the players preferences are (from best to worst) lone defection, mutual cooperation, mutual defection, and lone cooperation. A "PD" is any situation where it is reasonable to suppose that the players would have preferences such that the PD is a prisoners'

[6] I realize that these claims are extremely contentious. I have relied on Chapais (2008), Gat (2006), and Pinker (2007). If the history of our species is radically different from what I am supposing, the relevance—but not, I believe, the validity—of my argument will be substantially altered. Thus, those who take a Rousseauian view that the pre-history of humans was a time when noble savages lived in peace and harmony can use my arguments to help explain why game theory does not completely explain our collective fall from grace. See Wein (2007a) for problems confronting new urbanites in the developing world.

dilemma. Note that the original story from which we all learned about prisoners' dilemmas is a PD. We are not just told the players' preferences; we are also told that these are based on something—a desire to stay out of jail or, if jail be necessary, to spend as little time in jail as possible. And, typically at least, we are given the length of each jail sentence.[7]

But if we think of jail time as something that can sensibly be agglomerated, we can distinguish between two sorts of PDs our players might find themselves in. They might be in a *productive* PD (or PPD) or in a *destructive* PD (or DPD). The difference is this: In a PPD, mutual cooperation is the collectively best outcome, whereas in a DPD (at least) one of either cooperate/defect or defect/cooperate is the collectively best outcome. Thus if the arrangement that the prosecuting attorney offers the players is like that of Figure 5 below, they are in a PPD, but if it is like Figure 6, they are in a DPD:

Column

		cooperate	defect
Row	cooperate	***1***,1	***3***,0
	defect	***0***,3	***2***,2

Fig. 5. A productive PD (in terms of years in jail)

Column

		cooperate	defect
Row	cooperate	***3***,3	***5***,0
	defect	***0***,5	***4***,4

Fig. 6. A destructive PD (in terms of years in jail)

In the situation laid out in Figure 5, the collectively best outcome is mutual cooperation (which yields collectively only 2 years in jail); hence 5 is a PPD. All other outcomes are worse (yielding longer collective times in jail).

But the situation in 6 is a DPD; the collectively best outcomes are defect/cooperate and cooperate/defect. Each of these yields only 5 years in jail and all other outcomes yield more time (6 years or 8 years).

Now, in fanciful cases like the ones above the distinction between PPDs and DPDs might not seem important or especially interesting. But in more realistic cases it can be very important. I turn to two such cases now.

[7] In my quick unscientific survey of the literature every discussion of the original prisoners' dilemma story was told in such a way that if was a productive PD. There was not a single case of it being a destructive PD.

First, imagine two islands named X and Y. Each island is populated with two groups, the Rows and the Columns. (On both X and Y the Rows and the Columns have strong tribal loyalties, so there are relatively few problems with cooperation among the Rows or among the Columns in either place. Every Row sees the interests of other Rows as ones she shares; the same is true for Columns.) Now suppose that both have agricultural economies and that their productive outputs depend on whether the Rows and Columns cooperate with each other. On both islands the islanders find themselves in a PD. But, due entirely to geographical and climactic factors, the situations differ with respect to how much food can be produced. Figures 7 and 8 lay out the differences:

Columns

		cooperate	defect
Rows	cooperate	***700***, 700	***25***, 800
	defect	***800***, 25	***50***, 50

Fig. 7. Island X’s agricultural production (in terms of bushels of wheat)

Columns

		cooperate	defect
Rows	cooperate	***100***, 100	***25***, 2000
	defect	***2000***, 25	***50***, 50

Fig. 8. Island Y’s agricultural production (in terms of bushels of wheat)

The inhabitants of both islands are in PDs. Both Island X and Island Y will be agriculturally much more productive if the islanders manage to avoid mutual defection. But note that only on Island X will it be the case that the island as a whole is most productive if the solution to the PD is mutual cooperation. Those living on Island Y are, through no fault of their own, in circumstances where if one group defects while the other cooperates the island as a whole will be collectively most productive. Put in the terms I have introduced, those people living on Island X are in a PPD, but those on Island Y are in a DPD. Or, to put it another way, while trying the idea of enslaving the other tribe is attractive to each tribe on both islands, only on Island Y does slavery make the island as a whole more productive than it otherwise could be.[8] Thus, given a few plausible assumptions about human

[8] Though he does not put it in these terms, Sen (1999) points to evidence that this may have been the situation in the United States of America prior to its Civil War. The North was in circumstances like my Island X and the South in those like my Island Y.

motivation, the tools needed to avoid slavery may well differ from island to island. Those on Island X face a hard problem of how to avoid their collectively worst outcome—which may well amount to a living that is near subsistence—and construct a mutually cooperative and collectively prosperous society. Those on Island Y face that same problem *plus* the problem of how to avoid falling into being seduced into seeking great wealth for some by means of repressing others. The problem might be even more serious in cases where the payoffs to the two tribes are asymmetrical.

Thus, it might be the case that those on Island X could use morality alone to both avoid mutual defection and ensure that both tribes are treated well in a system of fair mutual cooperation. If they understand their situation correctly they will see that mutual cooperation makes the both as members of their own tribe and as islanders better off. Indeed, as islanders they are collectively as well of as they can be. But for those on Island Y—and this is just speculation on my part—morality may not be enough to have the two tribes cooperate. Since the collectively best situation is where one tribe gets the other to cooperate while it defects, it may be that a good dose of religion and superstition may be needed to reach the same level of cooperation that those one Island X can attain without these crutches. But note that if we just talk in terms of prisoners' dilemmas we cannot point to any difference between the two islands. Hence we have to assume their problems are the same and, consequently, what served as a solution for one island would serve as a solution for those on the other island. But, that may not be the case.

I do not mean to claim that DPDs necessarily require "stronger" measures to enable the participants to avoid sub-optimal outcomes. Let us turn to a more familiar case. Anne and Andrew are married professionals and, like many of their neighbours and friends, they have a young child. Each of them aspires both to be a good parent and to be a success in their professional life. Let us suppose that being successful amounts to becoming a partner in the firm at which they are employed. "Making partner" requires many long hours of hard work. Raising a child—at least in the social circumstances in which Anne and Andrew find themselves—requires many hours of ferrying the child from music lessons to ballet class to soccer practice to the orthodontist to second (and third?) language classes to swimming lessons and so forth. Anne and Andrew each love their child dearly and want all these things for her. But each would prefer that the other do most of the child ferrying, while she or he just had, say, a special lunch out with the child each weekend (both to bond with the child and to give the other parent some time alone). Now it is possible that if they shared the childcare duties equally, each of them would make partner but as both of them realize, this is extremely unlikely. Indeed, it is only if one of them does most of the child-ferrying that one member of the couple (*viz*., the one who does not do much child-ferrying) make partner. Each prefers that she (or he) be the one who makes partner while the other does most of the childrearing. But, loving their daughter as they both do, they also prefer to share the childrearing rather than have their lovely daughter not be properly taken care of. Further suppose that if each shares the work of bringing up their child, they will each earn $100K, but that if one does the childrearing she will make only $50K, while the other (who then becomes a partner in the firm) will earn $500K. If both ignore the child (that is, if they let her grow up with only the level of attention that they each received from their parents), each will be so racked by guilt that she (or he) will only earn $75K.

Anne and Andrew are in a DPD. From the point of view of the household they will earn the most money if one sacrifices her or his career so that the other will have a better chance at success. And the child will—by their lights at least—be better off. With $550K coming in the household can afford to have the pool heated with solar panels and the swimming instructor come to their pool, to get new soccer shoes every month instead of just each year, to have her learn both Spanish and Mandarin, and so forth.

Andrew

		cooperate	defect
Anne	cooperate	***$100K***, $100K	***$50K***, $500K
	defect	***$500K***, $50K	***$75K***, $75K

Fig. 8. Anne and Andrew try to make partner while being good partners.

Some aspects of this DPD are hard to resolve.[9] But other aspects are (at least partially) amenable to rather simple solutions (such as changes in divorce law and the introduction of child-support rules). What is more difficult is to figure out not just how to protect the partner who sacrifices her career success to her child's success but how to organize our society so that fewer couples are confronted with choices in such socially unwelcome and unfriendly circumstances. Ideally, those couples raising children who find themselves in situations of partial conflict should, so far as is possible, be in assurance games rather than prisoners' dilemmas. But, when we cannot figure out how to arrange society so that they are in assurance games, we at least want them to be in productive PDs rather than destructive ones. (Changing the tax structure or the pay system within partnerships so $144,999 was the maximum possible level or earnings would change destructive PD Anne and Andrew face into a productive one.)

Note that if we all adhere to the strictures of game theory—use only ordinal preference rankings and never allow for interpersonal comparisons—the (relatively simple) problems I have outlined in this section do not even arise. To the contrary, they would be completely invisible. Of course, many of the problems we collectively face are much more complex. They will require all our resources both to properly represent them and so that we can find the best arguments for how to solve such collective problems. Game theory, combined with argumentation theory, offers a great deal of promise in this area (see van Eemeren and Grootendorst 1984). But we need to find a way to be less strict, without being lax, regarding what epistemic constraints we impose when representing such problems.

It is, of course, very difficult to make non-misleading generalizations about a topic as complex as the trends in payoffs in strategic interactions among billions of humans. Whether we are entering a period during which collective cooperation on roughly

[9] Here I think of the many non-obvious causes feminist research has uncovered for why it is so much more likely that Andrew will be making partner while Anne is busy SUVing their child from one important "activity" to another and all the non-obvious ways in which this harms them both (Wein 2007c).

fair terms is becoming easier to achieve or whether the benefits of exploitation and oppression will come to be even more attractive is immensely difficult to pin down. The recent spectacular progress made by societies which until recently suffered from violence, oppression, and systemic lack of education gives grounds for hope. But the survival of slavery, and the recent signs of grown in the sex slave industry, is sad reminder that creating societies that are cooperative ventures for mutual advantage for each is no easy task. It will take a careful mix of our devices for fostering and sustaining cooperation to enable us to create the sort of world where more of us live fulfilling lives. We need to understand the full nature of the problems we face if we are to enhance the probabilities of hitting upon appropriate solutions.

7. CONCLUSION

Our world is becoming a safer and better one. But promoting and protecting the improvements that so many of us enjoy (and to which that so many more of us aspire) will require careful use of our talents at constructing social arrangements which treat each of us with the concern and respect that is appropriate. Game theory can play a role in helping us in these tasks. But we need to liberate it from some relics of its positivist past. As Scott Shapiro has shown, humans sometimes need institutional and organizational plans when they are confronted with moral problems whose solutions are contentious, complex, or arbitrary (Shapiro, 2011). Shapiro’s primary concern is with showing how this way of looking at things provides us with a firm foundation for a scientific sociology of legal systems. He is only tangentially interested in which planning systems work best in which circumstances. Yet one of the most difficult questions we confront in planning for planning is whether to use law or some other social device to solve the problem at hand. My contention is that by removing some of the epistemic constraints on how we describe the conflicts that our circumstances, combined with our rationality, produce, we will be better able to use the powerful resources game theory provides to help us locate and describe our problems so that we have a better chance of avoiding or removing them.

ACKNOWLEDGEMENTS: I thank Thea E. Smith for her help and the Dean of Arts at Saint Mary's University for generous financial support.

REFERENCES

Cave, E. M. (1998). *Preferring Justice: Rationality, Self-Transformation and the Sense of Justice*. Boulder: Westview Press.

Chapais, Bernard. (2008). *Primeval Kinship: How Pair-Bonding Gave Birth to Human Society*. Cambridge: Harvard University Press.

Danielson, P. (1992). *Artificial Morality: Virtuous Robots for Virtual Games*. New York: Routledge.

Eemeren, F.H. van and Grootendorst R. (1984). *Speech Acts in Argumentative Discussions*. Dordrecht: Foris Publications.

Gat, Azar (2006). *War and Human Civilization*. Oxford: Oxford University Press.

Gauthier, David (1986). *Morals by Agreement*. New York: Oxford University Press.

MacIntosh, P.D. (1992). Preference Revision and the Paradoxes of Instrumental Rationality. *Canadian Journal of Philosophy* Volume 22, No. 4, December.

Mintoff, J. (2000). Is Rational and Voluntary Constraint Possible? *Dialogue* Volume 39.

Pinker, S. (2007). A History of Violence. *The New Republic* March 17th edition.

Sen, A. (1999). *Development as Freedom.* New York: Alfred A. Knopf.

Sen, A. (2009). *The Idea of Justice*. Cambridge, MA: Harvard University Press.

Shapiro, S. J. (2011). *Legality*. Cambridge, MA: Harvard University Press.

Wein, S. (1985). Prisoners' Dilemmas, Tuism, and Rationality. *Simulation and Games*, Volume 16. # 1.

Wein, S. (1997). Feminist Consciousness and Community Development. *The International Journal of Social Economics,* Volume 24, Issue 12.

Wein, S. (2007a). Urbanization in the Developing World. *The Dalhousie Review*, Volume 87, No.1, Spring.

Wein, S. (2007b). From MADness to SANity. *Peace Review: A Journal of Social Justice*, 19 (2).

Wein, S (2011). Assuring Cooperation: From Prisoners' Dilemmas to Assurance Games to Mutual Cooperation. *Proceedings of the 7th Conference of the International Society for the Study of Argumentation,* Amsterdam: ISSA.

Commentary on “PRODUCTIVE *VERSUS* DESTRUCTIVE COOPERATION” by Sheldon Wein

RADU NECULAU

Department of Philosophy
University of Windsorn
401 Sunset Avenue, Windsor, ON N9B 3P4
Canada
neculau@uwindsor.ca

I am grateful to Sheldon Wein for confirming my suspicion that what is interesting about the prisoner’s dilemma (PD), the stag hunt, and other game-theoretical scenarios is precisely what is left out of the more narrowly specialized discussions of such scenarios. Now, it is not hard to understand what makes game theory such an attractive framework for social analysis. There is much to admire in the simplicity and elegance of its formal structures and it is easy to see how this reinforces the normative authority of whatever outcomes they produce. The clarity of perspective game theory affords us is a commendable quality in an age of post-enlightened cultural immaturity and ethical confusion. The conceptual content that is derived from constructive applications of PD may also provide us with something a bit more coherent than any of the prevailing versions of the Western evaluative consensus. However, what the non-specialist finds truly interesting about this fairly limited expression of social rationality is the response to the creative pressure put on the basic model by various components of its underlying normative infrastructure. In my commentary, I will briefly discuss three of these components—the evaluative, the will forming, and the socially integrative dimensions of rational choice—and in the process also highlight what I take to be the most inciting parts of Wein’s paper. In so doing I also hope to provide further support for Wein’s critique of the epistemic constraints that limit game theoretical approaches to complex cases of action coordination or social conflict.

The first component concerns the evaluative dimension of assurance seeking and cooperative action. Game theorists are right not to overburden their formal schemes with the task of accounting for personal and interpersonal value preferences. They are also correct to avoid ranking these preferences based on qualitative distinctions that assume other, perhaps also questionable evaluative schemes. This is not the point of such limited exercises in choice rationality and it is also not their strongest suit. To illustrate this, consider any PD-like situation in which two parties hold settled and transparent views on what is worth sacrificing one’s life for. When they start thinking about their options and considering their choices, they usually assume, rightly or wrongly, that they have enough knowledge of the other party’s evaluative preferences and commitments to know what that party is more likely to choose in an identical situation provided that she employs correct reasoning. And so, if the first party chooses, for instance, what she initially takes to be the most advantageous option for both, and if this turns out to be based on wrong motivational and evaluative assumptions about what the other party actually chooses (say, lone defection), the resultant choice would have to be characterized as an error in judgment. The error here would be caused by incorrect or incomplete information about the other party’s initial orientation to value, or about her actual strength of will or, in some

Zenker, F. (ed.). *Argumentation: Cognition and Community. Proceedings of the 9th International Conference of the Ontario Society for the Study of Argumentation* (OSSA), May 18-21, 2011. Windsor, ON (CD ROM), pp. 1-4.

more complicated cases, about her true ability to reason consistently in accord with her expressed orientation to value or her willpower.

Yet I wonder if this kind of explanation can account for the occurrence of such errors in any except the most simplistic types of PD scenarios. Isn't there more to such scenarios than merely optimally selecting one from the several possible value-neutralized outcomes? Aren't we ignoring a more complex process of value-interpretation in which one's value orientation or preference is actually clarified by the choices one makes? Doesn't this also help to articulate the meaning of the value in question? And could it be that values either acquire or enhance their normative force through such interpretations or through deliberations over which outcome is more consistent with values that are implicit in what one takes to be (in) one's own interest? In many cases, it is this force itself that turns values into effective grounds for action (thus enhancing their appeal). If so, shouldn't we focus less on the outcome of the deliberation in PD situations and more on the evaluative aspects of such deliberations that help clarify the choice, identify its grounds, and perhaps anticipate some of its effects on the complex web that ties together individuals and groups in their world of shared and historically developed meaning? In typical discussions of standard PD cases, the goal always seems to be to find ways of getting out of a tight spot or reaping the highest reward in the most efficient way possible. But many times the deliberation itself is an opportunity to test one's allegiance to a value or to norms that are allegedly derived from it. It is also a commitment to an interpretation of value or to a possible way of applying a norm to a particular situation where other values are also involved. In many such cases, what appears to be an error of judgment in standard discussions of PD is actually a form of evaluative error-reduction, as in finally figuring out one's values or one's interpretation of value. Perhaps the lone defector in a PD situation rather than cynically calculating what's to her advantage is in fact experiencing an epiphany of sorts that endows her with greater moral insight.

This brings me to the second and closely related component, which is what social philosophers call the will-forming effect of practical judging in PD-like situations. The will in question is both personal and interpersonal, that is, social or group-based, and the will-forming dimension of rational choice can be explained along the following lines: When agents make choices that are bound to have a considerable impact on their lives and the lives of others, and when the possibility of error forces them to consider the complex structure of willing that is involved in such apparently simple decisions as choosing one object over another (a structure that includes ends and means, values and duties, deliberation and action), they change in important ways. They acquire a clearer sense of who they are not only in terms of the wider horizon of value that provides their choice with meaning (as noted in my discussion of the first component), but also in terms of how this choice impacts others, whose desires, expectations, sense of selfhood, etc. play an equally important role in deliberation. When agents take these into account, the 'we' gradually replaces the 'I' and a we-attitude (as Raimo Tuomela calls it) supplants the initial I-attitude. In this transition from a single perspective to a common perspective (or at least a more inclusive one), a new and different individual will emerges, as well as a collective one.

All this is rarely captured by the often simplistic and highly reductive conception of agency that is used in game theory. Game-theoretical accounts derive their conception of personhood from the notion of utility maximization, which means that the agent's capacity for rationality is both established as well as exhausted by the act of choosing from different possible outcomes. Now, most choices people make and many of the PD-like situations they find themselves thrown into have no transformative effects that would lead to a qualitatively different will or to a we-attitude that could signal the presence of a common will. In such situations, the notion of agency used by rational choice theorists seems to be perfectly adequate. But the meaningful cases, the ones that are not reducible to stochastic descriptions, are not of this type, and Wein in his paper gives us at least one example of such a case, the example of the professional couple that must cooperate destructively in order to maximize the payoff for their entire family. This example contains all the necessary elements for explaining the development of an individual's will that is not only practically mindful of others (or prudent) but also understands itself in terms of others; that is, a will that coordinates with these others and subordinates its ends to those acceptable to all. The example helps us explain the emergence of a common will through attitudes of mutual acknowledgment. It also highlights the importance of shared values that make sacrificing one's private goals the only desirable and thus rationally acceptable end. (The Hobbesian appeal to authority that Wein mentions in his paper could also be understood as an appeal to the implicit good that justifies the institution of monopolies on the use of power.) Finally, the example captures what is essential about the creation of a collective self-image ("family values") that can be culturally internalized and socially reproduced through policies that aim at the public good and also help define what the public good is.

It is in this context that Wein's distinction between productive and destructive forms of PD shows its importance, and we realize this is as soon as we ask the following question: destructive for whom? This distinction seems to capture the motivational gap between the centripetal force of collective ends that pull individuals together and the centrifugal tendencies of pursuing private goals and desires. It may also signal the presence of a conflict of values that fractures the common will between orientations to different and perhaps incompatible goods. Whatever the case, this example presents us with an important instance of interpersonal will formation that goes beyond the contractualist or corporate conception of action coordination many rational choice theorists tend to assume. In addition to this, it shows us how individual wills reemerge from a form of collective agency that has lost its cohesion as a result of choices made outside the common value horizon.

The third component of the wider normative infrastructure that puts additional pressure on the narrow deployment of strategic choice models to social analysis is revealed by the study of complex, large-scale action coordination that reduces choice to performing predetermined functions within the various administrative systems of the modern world. This operation of substitution effectively relegates personal interaction to a species of structural or functional adjustment, and by ignoring the specific demands and pressures of system integration game theory ends up as its first enabler. The distinction productive-destructive PD effectively captures this reality. For what is destructive in the example of the family that chooses strategically is precisely what is system-optimal on a functionalist analysis of resource allocation.

Can one save rational choice from the destructive effects of system integration? To answer this question, it may be worth recounting that Rousseau's discussion of stag hunting as a problem of action coordination belongs in a wider philosophical account of social integration. Social integration or creating the general will, according to Rousseau, can be achieved in one of the following two ways: by collectively pursuing the same object or by teaching individuals how to harmonize their interests and speak with a universal voice. The first one requires the presence of assurance providing institutions of rule enforcement, or authority. The second one demands the affirmation of a different kind of social autonomy, or communicatively playing a re-assurance game. I am not sure if these two are mutually exclusive. However, traditional game theory, as Wein presents it, seems to favor something closer to the first strategy, which makes it difficult for it to explain how choice disappears in the monotonous system manufacturing of social objects. The alternative Wein himself seems to favor looks at choice in relation to a theory of social rationalization that may indirectly rely on a conception of communicatively achieved interpersonal harmonization. It would be interesting to know where exactly he stands on this particular issue.

Reply to Radu Neculau: Deconstructing productive cooperation—further reflections on cooperative plans

SHELDON WEIN

Department of Philosophy
Saint Mary's University
Halifax, Nova Scotia, B3H 3C3
Canada
sheldon.wein@gmail.com

1. INTRODUCTION

In "Productive *versus* Destructive Cooperation" I show that, simply by slightly enriching the available information beyond that which contemporary game theory usually admits, we can easily adopt standard game-theoretic models that will display in a concise and comprehensive manner key features of the nature of some of our more pressing social problems. Radu Neculau, in his insightful commentary, raises several difficult philosophic questions about the basic assumptions on which contemporary game theory (and the suggestions in "Constructive *versus* Destructive Cooperation") rests. Since his commentary is nicely organized around three topics, I will use his headings in my reflections. I should note, however, that the concerns Neculau raises are serious and I cannot do them justice here. Furthermore, even were we to conclude that the ground on which game theory rests is faulty, this should not prevent us from making use of the devices game-theorists have developed (and which I have suggested modifying) when such uses will help us easily achieve a clearer understanding of how to deal with various social problems.

Perhaps this is an appropriate analogy. Suppose we were to find that the philosophic objections to bi-valence were greater than the arguments in its defence. This would certainly give many of us pause. But it would be no reason to abandon, say, truth tables for those cases where they can elegantly illuminate the character of various logical connectives. So, even were we to find the conception of rationality used by contemporary decision theory to be in some respects inadequate, this should not keep us from utilizing—at least in those circumstances where we can do so without misrepresentation—the many ingenious models game-theorists have developed.

2. EVALUATION

In the course of refection on what we should do, given our values, we frequently come to think we should change our values (or to realize that our values are not what we thought they were). Can game theory help us evaluate our values? I think it is obvious that it can.

Sometimes one has what one wants. Other times to get what one wants one must change things. Sometimes one has to change other parts of the world (perhaps including other people and their desires and attitudes) to get what one wants. Still other times the best part of the world to change is one's own self—be it changing one's conception of rationality, or adopting new preferences, or altering one's values. Game theory helps people recognize when what they need to change is their circumstances or values or the

Zenker, F. (ed.). *Argumentation: Cognition and Community: Proceedings of the 9th International Conference of the Ontario Society for the Study of Argumentation* (OSSA), May 18-21, 2011. Windsor, ON (CD ROM), pp. 1-4.

values of others (or, as is most typical, all three) to increase the probabilities that our actions will have outcomes that are Pareto-superior to the status quo. Knowing which sort of situation we are in—whether we are in say a productive PD or a destructive PD or some form of assurance game, to take the examples used in my paper—can help us decide whether we should change our values *and* what social devices (a shared religion? new morals? a renewed commitment to our tribe? team spirit? legal sanctions? or etiquette?) we should use to reinforce and promote our newly acquired values. How good is decision theory alone at helping people make collective decisions? Likely it is not very good. (Try standing on a cold street corner with a group of decision theorists trying to make a collective decision about which restaurant to go to—and whether to wait for X before leaving for the restaurant—if you wish to confirm this!) Furthermore, decision theory helps us see the ways in which such questions are empirical questions. An example: Since developing team spirit is obviously a good way to ensure that everyone cooperates in assurance games, one might think—as I confess I did—that having people play team sports likely enhances moral development. But Catherine Hundleby pointed me to Shannon Hogarth's survey of research, which shows that my simplistic view was, well, simplistic. Matters are much more complicated than I had supposed.

3. WILL FORMATION

How can those who adopt the minimalist account of rationality favoured by contemporary decision theory ever hope to account for the fact that we humans are able move from the "I-attitude" to the "we-attitude" (as Neculau so nicely puts it)? This is a real difficulty, and my suggested modifications to what information decision theorists should be allowed does not address this problem.

To see that it is a real problem one need look no further than the trouble Hobbes has in this regard. Hobbes, who along with Descartes is making a radical break with the scholastic tradition, defines the will by saying "In Deliberation, the last Appetite, or Aversion, immediately adhaering to the action, or to the omission thereof, is that wee call the WILL; the Act, (not the faculty,) of *Willing*" (*Leviathan*, Chapter VI, § *The Will*). (He goes on to observe that, given this view of individual wills, animals that deliberate must be said to will. He objects to the scholastic view that the will is "*Rationalle Appetite*," observing that this makes voluntary action against reason impossible.) This account of the will leads Hobbes to hold that "a Commonwealth hath no Will, nor makes no Lawes, but those that are made by the Will of him, or them that have the Soveraign Power" (*Leviathan*, Chapter XXXI, § *All Attributes depend on the Lawes Civil*). If this were the best we could do to explicate (and justify) human capacities in this regard, we would face the difficult choice of either abandoning our conception of rationality or holding that many values we hold dear are chimerical.

Happily we can do better. Recently, Scott Shapiro—by combining insights from H. L. A. Hart's critique of Austinian/Hobbesian legal theory with Michael Bratman's work on planning and shared cooperative activities—has developed a promising account of how rational authority could develop even given only a minimalist account of human rationality. (Shapiro's *Legality* is primarily concerned with explaining and justifying legal authority—legal systems are, according to Shapiro, shared plans we use to solve moral problems whose solutions are complex, contentious, or arbitrary—but the prospect of extensions to the we-attitude in other areas seems viable.) So we have the basis for seeing how "a new and different individual will emerges, as well as a collective one".

4. INTEGRATION

Neculau reminds us that social integration, the attainment of what Rousseau in his brilliant and maddeningly elusive way called the general will, can be reached "by collectively pursuing the same object or by teaching individuals how to harmonize their interests and speak with a universal voice. The first one requires the presence of assurance providing institutions of rule enforcement, or authority. The second one demands the affirmation of a different kind of social autonomy, or communicatively playing a re-assurance game." Neculau asks where I stand on this matter.

This and related problems have long vexed me. In Wein 1987, I argued that even pure altruists face prisoner's dilemmas. (In effect, I was arguing that the second way is not sufficient.) And in Wein 1997 I argued that, given recent changes in technology, the first option was looking more like the second. More recently (Wein, 2007a), I argued from a position that assumes the two ways of looking at these things may not be as distinct as we usually suppose. In fact, it useful to remember that in the stag hunt story, Rousseau actually has his hunters adopt a conception of rationality which is more minimalist than even that Hobbes's contractors have in the *Leviathan*. (For an argument that Rousseau's stag hunters have a conception of rationality closer to the minimalist conception adopted by contemporary game theory, see Wein 2007b, and for an argument that Hobbes's agents are closer to embodying Rousseauian autonomy than is usually supposed, see Venezia, forthcoming.) So, perhaps there is no answer to where I stand. It might be better to say on this issue I am inclined to stagger!

5. CONCLUSION

The main point of "Productive *versus* Destructive Cooperation" is to show that, by slightly enriching the standard strictures contemporary game theory adopts, we can easily use some of its devices to more simply and accurately represent some of the problems we solve. I do not claim that sophisticated game-theoretic accounts of the situations I discuss can be provided. But I do think that such accounts will be so complex that my "less pure" models will be more illuminating. Consider just one example. The government of Canada has recently floated the idea of allowing couples to "income split" when calculating taxes. This has been presented as a unifying and quasi-egalitarian idea, but, as the case of Anne and Andrew shows, it would likely have the effect of preserving just those sorts of "family values" most of us hope to undermine. Further use of enriched models of the type I suggested could serve to make it easier for everyone to understand the nature of the sorts of choices we face and, consequently, for more people to have informed input into the conversation we need in order to make our society a genuine cooperative venture for mutual advantage. Thus I continue to urge those who seek ways to represent the circumstances in which our collective action problems arise to reconsider before blindly accepting the severe epistemic constraints that decision theory endorses.

ACKNOWLEDGEMENTS: I thank the Ontario Society for the Study of Argumentation for organizing such an interesting conference, my audience for probing questions (and especially those who continued to press and help me well after the formal session had ended), and Radu Neculau for sharing his insights on these difficult matters. I am grateful to the Dean of Arts at Saint Mary's University for generous financial support.

REFERENCES

Hogarth, Shannon (2008) *Moral reasoning ability in Canadian interuniversity athletes*. (MA thesis) Windsor: University of Windsor.

Shapiro, S.J. (2011). *Legality*. Cambridge, MA: Harvard University Press.

Venezia, L. (forthcoming) *Political Authority and Political Obligation in Hobbes's Leviathan*. (Ph. D. thesis) Tucson, AZ: University of Arizona.

Wein, S. (1985). Prisoners' Dilemmas, Tuism, and Rationality. *Simulation and Games*, Volume 16. # 1.

Wein, S. (1997). Feminist Consciousness and Community Development. *The International Journal of Social Economics,* Volume 24, Issue 12.

Wein, S. (2007a). Urbanization in the Developing World. *The Dalhousie Review*, Volume 87, No.1, Spring.

Wein, S. (2007b). From MADness to SANity. *Peace Review: A Journal of Social Justice*, 19 (2).