Meta-games:

An analysis, interpretation, and application to the Theory of Mind through a consideration of the Prisoner's Dilemma

Abstract:

Howard Nigel's theory of Metagames was developed to provide new insights into the notions of "rationality" and its various paradoxes. When the Metagame theory is applied to the Prisoner's Dilemma, it yields a theoretical result traditional game theory does not anticipate, but intuition and empirical evidence suggest. This paper will discuss the Prisoner's Dilemma, its meta-game, its meta-meta game, and its 3-person meta-meta game. The discussion will present the Metagame framework as a possible model for a person's "Theory of Mind" in matrix game situations, and then analyze both the strengths and the drawbacks of this model.

The phrase "Theory of Mind" (ToM) is used in different ways to refer to distinct areas of investigation. There are the general theories that describe how people think (functionalist theories such as the *Computational Theory of Mind*, and brain-mind identity theories are of this kind). This paper, however, will use the term *ToM*[1] to refer to a specific cognitive capacity: the ability to consider and model another person's beliefs, desires and thoughts[2]. The phrase *ToM* in this context refers to a person's own "theory" for another person's mind. For example, if I were playing a game of Chess, I will, whether consciously or not, develop a *Theory of Mind* for my opponent to model his future moves. Beyond my understanding of the explicit rules of the game, my *ToM* for my particular opponent will also govern how I act. The sum of my observations helps create a *ToM* that I will draw on throughout a game. Individual *ToMs* thus are not objective constructs concerning thinking in general (as the general theories do); they are individual-specific, and may be related to our capacities for empathy and *mental simulation*[3]. This paper will investigate how a *ToM* is relevant in a specific kind of conflict: the Prisoner's Dilemma. And discuss how a mathematical framework, previously developed to investigate the Prisoner's Dilemma, is fundamentally related to current investigations in *ToM*.

The Prisoner's Dilemma (PD) is both one of the most compelling and relevant discoveries of Game Theory[4]. The PD has been used to model conflicts ranging from nuclear arms races to oligopolistic competition, and a game-theoretic analysis into the PD has even been used to justify preemptive war. The PD that is relevant to our current

[1] Though *ToM* will be used, this paper is interested, precisely, in a *Theory for Theories of Mind.*
[2] Though there are arguments for the existence of the ToM capacity in primates.
[3] Gordon, 1986

investigation in *ToM*, however, is on a smaller, more individualistic scale[5]. Past experiments that have put two people in a PD (represented by a matrix game) has produced results that differs greatly from what traditional game theory would deem "rational". A significant amount of work into the PD has involved creating new definitions and measures of ratonality, and introducing different interpretations of the conflict. These new constructs have been developed to reconcile the central "dilemma": despite the fact that mutual defection is the most "rational" outcome by traditional measures[6], mutual cooperation is preferred from both individuals, and in fact, occurs quite frequently empirically. When Nigel Howard introduced the theory behind *Metagames*[7], however, he was able to retain the standard measures of rationality while providing a legitimate theoretical explanation for the intuitive conclusion that mutual cooperation is in fact "rational" in its own sense.

Howard's metagame approach is general in theory, but we will only consider it here as it is relevant to the PD. Instead of a standard PD, the metagame of the PD involves a Player 1 (P1) having the standard choice between Cooperating (C) or Defecting (D). And a Player 2 (P2), choosing a "meta-strategy", which are strategies that are contingent on P1's choice. Since P1 pickes one of two strategies (C or D) and P2 can respond to each of P1's strategy one of two ways (C or D), P2 has a total of 4 meta-strategies as follows:

I. If player 1 cooperates, then defect; if player 1 defects, then defect
II. If player 1 cooperates, then defect; if player 1 defects, then cooperate
III. If player 1 cooperates, then cooperate; if player 1 defects, then defect
IV. If player 1 cooperates, then cooperate; if player 1 defects, then cooperate

[5] Though it is conceptually possible for an individual to have a *ToM* about another nation or marketplace competitor, we will only consider the case when a ToM regards another person.
[6] Mutual Defection is both an intersection of dominant strategies for both players, and a Nash Equilibrium
[7] Howard, 1971

Player 2's four strategies may be nicknamed I. defect regardless, II. do the opposite, III. do the same, and IV. cooperate regardless. The metagame PD thus yields the following set of outcomes:

		Player 2			
		I	II	III	IV
Player 1	C	CD	CD	CC	CC
	D	DD	DC	DD	DC

Which in turn yields the following payoffs[8]:

		Player 2			
		I	II	III	IV
Player 1	C	(1,4)	(1,4)	(3,3)	(3,3)
	D	(2,2)	(4,1)	(2,2)	(4,1)

An analysis of this metagame shows that the only Nash Equilibrium is when P1 chooses D (bottom row), and P2 chooses I (1st column); this results in the mutual defection outcome with the corresponding payoff of (2,2). So this metagame actually does not produce a new equilibrium, and the equilibrium between P1's Defect, and P2's "defect regardless" is maintained.

The metagame approach starts producing unique results once we consider the "meta-meta game". In the meta-meta game, P1 chooses from one of the four meta-strategies described earlier; P2, however, chooses from one of 16 "meta-meta" strategies. A meta-meta

[8] The numbers used are ordinal, they are used only to rank preferences

strategy is not contingent on P1's decision of C or D, but on one of the four (I, II, III, IV) metastrategies that P1 chooses from. P2's meta-meta strategies thus are as follows[9]:

1. I → D; II → D; III → D; IV → D
2. I → D; II → D; III → D; IV → C
3. I → D; II → D; III → C; IV → D
4. I → D; II → C; III → D; IV → D
5. I → C; II → D; III → D; IV → D
6. I → D; II → D; III → C; IV → C
7. I → D; II → C; III → C; IV → D
8. I → D; II → C; III → D; IV → C
9. I → C; II → D; III → D; IV → C
10. I → C; II → D; III → C; IV → D
11. I → C; II → C; III → D; IV → D
12. I → D; II → C; III → C; IV → C
13. I → C; II → D; III → C; IV → C
14. I → C; II → C; III → D; IV → C
15. I → C; II → C; III → C; IV → D
16. I → C; II → C; III → C; IV → C

The meta-meta game thus yields the following set of outcomes:

P1 (meta) \ P2 (meta-meta)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
I	DD	DD	DD	DD	DC	DD	DD	DD	DC	DC	DC	DD	DC	DC	DC	DC
II	CD	CD	CD	DC	CD	CD	DC	DC	CD	CD	DC	DC	CD	DC	DC	DC
III	DD	DD	CC	DD	DD	CC	CC	DD	DD	CC	DD	CC	CC	DD	CC	CC
IV	CD	CC	CD	CD	CD	CC	CD	CC	CC	CD	CD	CC	CC	CC	CD	CC

This set of outcomes, in turn yields the following payoffs[10]:

P1 (meta) \ P2 (meta-meta)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
I	22	22	22	22	41	22	22	22	41	41	41	22	41	41	41	41
II	14	14	14	41	14	14	41	41	14	14	41	41	14	41	41	41
III	22	22	33	22	22	33	33	22	22	33	22	33	33	22	33	33
IV	14	33	14	14	14	33	14	33	33	14	14	33	33	33	14	33

[9] Each strategy is a set of four "If-then" statements separated by the semi-colons. For example, I → D means "If P1 chooses 'I', then defect". Since P1 has 4 choices (I, II, III,IV), each of P2's strategy must have 4 separate "If-then" statements to account for the 4 different possibilities he must 'respond' to.

[10] Instead of (j,k) denoting the payoffs to the 1st and 2nd player, respectively, we are simply using *jk*

An analysis of this meta-meta game reveals three Nash Equilibriums. When P1 selects "I" and P2 selects "1", the familiar mutual defection occurs (as it did in the meta-game and in the standard PD). When P1 selects "III" and P2 selects "3", however, there is a Nash Equilibrium of mutual cooperation; this new equilibrium also occurs in the intersection of P1's "III" and P2's "6" strategies, respectively.

So what do these new equilibriums mean? And how is Howard's theory of metagames relevant to a Theory of Mind? Firstly, the fact that there is a new equilibrium at mutual cooperation is significant for several reasons. There has always been a lack of theoretically-convincing arguments that advocate cooperation as a "rational" strategy. Howard's theory of metagames not only makes mutual cooperation a *stable outcome*, it also accounts for elements of *psychological game theory* that traditional game theory is unable to consider. The fact that the two players involved in this conflict (the PD) are human beings, means that their theories of mind about each other will certainly play a factor in their decision-making. Although the objection that the meta-metagame PD is fundamentally different from the standard PD is valid, this fact does not imply that the findings of the meta-meta game are not relevant. Most important in introducing new frameworks to interpret the prisoner's dilemma is that the original conflict is preserved. Manipulations of the matrix that represents it or the ordinal preferences(payoffs) threaten the integrity of the conflict; Howard's metagames, however, does neither. Two individuals who have ToM's concerning each other could, presumably, interpret the game in such a way that the meta-metagame framework becomes *an appropriate model* for their respective *ToMs* for one another. If for example, P1 was employing 1st-order reasoning, he would see four possible strategies to choose from (strategies I through IV). Since P1 is aware that P2 is rational and wants to maximize his own payoffs as well P1 thus might reason "since we both want to maximize

our payoffs, I should choose a strategy that would retain the possibility that of the outcome that maximizes our combined payoffs: mutual cooperation"; thus P1 would eliminate strategies I and II. P1 could then reason "I don't, however, want to leave myself open to the possiblity of being exploited. And since III can always ensure a payoff that is at least as good as what IV can ensure, I will select III[11]".

Since P2 believes that P1 is employing 1st-order reasoning; P2 could thus conclude that P1 is deciding between the strategies I, II, III, or IV. The strategies 1-16 are thus appropriate representations of P2's choices. P2's thought process might be as follows "If P1 chooses I (defect regardless), I certainly will not employ a strategy where I will cooperate and be exploited", therefore P2 would eliminates strategies (5, 9, 10, 11, 13, 14, 15, 16). P2 could continue and reason "If P1 chooses II (do opposite), I again won't select a strategy where I would end up cooperating, since I would then receive the worst payoff", and so P2 then eliminates, of the strategies remaining (4, 7, 8, 12). P2 then would reason "If P1 chooses III (do same), then I would want to cooperate, since I am really choosing between either CC or DD, and mutual cooperation is better than mutual defection" and so he eliminates (1, 2). Note that P2's two remaining strategies (3 and 6) are the two (and only two) strategies that contain the mutual cooperation outcome as a Nash Equilibrium. Either one of those two choices would thus yield the mutual cooperation outcome (since P1 will, rationally, play III), but it is worth noting that P2 could further reason "If P1 chooses IV (always C), then I would prefer a strategy that defects so as to exploit him and receive the best payoff"; and thus eliminate strategy 6[12].

[11] Using our standard measure of rationality, III is a dominant strategy over IV

[12] Note that, although a Nash Equilibrium is contained in 6, P2 can still prefer 3 to 6 (and in fact does) because the outcome that make 3 preferable to 6 is away from the outcome of the Nash Equilibrium. Since P1 would not rationally choose IV, however, this last line of reasoning is irrelevant.

The lines of reasoning I have presented are certainly not the only lines that can be deemed "rational". As Nigel notes, the concept of rationality "breaks down" under different conditions, and so the same definition of rationality can, employed with different logic, yield different results (in fact, this is the very nature of the paradox in the PD). The reason I have employed that specific logic, however, is to demonstrate how the Nash Equilibriums found in the meta-meta game are not only theoretically valid, but intuitively valid as well. The logical progressions of reasoning by a 1st-order and 2nd-order player not only mirrors the elimination of "unintuitive" strategies in the meta-meta game, it also concludes that mutual cooperation is a rational outcome both intuitively, and when the meta-meta game is considered, theoretically.

Using the fundamentals of the meta and meta-meta game for the 2-person PD, we can now extrapolate, and consider a 3-person meta game. A 3-person PD is defined as having the following ordinal payoffs:

1. DCC
2. CCC
3. DCD = DDC
4. DDD
5. CCD = CDC
6. CDD

Where the payoffs are ordered according to the player whose decision is the 1st of the 3 letters. For example, in outcome 1, the 1st player receives the best payoff, since both the other players cooperated and he defected. The 2nd player (corresponding to the 2nd letter), meanwhile, receives the 5th payoff (CDC) because from his perspective, he cooperated while one other player cooperated and one other player defected. The 3rd player's payoff is similarly defined.

For a 3-person meta game, let the non-meta player (0th-order) be P1, with m_1 strategies (in general, $m_1 = 2$, because that is the standard PD). The meta player (P2) then has $2^{(2^{(p_1p_3)})}$ strategies (m_2), with each strategy having $2^{(p_1p_3)}$ elements (where $p_1=p_3=2$,

the number of "distinct resolutions", and an "element" is a single "If-then" statement). A "distinct resolution" is defined as an output (for all players, either C or D). Notice that m_2 depends on p_1 and p_3; m_2 cannot possibly depend on m_1 and m_3 (the number of strategies P3 has) because m_3 depends on the number of strategies P2 has (m_2)[13]. The distinction between the input p_i (a strategy) and the output m_i (a distinct resolution) is important not only on theoretical grounds, but on its implications for *ToM*.

In this 3-person meta-meta PD, the players 1 and 2 have the following strategies:

P1	P2		"nickname"
C	I.	1C3C → C, 1C3D → C; 1D3C → C, 1D3D → C	C regardless
D	II.	1C3C → C, 1C3D → C; 1D3C → C, 1D3D → D	C unless both D
	III.	1C3C → C, 1C3D → C; 1D3C → D, 1D3D → C	
	IV.	1C3C → C, 1C3D → D; 1D3C → C, 1D3D → C	
	V.	1C3C → D, 1C3D → C; 1D3C → C, 1D3D → C	C unless both C
	VI.	1C3C → C, 1C3D → C; 1D3C → D, 1D3D → D	Same as 1
	VII.	1C3C → C, 1C3D → D; 1D3C → C, 1D3D → D	Same as 3
	VIII.	1C3C → D, 1C3D → C; 1D3C → C, 1D3D → D	D if 1&3 are same, C otherwise
	IX.	1C3C → D, 1C3D → C; 1D3C → D, 1D3D → C	Do opposite of 3
	X.	1C3C → D, 1C3D → D; 1D3C → C, 1D3D → C	Do opposite of 1
	XI.	1C3C → C, 1C3D → D; 1D3C → D, 1D3D → C	C if 1&3 same, D otherwise
	XII.	1C3C → C, 1C3D → D; 1D3C → D, 1D3D → D	D, unless 1& 3 C
	XIII.	1C3C → D, 1C3D → C; 1D3C → D, 1D3D → D	
	XIV.	1C3C → D, 1C3D → D; 1D3C → C, 1D3D → D	
	XV.	1C3C → D, 1C3D → D; 1D3C → D, 1D3D → C	D, unless 1& 3 D
	XVI.	1C3C → D, 1C3D → D; 1D3C → D, 1D3D → D	D regardless

Where "1C3C → C" is a single "element" that reads "If Player 1 cooperates and Player 3 cooperates, then cooperate" and each element is separated by a comma. Roman Numerals "I - XVI" denote P2's 16 strategies, and we will simply use "C" and "D" to denote both the strategy and distinct resolution of P1. A single strategy from P3, since it is contingent on both the strategies of P1 and P2, is as follows

A. 1C2I → C, 1C2II → C, 1C2III → C, 1C2IV → C, 1C2V → C, 1C2VI → C, 1C2VII → C, 1C2VIII → C, 1C2IX → C, 1C2X → C, 1C2XI → C, 1C2XII → C, 1C2XIII → C, 1C2XIV → C, 1C2XV → C, 1C2XVI → C; 1D2I → C, 1D2II → C, 1D2III → C, 1D2IV → C, 1D2V → C, 1D2VI → C, 1D2VII → C, 1D2VIII → C, 1D2IX → C, 1D2X → C,

[13] m_3, the number of strategies P3 has, depends on m_2 because this is a meta-meta game, with P3 being the meta-meta player

1D2XI → C, 1D2XII → C, 1D2XIII → C, 1D2XIV → C, 1D2XV → C, 1D2XVI → C (Always C)

Since each strategy from P3 has 32 elements, P3 has 2^(32) strategies, far more than that which can be listed here. We do not, however, need to consider all the strategies from each player in order to analyze this game.

Let us consider first how we might arrive at a Nash Equilibrium to this game without exhaustively mapping it out. If one were to take the point of view of each player, one can then "build" an optimal strategy by considering each possibility (each "If", what the other two players might do) and then deciding whether a D or C would yield a higher payoff. In doing so, it quickly becomes clear that the optimal strategy is one that corresponds to the nickname "cooperate if, in doing so, the other players will also both cooperate. Defect otherwise". This nickname makes evident why Nash Equilibriums of mutual cooperation[14] exists in n-person (n-1)-meta games. By making one's strategy choice not independent of the relevant element in another person's strategy choice, meta-games possess a facet of "causality". P3, being rational, will thus select a strategy such that if P1 cooperates and P2 cooperates, he too will cooperate (if defecting meant another player would defect as well). By choosing a strategy that defects in every other situation, P3 is making sure that the other players won't have an incentive to defect; there is thus a Nash Equilibrium at the intersection of the dominant strategies that we "build"[15] (no player would want to move away from the equilibrium because doing so would directly "cause" another player to move away from it as well, thus producing a mutually less-desirable payoff). I therefore argue that a Nash Equilibrium exists when P1 cooperates, P2 chooses XII (C only if both P1 & P3 C, D otherwise), and P3 chooses "K", defined below:

[14] Mutual cooperation among all n players
[15] This is not to say there are not other Nash Equilibriums as well

K. 1C2I → C, 1C2II → C, 1C2III → C, 1C2IV → D, 1C2V → D, 1C2VI → D, 1C2VII → C, 1C2VIII → D, 1C2IX → D, 1C2X → D, 1C2XI → C, 1C2XII → C, 1C2XIII → D, 1C2XIV → D, 1C2XV → D, 1C2XVI → D; 1D2I → D, 1D2II → D, 1D2III → D, 1D2IV → D, 1D2V → D, 1D2VI → D, 1D2VII → D, 1D2VIII → D, 1D2IX → D, 1D2X → D, 1D2XI → D, 1D2XII → D, 1D2XIII → D, 1D2XIV → D, 1D2XV → D, 1D2XVI → D

Likewise, other (and all beyond DDD) Nash Equilibriums exist where all three players can "coordinate" a C, and where if any one person moves unilaterally away from that equilibrium to a strategy that yields a "D", the other meta-player has a strategy that would produce a "D" as well.

When trying to use our meta-game analysis to model *ToMs*, we do encounter a few problems. Firstly, P1 has no basis whatsoever to make his decision. By definition, a myopic player is one that is unable to consider how another player's thoughts and beliefs plays a role in his own payoff. A myopic player can therefore only consider his own desires; in order for P1 to make a choice, he must have some kind of payoff immediately available (but since the payoff is contingent on what P2 and P3 decide, which in turn is contingent on what P1 decides, P1 has no basis for making this decision without developing some kind of *ToM* about the other two players). Another limitation of this model is that it requires that the three players (1, 2, 3) to be 0^{th}-order, 1^{st}-order, and 2^{nd}-order thinkers, respectively. This is a very specific situation, and so this model is not applicable to all n-person Prisoner's Dilemmas (or even all 3-person PDs). A final, important point to consider is that, in applying this model, we have made the assumption that a particuar player's *ToM* is indicative of the actual order of reasoning another player is employing. The number of strategies the meta-meta player has in this game is not derived objectively from a measure of the number of strategies the normal and meta player are actually considering. Even if a 2^{nd}-order player is actually employing a *ToM* modeled by a meta-meta player, then, for our theoretical results to be valid, his subjective

belief of the strategies the other players are considering must be the objective reality. Our analysis, interpretation, and application of Meta-games to *ToM* will be therefore be invalid if the *ToMs* are not themselves indicative of what other players are actually considering.

Future studies into *ToM* using Metagame Theory should expand beyond the Prisoner's Dilemma. Since the fundamentals of a metagame are, intuitively, very similar to the recursions of logic employed in higher order reasoning, the applicability of metagame theoretical results to *ToMs* during other matrix games should be investigated. One of the primary drawbacks of Metagame theory is that it is unable to account for the situation where two or more players are applying the same meta-level. If two players in a prisoner's dilemma are both 2nd-order thinkers, metagame theory currently has not prediction for what kind of reasoning would take place. Until these limitations in metagame theory are addressed, it is unlikely that the theory of Metagames will be used to model more situations where *ToMs* are relevant.