Private Rationality and Public Reasonableness: The Rational Interactor in Game Theory and the Law*

Bruce Chapman
Faculty of Law, University of Toronto
bruce.chapman@utoronto.ca

I. Strategic and Rational Interactions

Game theory offers us an account of how rational agents interact in strategic situations. The situations are strategic in that the determination of rational conduct for any one agent will depend upon what that agent believes another agent in the interaction might do. Of course, this other agent will also typically be working out her own strategy in light of the beliefs she has about the first agent and what the first agent might do. Thus, a rational agent needs to think not only about what to do, but also about what another agent is thinking about what to do and, further, about what that other agent is thinking about what the first agent is thinking about what to do. And so on. In non-cooperative game theory this thinking, while quite sophisticated, and having as its subject matter the thoughts and actions of another, remains very private. Each agent typically works out what to do, or how to interact with another, as a matter of private rationality.

All this is familiar. What is less familiar, perhaps, is that the law and legal theory also contain an account of how rational agents interact. However, it is an account of how rational agents interact, and how they understand their interaction, under the idea of public (or objective) reasonableness. More straightforwardly, we might say that law and legal theory offer an account of how "reasonable persons" interact. However, while this terminology might sound more familiar, it runs the risk of suggesting only that law adds some normative assessment of a rational agent's conduct into the mix, something in

* *forthcoming* **in American Philosophical Association Newsletter *Philosophy and Law* (Spring 2004)**

which non-cooperative game theory, as a predictive or descriptive tool, would not claim to have much of an interest. But in this paper I want to emphasize the public or objective nature of reasonableness as something distinct from private rationality, and argue that, in taking this more public orientation, the legal account provides a different interpretation of what an interaction between rational agents *is*, not simply how it is assessed. Indeed, I shall argue that the law's account has important advantages over the game theoretic version for *explaining* the levels of coordination and cooperation that we observe amongst rational agents.[1]

Effectively I will be arguing that while game theory can claim to offer an account of how rational agents interact, unlike law and legal theory, it cannot make the further claim that it provides an account of rational interaction. As I will be suggesting this is a crucial deficiency in game theory, it is worthwhile having some sense early on of what I mean by the distinction. Certainly a good deal of highly sophisticated reasoning and rationality is exercised by game theoretic agents. The thoughts that occur to rational players in a strategic game ("I think that she thinks that I think, etc.") are often so complicated that it is sometimes difficult even to articulate them. Moreover, the subject matter of each player's thoughts is the thoughts of another, and so in this sense the reasoning might even be thought to be "social".[2] But how "social", or public, are they, really? A player's rational thoughts, once replicated, are "socialized" in a sense, in that they become the thoughts of all the other players at some level. But in another sense what is replicated or socialized is still only the stuff of *individual* thoughts. When I am thinking about what you are thinking, there is definitely something *inter*subjective going on. But, as a different point of emphasis, we have also to concede that our mutual

intellectual engagement is still only inter*subjective*. Our private thoughts are overlapping, perhaps, in that they are "of each other" or "of each other's thoughts", but there is no real rational *interaction*, at least if we mean by interaction the idea that we meet one another in some public space.

Of course, the game theorist will say that interaction is at the very core of the theory of games. The players interact when each chooses a strategy, say a row or column in one of the familiar matrices depicting a two person game, and these different choices combine, or interact, to produce a final outcome, that is, some given cell within the overall array of possibilities. What could be more interactive than that? But my claim is that, however interactive this may be, it is not an account of rational interaction. For when there is interaction between players in the theory of games, it is causal rather than conceptual; each player simply determines a part of the world (a row or a column) as a matter of individual choice, and the determination of each part determines the whole as a causal matter. But there is no interaction in a shared conceptual space. So when there is interaction it is not *rational* interaction. On the other hand, when there is something rational, the rational is, perhaps, intersubjective, but it is not interactive. The players think, of course, and even think through each other's private thoughts, but they never think *together* in some more public, or objective, conceptual space. Thus, when there is something rational, it is not interactive, and where there is something interactive, it is not rational. There is never, therefore, any moment where there is a rational interaction.

I am sure that all this will continue to strike some readers as somewhat mysterious, especially the idea that agents can "think together", something that risks conjuring up the (non-sensical) notion of a "group mind".[3] Perhaps it will help to see

how the law provides the very sort of account of a rational interaction that I think is lacking in the theory of games. I turn to this in the next section. With this understanding of the law's account in hand, we will be in a better position in section III to see, by way of some simple examples, how such an account can help in the theory of games. The paper concludes in section IV.

II. The Objective Standard of Reasonable Interactions

In his important discussion of common law liability for unintentional harm, Oliver Wendell Holmes reminds us of the special sort of interest that law has in the idea of individual responsibility, an interest that it does not always share with ethical theory. Holmes asks us to consider the defendant in a tort action who has done the best he can to avoid injury to the plaintiff but, because of his particular ineptitude, has not been successful in doing so. On any ethical standard, Holmes suggests, it would be hard to fault the defendant for injuring the plaintiff; how can there be moral fault in doing everything one can to avoid such an injury? Yet, in a passage that is now well known and much quoted for its rejection of the relevance of these subjective abilities for a judgment of legal fault, Holmes remarks:

> If…a man is born hasty and awkward, is always having accidents and hurting himself or his neighbors, no doubt his congenital defects will be allowed for in the courts of Heaven, but his slips are no less troublesome to his neighbors than if they sprang from guilty neglect. His neighbors accordingly require him, at his

> proper peril, to come up to their standard, and the courts which they establish decline to take his personal equation into account.[4]

The neighbors' standard, of course, is the objective standard of the reasonable person, and, while Holmes thought (in his typical fashion) that such a standard was ultimately justified because it was more conducive to the public welfare, the more general point was that law concerns itself with what is appropriate as a standard of behavior for the man *in interaction with his neighbors* rather than what is fair as a subjective matter to the man considered on his own.

This difference in what is the proper concern of law as distinct from ethics is also, of course, much emphasized by Kant. In his *Doctrine of Right*, or that part of *The Metaphysics of Morals* which deals with his philosophy of law, Kant argues that law concerns itself only with what he calls the "universal principle of Right", or the coexistence of everyone's *freedom* in accordance with a universal law. "[A]nyone can be free", says Kant, "as long as I do not impair his freedom by my *external action*, even though I am quite indifferent to his freedom or would like in my heart to infringe upon it." By contrast, "[t]hat I make [the universal principle of Right] my maxim to act rightly is a demand that ethics makes on me."[5]

Thus, across a broad range of theories, and over a significant span of time, there has been agreement that law (or, at least, private law) attends to what is right between the parties linked by an interaction rather than what is right or ethical as a matter internal to one of the parties. Law is not concerned with a person's thoughts in so far as they do not impact upon another, that is, in so far as they are not acted upon and do not have any

potential for interaction. And when they *are* acted upon, and have some interactive effect, what does matter for law is not what an individual might mean to do as a private matter, but what she does as a public matter, that is, what she (actually) does under a publicly accessible, or (objectively) reasonable, understanding.

This is true for law in general, but it is particularly important for contract and consent, where parties set out to do something together, that is, when they choose to engage in cooperative activities.[6] For example, whether two parties have a contract for the sale of, say, "new oats" or "old oats",[7] will not depend on whether there is a meeting (or overlap) of their (private) minds on this issue. Rather, the court will attend to the most plausible public understanding of the transaction and deem the contract to be for "old oats" if that is the most (objectively) reasonable meaning of its terms in the context in which contracting occurred. Indeed, even where subjectivity does seem to be important, for example, in the criminal law, what the accused will have to attend to as a subjective matter (as a matter of, say, honest belief implicating subjective states of mind like intent, knowledge, recklessness, etc.) will be a public or shared (objective) understanding of what the concept of right conduct requires. Thus, it will not be enough in the case of a sexual assault, for example, for the accused to say (even honestly) "I thought (in my mind) that *in her mind* she was consenting" if there was no reasonable, or public, manifestation of that consent.[8] The accused's appeal to his thoughts about her thoughts, while exemplifying the same intersubjectivity that is characteristic of strategic thinking in the theory of games, is inadequate because for law the subject matter of his (subjective or honest) belief is insufficiently objective. What the accused needs to be able to establish is

that he had an honest belief in a reasonable manifestation of her consent, that is, he needs to be thinking about her consent as a publicly comprehensible matter.[9]

This means that, under law, two parties who are acting together will have the separate individual actions that make up their cooperative activity linked conceptually under some objective or public understanding. Thus, if each party is to understand what her separate obligations are, or what (in law, at least) she should do, she will have first to consult that shared understanding of *what it is* that they are doing together ("Is this a contract at all?" "Is this a contract for old oats?"), and only then ask what she should do under that shared understanding of the cooperative venture. This may seem obvious enough, but it is important to appreciate that, unlike for game theory, this *does* mean that the interaction of the parties is a rational interaction. Each party orders her individual action in the cooperative venture as a part of a conceptual whole that likewise also orders the individual action of the other party. In this sense the parties first meet or interact together in some common conceptual space. Of course, it will be the individuals themselves who ultimately act out their respective parts of the cooperative plan as understood in this public way, and at this point the interaction (part with part) might appear to be only an interaction between rational individuals (as in the usual game) and not a rational interaction. But it is precisely because each action (while individual in the causal sense) is ordered by a shared or public conceptual scheme (part with whole) that the interaction of these (rational) individuals is a rational interaction or, as law would articulate this idea, a *reasonable* one. It now remains to see how this idea of a reasonable interaction might be helpful in the theory of games.

III. Coordination and Cooperation: Rational and Reasonable

Consider the following very simple two person game in which two friends, Row and Column, would like to meet for lunch at one of two restaurants, *Andy's* or *Bob's*. Unfortunately, they have made no prior arrangement and must choose without the benefit of having already agreed about where to go. In Figure 1 this choice is represented for each of the friends as the choice between A (for *Andy's*) and B (for *Bob's*), with Row choosing between the two rows and Column between the two columns. The payoffs to each friend are indicated by the numbers in each cell of the matrix, with the first number being the payoff to Row and the second the payoff to Column. Each friend is assumed to know this representation of their situation and that each is rational. Further, all of this knowledge is assumed to be *common knowledge* (that is, not only does each know the game form representation and that he or she is rational, but also each knows that each knows this, each knows that each knows that each knows this, and so on).[10]

Figure 1

		Column A	**Column** B
Row A		4, 4	0, 0
Row B		0, 0	2, 2

It might seem that there is a reasonably obvious thing for each friend to do here. Since they both much prefer to lunch together at A to lunching together at B, then each should choose A. But, strictly speaking (at least according to strategic reasoning), that is not what the game form matrix shows. Row should indeed choose A, *but only if Column*

also chooses A. Otherwise, she should choose B. For her to choose A when Column chooses B would result in one of her two *least* satisfactory outcomes. The problem, of course, is that she does not know what Column has chosen to do. Indeed, as she thinks about it a little more, she will conclude that Column will not yet have chosen a restaurant himself. For he is working out the same problem at his end and he too is stymied; alas, until she chooses, he cannot choose either! This is the impact of seeing the situation as one of strategic interdependence.

Notice that there are views that Column could have about Row which would avoid any problem for Column in working out what to do. For example, if it was Column's view that Row would simply, and somewhat thoughtlessly (compared to Column), go to the restaurant where she most preferred to lunch with Column, namely at A, then Column would have no difficulty conditioning on Row's choice of A and would choose A himself. Note that Row does not have to actually *be* this sort of parametric thinker; it is enough that Column *thinks* she is. And if Row knows that Column thinks this of her, then she too will head for A.

But, under our assumptions, none of this is possible because Column gives Row more credit as a rational actor than that. Column recognizes that Row is every bit as rational as he is and, therefore, in this mirror image situation, he imagines her to be working out a mirror image problem. In this limited respect, therefore, Column thinks of his own reasoning as "social" (or at least "socialized"); it is replicated by other rational agents in the same situation, and that replication is reintroduced into his own thinking as something that he thinks about. Indeed, that is at least partially responsible for what is so paralyzing here. For when he thinks that like-minded Row is thinking about what he is

thinking and, further, is thinking about what he is thinking that she is thinking, he realizes that there can, as yet, be no choice by Row (for example, to choose A) upon which he can condition his own choice (to choose A).[11]

However, consider the following variation on this game, a variation that allows Row to unambiguously choose A as a rational matter, that is, in a manner consistent with the rationality assumptions that are typical of game theory. In Figure 2 the payoffs are changed for Row and show that she would like to eat at A regardless of what Column might choose to do. Of course, just as in Figure 1, in Figure 2 Row would most prefer to eat at A together with Column. Column's payoffs are unchanged; he would prefer most to eat together with Row at A, but failing that, just as before, he would prefer to eat with her at B rather than eat alone.

Figure 2

		Column A	**Column** B
Row A		4, 4	3, 0
Row B		0, 0	2, 2

In this game Row will head for A regardless of what Column chooses to do. We say that Row has a *dominant* strategy to choose A. Moreover, Column (knowing the structure of the game form matrix) knows this of Row and, therefore, knows that he should go to A to meet her. (The game is solved by the method of *iterated dominance*; first, Row chooses A by strict dominance and, second, Column chooses A, since, given that Row chooses A, A now dominates B for Column as well.) Moreover, while Row need not think all that

strategically in order to decide what to do, Row likewise knows all this of Column, and so can predict quite easily, and happily, that, as she heads for A, she will meet Column there.

But now consider a variation of the Figure 2 game that has a much less happy result. In Figure 3, the game has begun to look a bit more like a (one-sided) prisoner's dilemma. The payoffs for each of the two friends at A have been reduced so that they now prefer, when they lunch together, to lunch at B. That is, while Row continues to prefer to lunch at A regardless of what Column chooses, when she lunches together with Column, she prefers to lunch at B. (Perhaps she considers the dishes at B are easier to share.) Column continues to prefer to lunch together to lunching alone, and, when lunching together, like Row, he prefers to lunch at B.

Figure 3

		Column	
		A	B
Row	A	1, 1	3, 0
	B	0, 0	2, 2

Can these two friends rationally get to *Bob's* together? It seems not. The same reasoning that took each of them to *Andy's* in Figure 2 takes each of them to *Andy's* in Figure 3. Given a choice of A by Column, Row would choose A; and given a choice of B by Column, Row would choose A as well. So she must, surely, choose A. What (at least in game theoretic terms) could possibly make her choose otherwise? And, given that she chooses A so rationally, how can Column rationally choose other than to go to A as well?

Thus, Row will go to A regardless of what Column chooses, and Column, knowing this of Row, will go to A to meet her.

While the result is an unhappy one for the two friends (that is, it is Pareto-inferior to the outcome where they lunch together at *Bob's*), it is hard to deny that it is the rational result for them given the structural nature of strategic reasoning. Strategic reasoning, it will be recalled, conditions the rationality of a choice of action by one agent on a choice of action by the other. Put differently, but equivalently for the game form matrix in Figure 3, the rationality of an action, say, the choice of a row by Row (or a column by Column), is the rationality of that choice of row (or column) *given* a certain column choice by Column (or row choice by Row). And, surely, it is tempting to ask, yet again: How could it be otherwise? Rationality for each of the two friends goes to the rationality of his or her individual actions, where Row chooses row by row, and Column chooses column by column. There can be no question of what it is to choose rationally in anything other than a strictly vertical (row by row within a given column) or horizontal (column by column within a given row) direction. Choices in a *diagonal* direction (or choices for Row outside a given column or for Column outside a given row) are simply not available *as choices* for any one individual.[12] Thus, if rationality or reasoning is to govern choices, and choices are ultimately made by individuals, then it is natural to think that the rationality of an action for one individual is the *strategic* rationality of that action *given* the choice of action by another.

However, while natural enough, strategic rationality is not the only way to think about how rationality and reasoning might govern individual choices in cases of social interaction. An alternative form of reasoning, one that is exemplified, I think, in the law's

idea of a reasonable interaction, while conceding that choices are ultimately made by individuals, might not assess the rationality of these individual actions directly, but could derive their rationality from a prior assessment of the rationality of a *pattern* of actions for the group of individuals taken as a whole. In other words, the rationality of a part – some individual action – is derived from the rationality of the larger whole of which it is a part.[13] I will (hoping that I can avoid some of the baggage that comes along with this term) refer to this alternative way of thinking or reasoning about one's individual action as holistic. But, as already suggested from our discussion of the law, one might also want to label this sort of thinking as reasonable.

This alternative form of holistic reasoning about individual choices, even when replicated across all the agents in identical situations within some social interaction, has important implications for the problems our two friends have been confronting in choosing to meet at a restaurant. Consider first the pure coordination problem that we observed earlier of our two friends, Row and Column, in Figure 1. The problem of coordination arises under strategic reasoning because what is rational for each agent to do depends so crucially on what the other agent chooses to do. Under strategic reasoning, each agent asks: what should *I* do (given my thoughts about what the other agent might do)? However, under the alternative more holistic approach, Row and Column each ask themselves, first: what is it that *we* should do here? (Each answers easily: We should meet. Where? At *Andy's*.) And then they ask themselves, second: and what is it that *I* do when we do that? (Just as easily each answers: I should choose A.)[14]

Of course, the first question does require each of the two friends to make a *diagonal* comparison in Figure 1, a comparison that neither friend can immediately

follow through on as a matter of individual choice. But the second question returns each to the issue of what action he or she should choose as an individual. In particular, each friend should choose that row or that column which allows each to do his or her part in the achievement of the outcome deemed most rational (here, meaning best in terms of preferences) for the group as a whole. However, while individual actions must *ultimately* follow the strict vertical and horizontal contours of the game in this way, a *prior* assessment of what is *rational* in individual action need not. It is a peculiar feature of the sort of reasoning that game theory contemplates of interacting agents that it simply assumes that the empirical requirements of individual action must somehow be reproduced in an individual's prior assessment of that action's rationality. But there is no obvious reason why our conceptual world should track what is possible in the causal world in just this way. The causal world or, more particularly, how individuals act upon or interact with each other in some shared physical environment, could be informed by, or track, some independent judgment that we make of our actions (together) as a purely conceptual matter.[15]

Consider now the problem in Figure 3, which is not, strictly, a pure coordination problem, although it shares the feature with Figure 1 that both friends would very much like to find a way to lunch together at one of the two restaurants rather than the other. Again, strategic reasoning identifies a dominant strategy for Row to choose A, and then, given (the predictability of) Row's choice of A, Column chooses A as well, this despite the fact that both friends can see this coming and would be better off each choosing B. However, under the alternative more holistic approach to reasoning through their predicament, each friend again asks himself or herself: what is that we should do here?

The answer for Column would appear to be as easy as it was before, only this time it identifies a different restaurant: We should meet, says Column. Where? At *Bob's*. (Whether the questions and answers are in fact as easy as this for Column, who must also, even under holistic reasoning, take into account how the situation has changed for Row, is a question I will return to momentarily.) But what about Row? Given her preferences over the four possible outcomes in the game, Row has some difficulty even articulating a common conception of what the two friends might do. She seems not to be able to say easily that "We should meet for lunch. Where? At *Bob's*", since her most preferred outcome is, first, to eat *alone* and, second, to do so at *Andy's.* This, after all, is one of the crucial differences between Figures 1 and 3. However, she also cannot easily say that "We should eat alone", as this contemplates the possibility that she might eat alone at *Bob's*, her *least* preferred outcome. What she has to say in answer to the question about what "we" should do is something less categorical and more particular, something like "We should eat alone, but only if I eat alone at *Andy's* and you at *Bob's*; otherwise we should eat together at *Bob's.*"[16]

This obviously leaves something to be desired as a categorical call to rational action for the two friends.[17] Moreover, it fails in this respect for two reasons. First, the call to action is so infused with particularity it hardly seems categorical at all; there really is no common conception here that informs action by each of the two friends. In other words, there is no common conception of their interaction (as reasonable) that informs their individual actions as rational (that is, that renders them "sensible" under some common conception), and it is this that we were seeking under the alternative more holistic approach. Second, given his preferences, such a highly particularized call to

action by Row is hardly likely to be acceptable to Column; after all, it calls for an interaction which results in one of his least preferred outcomes. Moreover, under our common knowledge assumptions, Row can anticipate this reaction from Column. Therefore, Row can anticipate that Column will fail to find this call to action rational, at least in the holistic sense, not only because it seems so unorganized by general categories of thought that they can share, but also because, even as a particularized call to action, it is so contrary to the preferences that Column has.

However, it is only the first deficiency that I really want to emphasize here. In part this is because one can imagine Row also saying to Column that *his* categorical call to action (to lunch together at *Bob's*) is very contrary to *her* preferences. Indeed, we could intensify her preference for lunch alone at *Andy's* by giving her a payoff of 5 rather than 3 for that outcome, all without changing the decisive structure of the game in Figure 3 (i.e., from the point of view of strategic reasoning the game would still have the same "solution" under iterated dominance), and then Row could say of Column that his categorical call to action is *more* costly to her than her particularistic call to action is costly to him (supposing that these payoffs were cardinally comparable utilities and that costs are measured as departures from each friend's most preferred outcome). So this contrary-to-preference notion of a non-rational call to action cannot do much work to distinguish Row from Column. But the non-categorical or highly particularistic nature of Row's call to action *does* seem to distinguish her from Column. Column's call to action can organize the respective contributions of each friend to the overall social interaction under some general category of thought that each can share ("What am I doing in choosing B? I'm doing my part in us getting together for lunch at *Bob's*."). But in Row's

call to action ("We should eat alone, but only if I eat alone at *Andy's* and you at *Bob's*; otherwise we should eat together at *Bob's*."), there seems to be no such general category or concept that can pose as the "whole" of which action by each friend is some part.

Notice too, if we return to the situation where Row's preference for lunching alone at *Andy's* is intensified to the point where she has a payoff of 5 in that outcome, that while the decisive structure of the game again remains the same from the point of view of strategic reasoning (i.e., it is solved by iterated dominance), now there *is* a holistic call to action that could be quite categorical for Row, and it is a call to action that would avoid the Pareto-inferior outcome where both friends lunch together at *Andy's*. Now she could say, "We should each choose our restaurant such that total overall utility is maximized," and then, by choosing A, go on to do her part in that shared conception of the friends' interaction (again supposing that these payoffs are interpersonally commensurable cardinal utilities). Moreover, if Column shared that conception of their interaction, then he too would know to do his part by choosing B. (This illustrates how Column's call to action under a shared conception of the interaction must take into account how the situation might change for Row; even though his own payoffs are unchanged after the payoff to Row is changed, the change in payoff to Row makes available to Column, as much as to Row, a *shared* categorical conception of the interaction, where Row eats alone at *Andy's*, that was not available before the change.) Again, this indicates that what does the real work here in a categorical or holistic call to action, and what avoids the Pareto-inferior outcome where each friend chooses A, is that *is* holistic or categorical, not that it is consistent (somehow) with the preferences of all the interactors.[18] After all, under an interaction where the shared conception is that each

friend does his or her part in the maximization of total overall utility, a shared conception that also helps these friends to avoid choosing the Pareto-inferior outcome, the result can be, as it is here, significantly contrary to Column's preferences. Note too that the result under this shared conception of their interaction is not merely a coordinated choice *between* Nash equilibria, but, more strongly, a coordinated choice of a *non*-Nash equilibrium.

These variations in the examples suggest, not surprisingly, that different shared conceptions of an interaction are available, and that these different shared conceptions will differently inform agents about what each should rationally (and individually) do (as his or her part) under that shared conception. The possibility of quite different shared conceptions will also suggest that there is a good deal of indeterminacy for an individual agent in thinking about what he or she should rationally do, something that might have one wondering whether our two friends are much helped, in their attempts to get to lunch together, by these more holistic ideas. The skeptic might conclude that the problem now is only that our two friends need to coordinate their sense of "shared conceptions", a problem that does not seem, perhaps, all that much more easy to solve than the problem of coordinating their individual actions more directly.

I do not intend to argue in any length against that criticism here, although even this conclusion does suggest that our social institutions, including our legal institutions, will have a quite different problem of coordination to address from what is more conventionally supposed.[19] However, it does strike me that this skeptical conclusion is also very likely wrong. There will be fewer degrees of freedom for individual action under the thought that one's (preferred) action must be disciplined, first, by a conception

of what it is that one is doing and, second, by a conception of what it is that *we* are doing, that is, by a conception of our interaction that is capable of being sensibly *shared*. The different categories and concepts that organize the actions of different individuals will have to fit together in a "sensible" way, and if these categories and concepts are shared, then it seems reasonable to think that the actions of the different individual actors will have to fit together in a sensible way as an interaction. At a minimum, if there are fewer degrees of freedom for rational individual action under shared conceptual schemes than there are under individual preferences, even the common knowledge of such individual preferences, then one might expect fewer problems of coordination and, further, as our examples have suggested, fewer problems of cooperation than the standard game theoretic literature suggests.

However, even if the skeptic were to concede this point as a theoretical matter, he might still want to argue that there is little evidence that rational agents actually interact in this way rather than the more strategic way that is contemplated by the game theorist. However, here I can be more definitive in my reply. There is good experimental evidence that agents in game theoretic situations actually do reason this way. Consider, for example, what Eldar Shafir and Amos Tversky discovered about how subjects play the familiar two-person prisoner's dilemma game, a game very similar to what appears in Figure 3.[20] In their experiment the rate of cooperation in the prisoner's dilemma was 3% when the subjects knew that the other player had defected, and 16% when they knew that the other player had cooperated. One might well have expected some rate of cooperation between 3% and 16% when the subjects were uncertain whether the other player had cooperated or not. However, when the subjects were confronted with this uncertain

situation, the rate of cooperation rose significantly to 37%, a number that cannot even be rationalized as some weighted average between the strategically formulated actions "cooperate given that the other cooperates" and "do not cooperate given that the other does not".

Shafir and Tversky attribute this pattern of responses to the different *conceptions* or *understandings* that a subject will have of her choice situation depending on whether she knows if the other player has already made his choice of strategy. When she knows that the other player has already chosen his strategy, whether it is to cooperate or not to cooperate, the subject thinks of herself as acting "on her own". *Given* the choice of the other player, she alone will determine the final outcome of the game. This encourages her to bring a highly individualistic perspective to bear on her choice of strategy, a perspective that leads her more naturally to choose against cooperation. However, in the uncertain situation, where all four possible cells of the prisoner's dilemma game are still very much in play, the outcome of the game is to be determined by a combination of the strategy choices of both players taken together. Shafir and Tversky argue that this provides for a more collective understanding of the situation, and from this more collective point of view the optimal strategy for both parties is to cooperate. Thus, say Shafir and Tversky, it is less surprising that cooperation is chosen more frequently in this situation.

These results support the argument that individual agents in a game theoretic situation behave differently depending on how they *conceive* of the choice they are being asked to make. So there seems to be more than preference and information variables involved in these choices. Moreover, these differences are such that when the choices are

presented to them in the causally interactive way that game theory most often contemplates (that is, when the choices are presented to them under the conception "this is still to be determined by our combined actions", or "it's up to us", rather than "this is to be individually determined" or "it's up to me"), the individual actors think about what to do as something that they should do together. That is, each thinks first under the diagonal comparison "what should we do?", and then lets that judgment inform how he or she acts individually. In this section I have tried to suggest, by way of some simple examples, that this can help these individuals both to coordinate their actions and to avoid certain non-cooperative outcomes that might make them all worse off.

IV. Conclusions

The analysis that I offer here is highly preliminary. The details are certainly far from worked out, and the approach may fail fundamentally, either as a way to understanding better why people coordinate and cooperate in game theoretic situations, or as an account of what is a reasonable cooperation in the law. I hope only to be suggestive about what might be possible under a quite different approach, one that law already makes available. But if I am right in making the suggestion, the returns are large. Not only are we closer to understanding why people coordinate across multiple Nash equilibria, that is, in the cases of pure coordination games or coordination games where there might be some conflict of interest (e.g., the game of chicken or the battle of the sexes), but also we might be able to understand better why people choose to cooperate and coordinate around non-Nash equilibria, something that challenges game theory as an account of social interactions at a very fundamental level indeed.

[1] Some of the empirical (experimental) literature that shows people have a tendency to coordinate and cooperate at higher levels than game theory predicts is nicely surveyed in John H. Kagel and Alvin E. Roth eds. *The Handbook of Experimental Economics* (Princeton: Princeton University Press, 1995).

[2] Ariel Rubinstein defines game theory as "an abstract inquiry into the concepts used in *social reasoning* when dealing with situations of social conflict"; see Ariel Rubinstein, "Comments on the Interpretation of Game Theory" *Econometrica* 59 (1991): 909, 909 (emphasis added).

[3] Philosophers are typically very careful to avoid any suggestion of a shared mental state, even when their arguments appear to bring them close to the possibility. For example, in a paper discussing 'shared intention', Michael Bratman cautions that "a shared intention is not an attitude in the mind of some super agent." See Michael Bratman, "Shared Intention" *Ethics* 104 (1993): 97, 99. John Searle makes a similar point in his essay "Collective Intentions and Actions", in Philip R. Cohen, Jerry Morgan, and Martha E. Pollack eds. *Intentions in Communications* (1990), at 404: "[T]alk of group minds … [is] at best mysterious and at worst incoherent. Most empirically minded philosophers think that such phenomena must reduce to individual intentionality."

[4] Oliver Wendell Holmes, *The Common Law* [1881] (Boston: Little Brown, 1963), at 86-87.

[5] Immanuel Kant, *The Metaphysics of Morals* [1797] (Mary Gregor trans.) (Cambridge: Cambridge University Press, 1991), *231.

[6] Arthur Ripstein, *Equality, Responsibility, and the Law* (Cambridge: Cambridge University Press, 1999), 201-17.

[7] See Smith v. Hughes (1871), L.R. 6 Q.B. 597 (a buyer refused oats that the seller had delivered on the grounds that he had meant to buy old oats, whereas the oats delivered were new). For excellent discussion of the case under the idea of a public rather than merely intersubjective understanding ("the meeting of minds" idea), see Brian Langille and Arthur Ripstein, "Strictly Speaking – It Went Without Saying" *Legal Theory* 2 (1996): 63, 76 .

[8] The facts of *Director of Prosecutions v Morgan* [1975] 2 All ER 347 (HL) come to mind. After an evening out, Morgan, the accused, invited some of his friends to come back to his house and have intercourse with his wife. Somehow, contrary to the truth, he managed to convince them that she would likely feign her resistance, but that she would be willing nevertheless. One issue was whether the friends' honest beliefs in her consent, no matter how unreasonable in the face of her manifest unwillingness, were enough to undermine the mens rea requirement for their assaults. The House of Lords, reversing the Court of Appeal on this point (although upholding the conviction of the accused), held that they were. Needless to say, the House of Lords decision on this point of law has been controversial. For good discussion, see E. M. Curley, "Excusing Rape" *Philosophy and Public Affairs* 5 (1976): 325-360.

[9] Sometimes this is expressed as a requirement that his belief in her consent be honest *and* reasonable. But that is problematic because it obscures too much the distinction that properly exists between the fault standards that are appropriate for criminal law and those that are appropriate for private law. It is more accurate to say that in criminal law the accused must have an honest belief in what is an objectively reasonable (public) manifestation of consent. Without that he is at least guilty of a reckless indifference to what right conduct requires as a matter of law. For discussion, see Bruce Chapman "Responsibility and Fault as Legal Concepts" *King's College Law Journal* 12 (2001): 212. Also, see Ripstein, *supra* n. 6, at 202-14.

[10] It has become (itself) conventional to credit David Lewis, *Convention* (Cambridge: Harvard University Press, 1969) with the first formulation of the concept of common knowledge, and Robert J. Aumann, "Agreeing to Disagree" *Annals of Statistics* 4 (1976): 1236-9, with the first rigorous formulation, demonstrating its fundamental importance for game theory. However, for a recent argument that suggests there are important differences between Lewis's account and that which has been imported into game theory by way of Aumann, see Robin B. Cubitt and Robert Sugden, "Common Knowledge, Salience, and Convention: A Reconstruction of David Lewis' Game Theory," *Economics and Philosophy* 19 (2003): 175-210.

[11]Formally, the problem here is that there are two (pure) Nash equilibria in this game. "A Nash equilibrium is an array of strategies, one for each player, such that no player has an incentive (in terms of improving his own payoff) to deviate from his part of the strategy array"; see David M. Kreps, *Game Theory and Economic Modelling* (New York: Oxford University Press, 1990), 28. In a single play game, the equilibrium notion shows up in the idea that only Nash equilibria can survive, under common knowledge of rationality, the recursive thought processes of individuals thinking through the rational thoughts of another. Thus, in Figure 1, there are the two possible Nash equilibrium outcomes (hence the indeterminacy for the friends' choice of actions), and *only* two possible Nash equilibrium outcomes. The 'only" will be important in Figure 3.

[12] See Christopher Woodard, "Group-based Reasons for Action" *Ethical Theory and Moral Practice* 6 (2003): 215, 217.

[13] *Ibid*., at 219.

[14] The holistic reasoning being contemplated here is akin to the sort of reasoning that Michael Bacharach has characterized as " we" thinking, something that he contrasts with the more usual "I/he" thinking that we see in game theory. See Michael Bacharach, " 'We' Equilibria: A Variable Frame Theory of Cooperation" (unpublished, June 24, 1997), 5. Robert Sugden's "team reasoning" is also closely related; see Robert Sugden, "Thinking as a Team: Towards an Explanation of Nonselfish Behavior," *Social Philosophy and Public Policy* 10 (1993): 69, and Robert Sugden, "Team Preferences" *Economics and Philosophy* 16 (2000): 175.

[15] I develop this argument more generally (that is, in a way that looks for its implications in the theories of individual choice and social choice as well as the theory of games) under the idea of "categorical reason" in Bruce Chapman, "Rational Choice and Categorical Reason" *University of Pennsylvania Law Review* 151 (2003): 1169-1210. All the discussion of the "causal" as distinct from the "conceptual" in the text at this point should not suggest that what is in play here is the same distinction that separates "causal decision theory" from "evidential decision theory". On the latter distinction, see Robert Nozick, *The Nature of Rationality* (Princeton: Princeton university Press, 1993), 41-63. The evidential decision theorist takes how another might behave (e.g., a Newcomb predictor, another similarly situated prisoner in the prisoner's dilemma, etc.) as evidence (sometimes) of how you should behave even though there is no causal connection between the other's behavior and the payoffs you face. This sort of thinking (while allegedly more "informed" by what others are doing) is still highly individualized or private, and differs, therefore, from what is being envisaged when your conduct as an agent is informed by a shared conception or understanding of what it is that you are doing as one part of a larger whole.

[16] I have tried to suggest elsewhere that the difficulties that some agents might face in articulating for others a shared conception of what they should do as a group might be useful if certain difficulties in social choice theory are to be avoided; see Chapman, *supra* n. 15, at 1195-1203, and Bruce Chapman, "More Easily Done Than Said: Rules, Reasons, and Rational Social Choice" *Oxford Journal of Legal Studies* 18 (1998): 293-329. The problem, as here, is that the agent cannot organize the particular action proposed under categories of thought that the other agents might share. "Eating alone", for example, is not unambigously (or categorically) better, or worse, than "eating together". In these earlier papers I relate this problem to some of the domain restrictions in social choice theory (in particular, Sen's "not between" value restriction) that are sometimes required if stable social choices (e.g., where there are no majority voting paradoxes) are to be ensured.

[17] Of course, no actual "call to action" is being contemplated here, as there is (by assumption) no actual conversation between the friends. Rather, what is being contemplated by each friend is only a thought experiment, albeit one that is a little more "public" than what goes on in strategic reasoning. The friend asks herself (silently) whether her conception of what they are (reasonably) doing (or what they ought to do) is something that the other friend can (reasonably, sensibly) be expected to share. Only then can the conception order what each will do together. Row should see that her call to action is not one that is likely even to occur to (or be conceived by) Column.

[18] This also shows that what is being proposed in this paper to solve coordination games where one of the Nash equilibria is Pareto superior to another is to be distinguished from the mere addition of something like a joint dominance principle. For examples of the latter approach, see David Gauthier, "Coordination" *Dialogue* 14 (1975): 195, and (for a cautious endorsement of this approach in limited circumstances) Kreps, *supra* n. 11, at 31. A shared conception of oneself as a total utility maximizer will avoid Pareto-inferior

outcomes without necessarily endorsing the relevant Pareto-superior outcome. In other words, under such a shared conception, some (but not all) agents might be made worse off.

[19] Two points need to be emphasized. First, when the law addresses coordination in the way suggested here, it does not address a coordination *game*, that is, the problem of how best to coordinate *actions* by individuals (e.g., say, by making some actions more salient than others). On this it contrasts with the interesting argument in Richard H. McAdams, "A Focal Point Theory of Expressive Law" *Virginia Law Review* 86 (2000): 1649-1729. Rather the law should address more general understandings or conceptions of action and interaction, developing and/or reinforcing these general understandings so that coordination games are solved as a matter of individual reasoning, through a particular shared conception, to the choice of an action. Second, when the law does this, it will help individuals to avoid more than just the problems modeled by coordination games, even coordination games where there is some conflict of interest (as in the game of chicken or the battle of the sexes game, both discussed by McAdams, *supra*). If the examples in this section are any indication, a shared conception of their interaction may well help the interacting agents to solve problems modeled by the prisoner's dilemma as well. That is, such an approach might encourage the agents to choose not only *between* Nash equilibria, but also to choose *non*-Nash equilibria.

[20] Eldar Shafir and Amos Tversky, "Thinking Through Uncertainty: Nonconsequential Reasoning and Choice" *Cognitive Psychology* 24 (1992): 442, 452-59.