Butler, D.J. (2012) A choice for 'me' or for 'us'? Using we-reasoning to predict cooperation and coordination in games. Theory and Decision, 73 (1). pp. 53-76.

A Choice for 'Me' or for 'Us'? Using We-Reasoning to Predict Cooperation and Coordination in Games*

By David J. Butler[#]
Economics Program,
Murdoch Business School,
Murdoch University, Murdoch, WA 6150, Australia

Key Words: Prisoner's Dilemma, Hi-Lo, We-Reasoning, Experiment

JEL classification numbers: C70, C91, D03, D81

Abstract
Cooperation is the foundation of human social life, but it sometimes requires individuals to choose against their individual self-interest. How then is cooperation sustained? How do we decide when instead to follow our own goals? I develop a model that builds on Bacharach's (2006) 'circumspect we-reasoning' to address these questions. The model produces a threshold cost/benefit ratio to describe when we-reasoning players should choose cooperatively. After assumptions regarding player types and beliefs, we predict how the extent of cooperation varies across games. Results from two experiments offer strong support to the models and predictions herein.

[#] d.butler@murdoch.edu.au Economics Program, Murdoch Business School, 90 South Street, Murdoch, Perth, Australia.

*Financial support from a UWA-administered Australian Research Council small grant is appreciated. Thanks also to participants at FUR Newcastle, 2010, seminar participants at George Mason University and the University of Arizona, to Martin Dufwenberg, Dan Houser, Paul Miller, two referees and the Editor for helpful comments and to Dahai Fu for statistical assistance.

1. Introduction

It used to be said that the laws of aerodynamics can prove the bumblebee is too heavy to fly. Magnan (1934) claimed to have proved it was “impossible”; but the bumblebee, being unaware of these laws, continues to fly. A similar situation appears to exist in some strategic settings for successful human cooperation and coordination and the predictions derived from game theory. The most well-known example of unpredicted cooperation involves the Prisoner’s Dilemma game, and (primarily for coordination) a game labelled ‘Dodo’ by Binmore (1992) but more often known as ‘Hi-Lo’ (e.g., Bacharach, 2006). Here we shall use ‘**A**’ for Hi and ‘**B**’ for Lo.

Figure 1: The Hi-Lo Game

		Player 2	
		A	**B**
Player 1	**A**	2,2	0,0
	B	0,0	1,1

The reason such a seemingly simple game has been widely debated is that while human players find it almost self-evident to choose ‘**A**’, game theory’s basic assumption of individual instrumental rationality (IIR) alone can not provide a coherent reason for a player not to rule out ‘**B**’ (see Bacharach 2006, Gold and Sugden 2007). This is because [**B**, **B**] is as much a Nash Equilibrium of the game as [**A**, **A**]. Gold and Sugden (2007) conclude:

> If we find that standard game-theoretic reasoning cannot tell players how to solve the apparently trivial problem of coordination and cooperation posed by Hi-Lo, we may begin to suspect that something is fundamentally wrong with the whole analysis of coordination and cooperation provided by the standard theory.

But as standard theory works well in many other applications, we need to understand why IIR is not the sole appropriate foundation on which to apply game theoretic methods. Identifying the set of games for which an alternative to IIR is needed is our first task. The games that interested Bacharach in his (posthumous) *magnum opus* (2006) included not just the one-shot, simultaneous-move, symmetric, two-player game of *Hi-Lo*, but also *Chicken*, *Prisoner's Dilemma*, *Stag Hunt* and an alternate form of Stag Hunt sometimes called *Tender Trap*. It is this set of games that will be the focus of the current paper[1].

With reference to Figure 2, the following payoff inequalities demarcate the scope of this set of games: **R**>**P** for mutual cooperation to exceed mutual defection; **R**>**S** to ensure cooperating alone is risky; **T**>**S** for the off-diagonal cells to penalise the cooperator.

Figure 2: The 2x2 PD Game

		Player 2	
		A	**B**
Player 1	**A**	R,R	S,T
	B	T,S	P,P

Of the 24 possible rank orderings only five satisfy these inequalities. One of these, **R**>**T**>**S**>**P**, is trivial because cooperation is a dominant strategy under any known mode of reasoning. This leaves four payoff rankings of 2x2 symmetric games: *Prisoner's Dilemma* (**T**>**R**>**P**>**S**)[2], *Chicken*

[1] Bacharach focused on 2-player games but took his basic arguments to generalise also to n-player versions. While he believes group identification is independent of group size (2006, p.112), the likelihood *all* players cooperate probably falls as n rises. Consistent with this, Van Huyk, Battalio and Beil (1990) found cooperation in a stag hunt game with 2 players was higher than for that game played with 16 players. We have not attempted to model or test this issue here, but their finding suggests it is a suitable issue for future work.

[2] Note the repeated-game condition 2R > (T + S) is omitted because these are one-shot games. Our interests also extend beyond the PD game. Although the (asymmetric) 'group' payoff is not maximized through mutual

(**T**>**R**>**S**>**P**), *Stag Hunt* (**R**>**T**>**P**>**S**) and *Tender Trap* (**R**>**P**>**T**>**S**). Other games intermediate between these types, for example **T**>**R**>**P**=**S**, are also of interest. A limiting case of *Tender Trap* as **T** is lowered to zero is the *Hi-Lo* game for which **R**>**P**>**T**=**S**=0.[3]

The Prisoner's Dilemma (PD) game is a notoriously barren landscape for cooperation; for this reason it is also an important benchmark against which the rationality of cooperation has been vigorously debated. Social psychologists use the payoff difference [**T**-**R**] to measure the greed incentive, and [**P**-**S**] to measure the fear motive; both are present in the PD game (Simpson 2006). In 'Morals by Agreement' (1986) Gauthier develops a controversial model he calls 'constrained maximization' which promotes the view that cooperation in a one-shot PD game is the uniquely rational choice. His claim was met with staunch opposition from game theorists (e.g. Binmore, p.133 in Gauthier and Sugden, 1993) because he advocated the choice of a strongly dominated strategy. This criticism also holds if a theory merely allows, rather than prescribes, cooperation; despite this, the intuition that cooperation is not obviously irrational has not gone away.[4]

However, obviously-dominated options are not often chosen in other domains where the dominant nature of the choice/strategy is equally transparent to players. Numerous studies of choice over lottery pairs have shown that despite the many systematic violations of Expected Utility Theory axioms, the one principle that is very rarely violated is *transparent* dominance. For example, Butler and Loomes (2007) make respect for this principle a cornerstone of their

cooperation if 2R < (T+S), 'we'-thinking may still be triggered for PD as for Chicken games, where often 2R < (T + S). See also footnote 10.

[3] Hi-Lo is not tested here because the expectation of near universal cooperation is not in dispute.

[4] The PD game is not the only 2x2 game to encapsulate the individual/collective conflict. Both Colman (1995) and Thaler and Camerer (2003 p.164) argue the Chicken game is in many ways more suited to investigating cooperative versus competitive tensions than the PD game because the 'fear' motive for defection is eliminated.

model of imprecise preferences. Why then is applying the ‘transparent dominance’ argument to PD games controversial?

In relying solely on IIR, game theory permits only the question ‘what should *I* do’; not one stemming from an alternate perspective ‘what should *we* do’ (Sugden 2003, p.167). If I take the ‘me-thinking’ view, my preferences, as represented by the payoffs, can be summarised as they are written in the normal form matrix and my choice will follow those utilities in the standard way. If the payoffs in the matrix are dollars, other sources of utility such as altruism may alter these dollar payoffs, possibly to the extent that a payoff-transformation changes the game so that its utility ranking no longer constitutes a Prisoner’s Dilemma. A desire to cooperate in the new game could not, of course, be used to justify cooperation in a PD.

But suppose when facing a particular configuration of dollar payoffs I instead perceive the problem as one for *us* and choose as a component of a dyad, when I consider this ‘dyadic goal’ sufficiently likely to be shared? If the configuration of individual payoffs leads me to group identify, the *level of agency* of my actions is transformed and I may use ‘we’-reasoning for my decision, not payoff transformation.

Theorists usually assume *any* sources of motivation can be combined together to produce these utility numbers and that the source of motivation they embody is irrelevant once the numbers are identified. It may then be argued that a game intended to be a PD, for example, is actually a Stag Hunt game when we include any symbolic meaning[5] cooperation in that context has for us. But there are profound conceptual reasons why agency transformation can not simply be reduced to payoff transformation in this way. As Nozick (1993, p.55) explains: “…if the

[5] Any ‘meaning’ a particular choice may have must clearly depend on the characteristics of the option(s) we reject, as well as the choices the other player faces. ‘Meaning’ in this sense can’t attach to a choice in isolation.

reasons for doing an act **A** affect its utility, then attempting to build this utility of **A** into its *consequences* will thereby alter that act and change the reasons for doing it".

Hollis and Sugden (1993) make a related argument. In Savage's (1954) theory, a preference between acts **A** and **B** derives from the consequences of those choices under each event. Hollis and Sugden point out that to allow these consequences to be slotted into *any* event or choice as Savage's theory requires the consequences (utilities) must be independent of the description of any *particular* event or choice. There is then no scope for "…reflection to pass judgment on the promptings of preference".[6] They use these arguments to conclude that:

> Savage's axioms serve to rule out accounts of motivation where the source or character of satisfactions affects the agent's attitude towards consequences or where principles enter into the description of acts. To this extent then, the axioms presuppose a particular account of motivation.

For decisions in experimental settings, violations of transparent dominance in the Prisoner's Dilemma exceeding 50% can be observed through suitable choice of dollar payoffs, in striking contrast to the violation rate for transparent dominance of 1-2% seen in choices over lottery pairs (which also use dollar payoffs). The difference between these cases is so striking it suggests that maybe half of all players do not view their decision to cooperate in the context of a one-shot PD as the choice of a transparently dominated option. If our aim is to model human behaviour, including of the many people who do not accept that a '**B**' choice is uniquely rational, agency transformation offers us an alternative way to reason from payoffs to decisions.

[6] Gauthier may have had a similar idea (Gauthier and Sugden 1993, p.186) when he claimed that standard theory: "…leaves no conceptual space between preferences [*the utilities in the matrix*] and choices, and a view that leaves no such space is simply too impoverished, in the distinctions that it admits, to provide a model of rational action." Bacharach's approach would differ as he accepts standard theory does offer *a* model of rational action, just not the *only* model of rational action for these games.

Bacharach modifies we-reasoning to recognise that people won’t always choose cooperatively, even under a ‘we’-frame, because they lack assurance others will we-reason; this broader perspective he calls circumspect we-thinking. Gold and Sugden (2007) present a formal statement of this concept as a reasoning schema. In this paper I shall assume hereon that circumspect we-thinking is the missing concept game theory needs to incorporate to predict human choice behaviour in these games.

2. Adaptive Origins

Any adaptation favouring human cooperation would need to be both nuanced and finely balanced against selfish concerns to survive the pruning and shaping of natural selection. Bacharach argues that if we recognize the diversity of encounters that early humans faced, no one game can serve as an adequate model for all kinds and contexts of possible cooperation. But selection for many game-specific special-purpose behavioural mechanisms is implausible. Instead, the evolution of a single, facultative, mechanism for promoting cooperation in related game situations is more probable, offering economies of scope in application across the set of relevant games.

Bacharach (2006, p. 111) argues:

> “… Dispositions to cooperate in a range of types of game have evolved in man, group identification has evolved in man, and group identification is the key proximate mechanism for the former. … Group identity implies affective attitudes which are behaviourally equivalent to altruism in Dilemmas, and it can explain what altruism cannot, notably human success in common-interest encounters”.

His perspective fits neatly within the popular ‘social brain hypothesis’ (Dunbar and Shultz, 2007). This hypothesis notes that early man’s ecological challenges were mostly solved socially, requiring both coordination with and flexibility in interactions with other tribal-group members. The capacity for language may also have been selected for because of its facilitation of cooperation in groups. Tomasello (in Tomasello *et al* 2005 p.690) argues that underlying human cognition is “an adaptation for participating in collaborative activities involving shared intentionality”; this shared, or ‘we’, -intentionality is he argues unique to humans. Any shared intentionality might make we-thinking not simply possible, but in many contexts, natural.

Social psychologists have arrived at a concept similar to ‘we’-reasoning which they call ‘transformation of motives’ (De Cremer and Van Vugt, 1999). Those authors investigated the consequences of enhancing players’ group identification on their contributions in a public goods experiment. They found an increase in contributions from those previously not contributing, but no further increase from those who were already contributing. The inference is that the latter group already perceived the decision as one for the ‘group’ so the manipulation had no further effect. If some forms of ‘cheap talk’ can enhance group-identification which then raises contributions from free-riders, such a ‘medium of social exchange’ may be more valuable than usually assumed.

Finally, our capacity in the marriage dyad to subsume our individual goals within the goals of the couple may be a consequence of the evolution of shared intentionality. There is evidence that the frequency with which a spouse uses the words ‘we’ and ‘our’ (problem) in discussing a partner’s health after an episode of heart failure, rather than ‘he’ and ‘his’ (problem) predicts the future health of the partner (Rohrbaugh *et al*, 2008).

I ignore in this paper constant-sum games (e.g., the Dictator game) that primarily invoke altruism or inequity-aversion rather than we-thinking. Dispositions for altruism and equality likely draw upon subtly different emotions and instincts than group identification. Price *et al* (2002) argue that any mental module adapted for reciprocity would not also generalize to 'fairness' games. Limited experimental evidence supports this conjecture. Brosig (2002) for example found little correlation between the subjects who cooperated in PD games, and the generous subjects in Dictator Games, as has other work by Butler *et al* (2011).

If the feelings behind man's intuitively we-thinking approach are adaptations embodying information accumulated over thousands of generations of tribal life, our choices today will draw upon and build on the messages transmitted by those feelings. In summary, our response to the payoff-configurations defining these games may, by agency transformation, be nature's way of helping *us* extract *our* mutual benefits from cooperation and coordination.

3. The Models

3a) Using IIR: The Standard Model

Let '**A**' be the cooperative choice and '**B**' the defecting choice as labelled in our experiments. Here, **A** is only chosen by a 'me'-thinker for self-interested reasons. In standard game theory we can assign a subjective probability to the strategy choice of the other player. Let p denote a 'me'-player's estimate of the chance another player will choose option **A**, and (1-p) choose option **B**, where $p \in [0, 1]$ (see Figure 3).

- Figure 3 here -

These games possess between one (PD) and three (e.g., Hi-Lo) Nash equilibria in pure or mixed strategies. Where it exists, the value of p that equates the expected values of options **A** and **B** so that he is indifferent between them, we denote by p^* and is the mixed-strategy NE in conjectures for the game:

$$p^* = \frac{(P-S)}{(P-S)+(R-T)} \qquad (1)$$

If he believes either $p > p^*$ or $p < p^*$, then either **A** or **B** could offer a greater expected value. We can think of p as his estimate of the probability that another player will choose '**A**'. In these games a 'me'-thinker will use (1) to decide his choice. In the PD game he may still expect some others to cooperate; but no matter how strong his belief that others will play '**A**', it will always pay him to defect and (if correct) receive '**T**'. He will never cooperate in any one-shot PD game, as no value of $p \in [0, 1]$ in (1) can ever make EV (**A**) > EV (**B**).

3b) The Circumspect We-Reasoning Model

Now let us adapt (1) for circumspect we-reasoning. We-thinking players are likely heterogeneous in their degree of we-strength, weighing the perceived risks and dyadic benefits involved. Let us assume a random player is drawn from a population represented by a continuum of circumspect we-thinkers. Denote this new dimension using a continuous variable, c (where $c \in [0, 1]$), representing the degree to which a player is willing to act as part of the dyad. A committed we-player using $c = 1$ chooses as if either 'we' will get **R** or 'we' will get **P** (although recognising there is no causation to guarantee these outcomes). She will then choose option **A,** as **R** > **P** in the games of interest in this paper. Similarly, if a player uses $c = 0$ she is indistinguishable from a 'me'-player and the new model reduces to the standard model in (1).

But the most interesting case is the contingent we-thinking player; using $0 < c < 1$, her circumspect approach leads her to balance her we-goals with an eye to self-protection against both ‘me’-players and potential ‘**B**’ choices by other contingent we-reasoners. How might we incorporate these assumptions into (1) in the most parsimonious way? As in (1), let ‘*p*’ denote a player’s estimate of the chance another player will choose option ‘**A**’, but where the player is now a ‘we’-thinker who also recognises the existence of ‘me’-types in the population.

For each of player 2’s options, player 1’s ‘*c*’ should measure the degree of increased decision-weight player 1 assigns to the ‘we’ choice-pairs (at the expense of the off-diagonal payoff-combinations). When $0 < c < 1$, increasing weight is given to **R** and **P** as *c* rises from 0 to 1. Her maximizing choice could then be either **A** or **B** depending on: i) her degree of we-thinking, *c*; ii) her beliefs regarding the likelihood of others’ cooperation, *p*; and iii) the particular payoff-values in the game: **R**, **S**, **P** and **T**. We may then represent the expected value of choosing **A** and of choosing **B** for such a player as follows.

----- Figure 4 here -----

Strictly speaking, it is not essential to subscribe to the logic of we-reasoning to make use of Figure 4. Even if one were committed to IIR and individual agency, a model that can predict the extent of error across this class of games would still be descriptively and practically useful. As an analogy, consider the way probabilities are transformed in decision theories such as cumulative prospect theory (Tversky and Kahneman 1992), where for example 4% + 4% might be treated as if it equalled 10% in the weighting of consequences. While probability transformations are popular because they can improve the descriptive fit of some theory, it is arguable whether the resulting behaviour should be counted as normatively rational.

Returning to Figure 4, the threshold value for *c* denoted by *c**, is that for which EV (**A**) = EV (**B**). For Player 1, by solving for *c* we find:

$$\boldsymbol{c^*} = \frac{p(\mathrm{T}\text{-}\mathrm{R}) + (\mathit{1}\text{-}p)(\mathrm{P}\text{-}\mathrm{S})}{} \qquad (2)$$

$$p(\mathrm{T\text{-}P}) + (1\text{-}p)(\mathrm{R\text{-}S})$$

By symmetry (2) also holds for Player 2. When a ‘we’-thinking individual’s $c > c^*$, her identification with the dyadic-goal is sufficient for her to choose **A** in that game. As her EV (**A**) > EV (**B**), her maximizing choice is to participate in bringing about [**A**, **A**] over [**B**, **B**]. Similarly, if her $c < c^*$, instead of playing her part in [**A**, **A**], she chooses option **B**. *Ceteris paribus*, the higher c^* is in any game, the smaller is the fraction of we-thinkers whose we-strength will satisfy $c > c^*$, and so a lower percentage of cooperative choices is predicted.[7]

The model in (2) can also be understood as follows. The numerator shows Player 1’s expected ‘cost’ from selecting the cooperative row, while the denominator shows her expected ‘benefit’ if Player 2 selects the cooperative column. *Ceteris paribus*, as the numerator falls (denominator rises), the costs of cooperation decline relative to the benefits, lowering the c^* threshold and raising the predicted percentage of ‘**A**’ choices. Model (2) then produces a threshold level for cooperation and coordination analogous to Hamilton’s ‘cost to donor/benefit to recipient’ rule for kin-selection theory (Hamilton 1964); if a player’s c exceeds this threshold, cooperation is justified. The model therefore fits comfortably with the underlying logic of both evolutionary biology and economics and is consistent with the five rules for the evolution of cooperation identified by Nowak (2006).

Interestingly, model (2) can be applied in other contexts if ‘c’ is given a different interpretation. For example, suppose a me-reasoning Player 1 were to play these games against a Player 2 drawn from a population consisting of proportion (1-c) of other me-reasoning players,

[7] Threshold value c^* is unique under a positive linear transformation of these payoffs. Indeed, List (2006) presents evidence from a TV game show which finds a significant minority of participants choose cooperatively in a situation analogous to the PD even when thousands of dollars are at stake.

and proportion *c* of computer programs that simply mirror back the choices of the 'me'-reasoning Player 1. The optimal choice can be calculated using precisely this same result.[8] One key difference however is that the weights attached to the payoffs in Figure 4 would then be probabilities, as a degree of causation between the 'me'-player's own choice and that of the column player would exist, which is not true for circumspect 'we'-thinking.

When applying (2) for a game with specified values of [**R**, **S**, **P**, **T**], we can simplify the equation for threshold value, *c**. Figure 5 gives two examples and then plots *c** as a function of *p*, *c**(*p*), using a unit square diagram in '*p*, *c*-space'. Here *c**(*p*) separates the area above the EV (**A**) = EV (**B**) line, the region where option **A** is preferred by a 'we'-thinking player (as one's $c > c^*$), from below where option **B** is preferred (if one's $c < c^*$).

------ Figure 5 here ------

Note that the EV (**A**) = EV (**B**) line has no horizontal intercept for PD games, as option **A** can never be optimal when a player's $c = 0$. To summarize:

Claim 1: 'Me'-thinking requires *p* in [0, 1] and $c = 0$ and predicts individual maximizing choices according to $p < p^*$ and $p > p^*$. Model (1) is used.

Claim 2: Circumspect we-thinking requires *p*, *c* in [0, 1] and predicts the circumspect we-maximizing choice is '**A**' if the (*p*, *c*) pair lies above Figure 5's *c*(p)* line; she will choose '**B**' otherwise. Model (2) is used.

3c) Assumptions for Predicting Cooperation

If we hope to make predictions for the frequency of cooperation that could hold true across a population for this set of games, we require simple assumptions on the use of 'we'-thinking and on

[8] See Section 5a for still another possible use of Model (2).

how *p* is distributed. As there is no greed motive for defection in Stag Hunt and Tender Trap (including Hi-Lo) games, all players will be assumed to use circumspect we-thinking to coordinate. For the game-types with a ‘greed’ motive, a meta-analysis of PD games by Sally (1995) suggests around half of subjects are potentially willing to cooperate; in their works both Sugden and Bacharach take this figure to be approximately correct. There is also extensive evidence from social psychology (summarised in Van Vugt and Van Lang, 2006) that measures of social value orientation generally find some 60% of people are disposed to maximise mutual gains (*pro-socials*), some 30% seek to maximise individual gain (*individualists*) and 10% to maximise relative gains (*competitive types*); the latter two categories can’t be separated by observing play in our experiments. In the experiments reported here the highest frequency of cooperation in a PD game is 55% which is consistent with that literature (see also footnote 14). Drawing upon this body of evidence the relevant figure for a typical subject pool is likely between 50-60%. For simplicity I use a figure of 50% we-reasoners to make predictions.[9] The other 50% are assumed to use model (1) in PD and Chicken games.

While these traits are usually taken to be broadly stable characteristics of individuals, this does not imply we can measure such traits precisely, or that an individual’s actual choices will always express an underlying trait consistently. We know from numerous individual choice experiments that subjects may reverse their preferences over lotteries within the course of a single experiment and there is no reason to assume choices in our experiments are any less stochastic. But nor would it be supportive of our model if the aggregate cooperation patterns were driven by wildly inconsistent choices at the level of the individual, as judged against the

[9] If the figure were substantially different, we would need to modify the ‘overall predicted %A’ column in Tables 3 and 4 accordingly. A substantially lower figure would pull that column closer to the ‘me-frame predicted %A’ value and a significantly larger value would push us closer to the ‘we-frame predicted %A’ value.

model's *c** values. While within-subject variability of *c* can not be fully investigated here as each subject completed just one experiment, we can at least see whether our subjects chose consistently when facing the same game a second time in the same experiment. Although the initial purpose of using some games twice in an experiment was to check whether aggregate cooperation rates were sufficiently stable for prediction to be meaningful, the data can also be used at the level of the individual. We can then see whether a game with a given *c** value is treated consistently by an individual. The results are reported in Section 4.

That player's are heterogeneous is a necessary assumption: first, to allow for both me- and we-thinkers to exist and second, to generate a distribution of *c* for when we-thinking is used. We have no theory to tell us how the frequency of use of 'we'-reasoning might otherwise vary with game parameters. Without such a theory I make the simplest and most tractable assumptions based on the stylized facts.

Assumption 1: In Prisoner's Dilemma and Chicken games, half of players' use 'we' reasoning (model (2)) and half 'me' reasoning (model (1)). For Stag Hunt, Hi-Lo and other games, assume all players use 'we' reasoning (model (2)).

Assumption 1 can easily be adapted if factors known to affect the generation of emotions behind group identity, such as the contextual framing and 'social distance' individuals experience in an experiment so require. A double-blind design for example may fail to trigger human social cues sufficiently to engage our latent tendency to group identification. These factors may lower the prevalence of we-thinking and therefore the level of cooperation/coordination achieved, just as manipulations to enhance group identity as in De Cremer and Van Vugt (1999) may raise it. An implicitly uniform distribution of *p* is also the simplest benchmark assumption to use.

Assumption 2: When a player uses 'me'-reasoning, the proportion of the *p*-line where EV (**A**) > EV (**B**) gives the predicted level of cooperation.

Assumption 3: When a player uses 'we'-reasoning, the fraction of (*p*, *c*)-space where EV (**A**) > EV (**B**) is the predicted level of cooperation.

Assumption 3 effectively flattens each game's *c*(p)* line (see Figure 5) at its mean *c* value for that game. This generates a ranking of the inherent tensions (or lack of 'harmony', *cf*. Zizzo and Tan, 2008) of each game, as measured by the degree of we-identification required for cooperation: the larger is *c**, the fewer 'we'-thinkers will possess sufficient dyadic-identification to risk participation in achieving [A, A].

Model (2) combined with assumptions 1-3 also predicts a consistently high level of coordination in the Hi-Lo game. For **R**=5 and **P**=1, (**S**=**T**=0), (2) predicts 98.2% will choose '**A**'. Reducing **R** to 3 would, *ceteris paribus*, lower predicted '**A**' choices to 95.3%, and if **R**=2, to 89.8%. Although reassuringly high, these calculations likely *understate* the true level of cooperation, because only a low *p* can support a **B** choice; in reality fewer players than is implied by our uniform *p*-distribution assumption are likely to hold such beliefs.

4. The Experimental Design and Results

4a) Experimental Design

Experiment 1 involved eighty one subjects playing 25 2x2 one-shot games at the Economic Science Laboratory at the University of Arizona. The games used are listed in Table 1. The payoffs of these games were chosen to reflect a variety of incentives to cooperate or defect, as measured by model (2).

- Table 1 here-

Six sessions were run, each with 12-15 undergraduate subjects. No subject participated in more than one session. Players were seated at computer terminals in booths separated by screen walls. Each session began with an introduction to the experiment, including an opportunity to ask questions. Players worked through the introduction and sample games with the administrator, who projected his computer screen onto a large white screen visible to all. The administrator read the instructions out loud to the players, to ensure common knowledge of the experimental conditions.

Each player was randomly paired with another player in the lab using a ticket sealed in an envelope previously placed in each booth. The set of envelopes contained two copies of each ticket number and the single ticket in each envelope was the subject's ID; her ID number matched one other person's, he or she being the pair. Players knew they had an ID which paired them with some random person from the same session, and understood they would later briefly meet face-to-face to determine payment. The identity of the other player was concealed until after the completion of the experiment, as the envelopes were not opened until all decisions of both players were completed. In other words, there were no opportunities for communication or feedback, minimising potential super-game effects. As a choice is more likely to be perceived as one for 'us' when there is symmetry of circumstance, only symmetric payoffs were used in these experiments.[10] Players also had only intermediate social distance from each other, and shared a common experience of being students seeking to earn money from Professors. In these senses at least, they constitute a well-defined group sharing a common goal.

An incentive-compatible payment method analogous to the random lottery incentive system was used for the 25 decisions to cooperate or defect. When the players had been paired, one of them

[10] Evidence suggests that cooperation rates are lower in asymmetric PD games, and lower still as the asymmetry widens (Marwell and Schmitt, 1975). Anomalous cooperation may be less common in society under asymmetric conditions if fewer people perceive such choices as a problem for 'us' or if assurance declines.

drew a ticket from a box containing a number from 1-25. This selected the game to be played for real. Each player's choice in this game was then retrieved, and they were paid according to their choice combination. Payments averaged US$16.50 per player, with a range from US$0 to US$36 (US$2 for each unit of payoff).[11] The experiment took approximately 1 hour of participants' time.[12] To observe any inherent variability in aggregate cooperation, three games were presented twice: in Table 1 compare games 9 with 23, 12 with 25 and 15 with 21.

Experiment 2 was held subsequently at the University of Western Australia. While the main focus of that experiment was to investigate the *causes* of individual differences in play which is not the topic of the current paper, the aggregate results from that set of games are included here to add to the weight of experimental evidence bearing upon model (2). Experiment 2 focused on Chicken and PD games; there were 20 of each type of game, for a total of 40 games in all. One hundred and three subjects played each of these 40 games.[13]

----- Table 2 here -----

Six games were repeated in Experiment 2: in Table 2 compare games 15 with 33, 18 with 34, 24 with 32, 2 with 36, 11 with 37, and 21 with 39. Finally, across Experiments 1 and 2, Experiment 2's games 9 and 10 were identical to Experiment 1's games 5 and 4, so any difference in cooperation across subject pools can be observed. Combining Experiment 2's 40 games with Experiment 1's 25 games, overall we have data from 65 games to investigate our claims.

[11] A couple of times there were an odd number of players in a session due to no-shows. One random player then received two envelopes so that his/her responses could be played against two people. That player would however only receive payment for the first pairing.

[12] After making their choice for each game, each player then answered four supplementary questions regarding that game. That data is not used in this paper.

[13] The same procedures were used as for Experiment 1 except that the Lab had no screens between computer terminals. As the venue we used was never crowded, this had little effect on privacy.

4b) Experimental Results

Experiment 1

Table 3 reports the predictions and percentage of players choosing option **A** in each game.

Table 3 here

The first column in Table 3 shows the c^* equations derived for each game in Table 1. The second column calculates the fraction of $[p, c]$ space where the cooperative choice is optimal.[14] For Stag Hunt and Tender Trap games this fraction constitutes the model's prediction for the proportion of cooperative choices. For the PD and Chicken games we use p^* to calculate the fraction of the horizontal axis where EV (**A**) > EV (**B**) and this is shown in column 3. In column 4 we then sum half of column's 2 and 3 for our predicted value, **x**, to account for both 'we' and 'me'-thinkers in those games.

The variation in %**A** across these 25 games, from 8.8% to 91.2%, shows a very strong correlation with our predictions. The Pearson correlation coefficient is $\mathbf{r} = +0.966$, significant at any level (giving $\mathbf{R}^2 = 0.933$). Encouragingly, not only the direction of change but the *level* of cooperation fits these predictions very well. A simple OLS regression offers further insight; regressing 'actual **A**' on the 'predicted **A**', here **x**, gives:

$$\text{Actual } \%\mathbf{A} = \underset{(t = 1.47)}{3.12} + \underset{(t = 18.66)}{0.887}\mathbf{x} \qquad F = 348.35$$

As the predicted value (**x**) ranges from 0% to 93.5% (see Table 3, column 4), the regression implies actual cooperation varies from 3% to 86.1%, only slightly under-predicting the 8%-91% observed. Because the dependent variable is a proportion, a generalised linear model (GLM) is more

[14] Integration by parts calculates the area of the definite integral between the equal EV line and the horizontal axis. As the result gives the fraction where **B** is preferred, we subtract this figure from 1 to obtain the fraction where **A** is preferred. PDF files of the calculations are available in the web appendix.

precise. Estimating with the GLM gives a predicted range for cooperation of 9.4% to 85%. The partial effect at the sample mean is 1.062, which compares with the fixed slope coefficient in OLS of 0.887. On an aggregate basis then, the new model performs very well in representing the balance of incentives for a population to choose **A** or **B** within a game. The three games played twice reveal only fairly minor differences of between 1 and 8 percentage points of cooperation, suggesting aggregate predictability. At the level of the individual, choice switching rates were 9%, 12% and 28%. If we then see how many switches each person had in these three cases, we find 47 of the 81 players were fully consistent, 26 switched once, 6 switched twice and 2 switched at each of the three occasions. So for some 90% of subjects, play appears to be fairly consistent.

Two minor caveats to our assumptions (and/or model) are noted. First, in the severest PD games (as measured by c^*), actual cooperation is consistently several points higher than predicted. Unconditional co-operators may be driving this finding (i.e., the distribution of c may contain a small peak at $c = 1$ rather than being uniform for we-thinkers). However such players are not strictly *circumspect* 'we'-thinkers.[15]

Second, four games have a cooperation level 10+ points below our predictions and all four are SH or TT games. The most likely reason is that unlike Hi-Lo, somewhat fewer than 100% of players use we-thinking for every SH or TT game. While these caveats may well play a small but genuine part in the solution to the puzzle of human cooperation and coordination, our emphasis on developing a parsimonious and tractable predictive model leads us to not pursue them here.

[15] At the completion of experiment 1 we presented subjects with a list of several possible ways to best describe their play across the set of games. Of the 81 players, 41 selected the disposition which comes closest to circumspect we-thinking, consistent with the ideas in this paper. However a further 5 selected the disposition closest to unconditional cooperation, suggesting 57% are pro-social. Nearly all others chose either the self-interest or 'maxi-min' descriptions.

Experiment 2

For the 40 PD and Chicken games in Experiment 2, the Pearson correlation coefficient across the 40 games is **r** = +0.961 (giving $\mathbf{R}^2$ =0.923). A simple OLS regression of actual cooperation on predicted cooperation shows:

$$\text{Actual }\%\mathbf{A} = 0.076 + 1.036\mathbf{x} \qquad F = 525.75$$

$$(t = 0.03)\ \ (t = 22.93)$$

As **x** here varies from 6.4% to 93.4% (see Table 4, column 4), actual cooperation is predicted to vary from 6.5% to 96.8%. Cooperation across these games actually varies from 7% to 89%. Using the GLM, the comparable prediction is from 10.8% to 91.1%. The partial effect at the sample mean is 1.266, compared with 1.036 for the fixed slope in OLS.

The six games that were played twice showed little aggregate variability: differences were once again between 1 and 8 percentage points of cooperation; implying cooperation rates are sufficiently stable for prediction, as in Experiment 1. At the level of the individual, choice-switching rates varied between 15% and 32%. Across the 6 opportunities, 31 of 103 subjects are fully consistent, 39 switched once and 18 switched twice, suggesting for 87% of subjects play is fairly consistent. This finding is strikingly similar to that seen in Experiment 1. For completeness, 9 subjects switched three times and 6 switched four times. No subjects switched either five or six times.

Experiments 1 and 2 Combined

A bigger challenge is to combine the predictions and results from both of these experiments; if the model is to be useful it should have explanatory power both within *and* across broadly similar subject pools and experiments. We now obtain a Pearson correlation coefficient of predicted against actual cooperation across the 65 games of **r** = +0.957 (giving $\mathbf{R}^2$ = 0.916), significant at any level. An OLS regression of actual on predicted cooperation now finds:

$$\text{Actual } \%\mathbf{A} = 1.39 + 0.974\mathbf{x} \qquad F = 690.23$$

$$(t = 0.71) \; (t = 26.27)$$

That is, as **x** varies from 0% to 93.5% (see Tables 3 and 4, column 4); cooperation is predicted to increase from 1.4% to 92.5%. Actual cooperation in fact varies from 6.8% to 91.2%. Using the GLM, the comparable predictions are 8.7% to 88.7%. The partial effect at the sample mean is 1.175, compared with 0.974 for the fixed slope in OLS.

The two games from Experiment 1 also used in Experiment 2 reveal very similar aggregate behaviour: just a 1 point and a 5 point difference in %**A**, on a par with the games played twice *within* an experiment and subject pool. Overall, 11 games were used twice, either within or across the two experiments. The mean absolute difference in percentage cooperation is 4.7 percentage points, low enough to suggest predictable and broadly replicable cooperation levels within and across similar subject pools for this class of games.

A scatter plot of %**A** against *c** for all 65 games is shown in Figure 6. The scatter plot, which separately identifies the experiment and game type for each observation, provides visual evidence of predictive success within and across game types. The overwhelming impression remains of how closely our combination of models (1) and (2) capture variations in actual cooperation in a population, both within and across game types.

---Figure 6 here---

5. Discussion and Conclusion

a) Related Models

i) The model proposed in (2) for circumspect we-thinkers can be linked to Bergstrom's (2002) evolutionary "assortative matching" model. Bergstrom (2002 p.212-14) defines x as the proportion of evolutionarily 'hard-wired' cooperators, or 'C-strategists', and (1-x) as the proportion of 'D-strategists'. He then defines $p(x)$ as the conditional probability that a cooperator is encountered, given that one is a cooperator, and $q(x)$ as the conditional probability a cooperator is encountered, given that one is a defector. This yields [1-$p(x)$] as the probability a C-strategist encounters a D-strategist and [1-$q(x)$] as the probability a D-strategist encounters a D-strategist.

He then introduces an "index of assortativity", $a(x) = p(x)-q(x)$, which he defines as "...the difference between the probability that one meets one's own type and the probability that a member of the other type meets one's own type". Bergstrom goes on to show from the difference between the expected values of cooperation and defection, that under assortative matching, the growth rate of the proportion of cooperators in the population will be given by:

$$\delta(x) = (\mathbf{S} - \mathbf{P}) + a(x)(\mathbf{R}\text{-}\mathbf{S}) + x(1\text{-}a(x))[(\mathbf{R}+\mathbf{P}) - (\mathbf{S}+\mathbf{T})] \qquad (3)$$

Suppose we set $\delta(x) = 0$, thereby equating the expected success of cooperation and defection, and then rearrange (3) to solve for $a(x)$. We can derive a threshold value for the assortative matching index for each game, based upon that game's payoffs, **R**, **S**, **P** and **T**, and the proportion x. After some algebra, re-interpretation of the variables and translation of notations, the threshold value for $a(x)$ can be shown to be identical to that which we found for c* in (2). So while the motivation, aims and interpretations differ, our models can be shown to share a common mathematical structure.[16]

[16] There is a still closer analogy between the hypothetical example given in Section 3b) and Bergstrom's model.

ii) Bacharach's (2006) model shares our goal of describing how circumspect we-thinking might be implemented in practice; that is, can it accurately describe variations in the *extent* of observed cooperation and coordination across a diversity of encounters that constitute the root of the puzzle? Bacharach took circumspect we-reasoning to 'function' (i.e., the [**A**, **A**] equilibrium will prevail over the [**B**, **B**] equilibrium) in an 'unreliable coordination context' with a probability ω (see 2006, p.132). Expressed in terms of the game's payoffs and translating to our notation, Bacharach's threshold value ω* can be rewritten as follows:

$$\omega^{*} = \frac{2P - (S+T)}{(R+P) - (S+T)} \qquad (4)$$

So, if the probability ω exceeds the threshold value ω* in (4), he predicts outcome [**A**, **A**] will prevail. The greater is ω*, the more likely the unreliable coordination context is to fail to function and so outcome [**B**, **B**] will prevail. Bacharach also suggests that ω can be interpreted as the probability a representative individual group-identifies (and therefore we-reasons) in some game. When ω is bounded between 0 and 1 it can be compared with model (2)'s parameter *c* for prediction purposes. But when the numerator is zero or negative, he adds the claim that all we-thinkers will bring about [**A**, **A**].

However, while the present paper endorses and in general draws upon the essence of Bacharach's ideas, this is not so for his model of circumspect team-reasoning. Bacharach's model does not lead to a ratio capable of a cost-benefit interpretation; indeed, for PD games with 'additive' payoffs, Bacharach's denominator is always zero, leaving his ω* undefined. More generally, when $2P \leq S + T$ [17], model (4) predicts 100% of we-thinkers cooperate every time in

[17] Bacharach (2006, p.152) explicitly recognises his model requires cooperation when this inequality holds, even if ω=0.

that broad range of PD games. If we assume half the population we-reasons in these games, then the overall level of cooperation would be fairly stable at about 50% of players across a broad sweep of parameter values within the PD structure, a prediction strongly contradicted by the data reviewed in Section 4.

5b) Other Approaches

The agency question has been approached from other angles by authors such as Ainslie (1992) and Ross (2007). Ainslie argues that we have sub-personal interests within the interests of the whole person and sometimes they may dominate our behaviour contrary to the long-term goals of the individual. For instance when making long-range plans I may set myself a weight-loss target for the New Year. But each day my short-run self must decide whether to cooperate with the plan or defect and eat junk food. A possible application of circumspect we-reasoning to these questions would see the short-run self taking the interests of the set of selves into account to degree c in (2). Some days may confront the short-run self with modest temptation, a low c^*, or severe temptation, a high c^*. The team in this case would be the set of sub-personal agents and 'we' is the individual's long-term interests.

Ross (2007) builds upon Ainslie's approach to offer a defence of the theorems game theory generates. Instead of applying to an individual person, he argues the theorems apply as a tautology to the interactions of agents and those agents are not human individuals. He maintains that what GT can prove for these sub-personal agents still matters to us, even though those results do not directly apply to the whole human individual. His approach also is not obviously inconsistent with the perspective of the current paper, although different issues have been addressed to date.

Other existing theories seem designed to apply to either a much broader (e.g., Tan and Zizzo, 2008) or a much narrower class of games than we require. For example, Dufwenberg and Kirchsteiger (2004) present a sophisticated model within the framework of psychological game theory, which was inspired by, and is an extension of, Rabin's (1993) paper. Their theory introduces the concept of 'sequential reciprocity equilibrium' (SRE) for a broad set of games. But when applied to our narrower set of simultaneous-move games, the SRE predictions are not well-suited to explaining variations in the frequency of cooperation observed in experiments 1 and 2.

Other authors such as Rappoport and Chammah (1965) propose a simple formula to predict cooperation, but only in PD games. Translating to our notation (and inverting the ratio for comparability with c^*) they used: $\mathbf{r_1} = (\mathbf{R}\text{-}\mathbf{P}) / (\mathbf{T}\text{-}\mathbf{S})$. While *ad hoc*, empirically this ratio captures the tensions in the PD game fairly well; however it is not sufficiently general for our purposes. Roth and Murnighan (1978) proposed an alternative ratio, in our notation: $\mathbf{e_1} = (\mathbf{T}\text{-}\mathbf{R}) / (\mathbf{T}\text{-}\mathbf{P})$. This ratio is equivalent to $c^*(p)$ if and only if $p = 1$, but once again it is not sufficiently general for our class of games.

5c) Conclusion

In this paper I build on Bacharach's insight that in some games a player's perceptions, as triggered by the configuration of individual payoffs, may transform their method of reasoning from the payoffs to a decision via agency transformation. As an explanation for actual choice patterns, Bacharach's predictions for when the [**A**, **A**] outcome will prevail are not consistent with observation. The current paper therefore proposes a different model of circumspect we-reasoning to capture the way such players balance the competing motives.

A related aim of this paper has been to derive predictions for the extent of cooperation in a given game so we might understand when cooperation is most and least likely to be sustained in human society. We find strong support from two experiments for our predictions. It appears likely that just one underlying cause of ‘excess’ cooperation spans the set of these games, rather than a series of different ‘causes’ each confined to a specific game-type. This ‘economy of scope’ for we-reasoning in turn offers support to the proposition that the capacity for we-reasoning has an adaptive origin.

In recent years scientists discovered that the bumblebee’s wing stroke creates a vortex effect, generating extra lift that had gone unrecognised in the physical laws used for calculations in earlier decades (Altshuler *et al*, 2005). These authors comment that since Magnan (1934), “…bees have symbolized both the inadequacy of aerodynamic theory as applied to animals and the hubris with which theoreticians analyse the natural world”. Although the ‘hubris’ of the theoreticians declined in the decades after Magnan, the paradox of bee flight was confirmed as recently as Ellington (1984) and only fully solved with Altshuler *et al*. This example provides us with a reminder to respect empirical facts and a caution against overconfidence in any prevailing scientific paradigm, including IIR. Might a deeper understanding of the possible levels of agency embodied during choice take us close to resolving the long-standing paradox of successful human cooperation and coordination? This paper aims to be a step towards that goal.

References

Ainslie, G. (1992), Picoeconomics, Cambridge University Press.

Altshuler, D., Dickson, W., Vance, J., Roberts, S. and M. Dickinson (2005), "Short-Amplitude High-Frequency Wing Strokes Determine the Aerodynamics of Honeybee Flight", *Proceedings of the National Academy of Sciences,* 102, 18213-18218.

Bacharach, M., (2006), "Beyond Individual Choice: Teams and Frames in Game Theory" edited by N. Gold and R. Sugden, Princeton University Press, Princeton and Oxford.

Bergstrom, T., (2003), "The Algebra of Assortative Encounters and the Evolution of Cooperation", *International Game Theory Review*, 5, 211-228.

Binmore, K., (1992), "Foundations of Game Theory", in Advances in Economic Theory, edited by J.J. Laffont, Cambridge University Press.

Brosig, J., (2002), "Identifying Cooperative Behavior: Some Experimental Results in a Prisoner's Dilemma Game", *Journal of Economic Behavior and Organization*, 47, 275-290.

Butler, D., Burbank, V. and J. Chisholm (2011), "The Frames behind the Games", *Journal of Socioeconomics,* 40, 103-114.

Butler, D. and G. Loomes (2007), "Imprecision as an Account of the Preference Reversal Phenomenon", *American Economic Review,* 97, 277-297.

Colman, A., (1995), Game Theory and its Applications in the Social and Biological Sciences, (2nd ed), Butterworth-Heinemann.

De Cremer, D. and Van Vugt, M. (1999), "Social Identification Effects in Social Dilemmas: a Transformation of Motives", *European Journal of Social Psychology*, 29, 871-893.

Dufwenberg, M. and Kirchsteiger, G. (2004), "A Theory of Sequential Reciprocity", *Games and Economic Behavior,* 47, 268-298.

Dunbar, R. and Shultz, S., (2007), “Evolution in the Social Brain”, *Science*, 317, 1344-1347.

Ellington, C.P., (1984), “The aerodynamics of hovering insect flight”, *Philosophical Transactions of the Royal Society of London B*, 305, 1-15.

Gauthier, D. (1986), Morals by Agreement, Oxford: Clarendon Press.

Gauthier, D. and R. Sugden, eds., (1993), Rationality, Justice and the Social Contract: Themes from ‘Morals by Agreement’. London: Harvester Wheatsheaf.

Gold, N. and Sugden, R., (2007), “Collective Intentions and Team Agency”, *Journal of Philosophy*, 104, 109-137.

Hamilton, W. D., (1964), “The Genetical Evolution of Social Behavior”, *Journal of Theoretical Biology*, 7, 1-52.

Hollis, M. and Sugden, R., (1993), “Rationality in Action”, *Mind*, 102, 1-35.

List, J. (2006), “Friend or Foe? A Natural Experiment of the Prisoner’s Dilemma”, *Review of Economics and Statistics*, 88, 463-471.

Magnan, A. (1934), “La Locomotion Chez les Animaux”, Hermann, Paris, Vol.1.

Marwell, G. and D.R. Schmitt, (1975), Co-operation: An Experimental Analysis, (New York, Academic Press).

Nowak, M. (2006), “Five Rules for the Evolution of Cooperation”, *Science*, 314, 1560-1563.

Nozick, R., (1993) The Nature of Rationality, (Princeton: Princeton University Press).

Price, M., Cosmides, L. and J. Tooby (2002), “Punitive Sentiment as an Anti-Free Rider Psychological Device”, *Evolution and Human Behavior*, 23, 203-231.

Rabin, M., (1993), “Incorporating Fairness into Game Theory”, *American Economic Review*, 83, 1281-1301.

Rohrbaugh, M., Mehl, M., Shoham, V., Reilly, E. and Ewy, G., (2008), "Prognostic Significance of Spouse *We* Talk in Couples Coping with Heart Failure, *Journal of Consulting and Clinical Psychology*, 76, 781-789.

Ross, D. (2007), Microexplanations, MIT Press, Cambridge.

Sally, D. (1995), "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958-1992", *Rationality and Society*, 7, 58-92.

Savage, L.J. (1954), The Foundations of Statistics, New York: John Wiley.

Simpson, B. (2006), "Social Identity and Cooperation in Social Dilemmas", *Rationality and Society*, 18, 443-470.

Sugden, R. (2003), "The Logic of Team Reasoning", *Philosophical Explorations*, 6, 165-181.

Thaler, R. and Camerer, C., (2003), "In Honor of Matthew Rabin: Winner of the John Bates Clark Medal", *Journal of Economic Perspectives*, 17, 159-176.

Tomasello, M., Carpenter, M., Call, J., Behne, T. and Moll, H. (2005), "Understanding and Sharing Intentions: The Origins of Cultural Cognition", *Behavioral and Brain Sciences*, 28, 675-691.

Tversky, A. and D. Kahneman (1992), "Advances in Prospect Theory: Cumulative Representation of Uncertainty", *Journal of Risk and Uncertainty,* 5, 297-323.

Van Huyk, J., Battalio, R., and Beil, R., (1990), "Tacit Coordination Games, Strategic Uncertainty and Coordination Failure", *American Economic Review*, 80, 234-248.

Van Vugt, M. and Van Lange, P. (2006), "The Altruism Puzzle: Psychological Adaptations for Pro-social Behavior", in M. Schaller, D. Kenrick, & J. Simpson (Eds.), Evolution and Social Psychology, Psychology Press, pp.237-261.

Table 1: The Games Used in Experiment 1

Game	Payoffs		Game Type
1.	12,12	0,10	Stag Hunt
	10,0	7,7	
2.	6,6	0,13	Prisoner’s Dilemma
	13,0	4,4	
3.	12,12	2,5	Tender Trap
	5,2	7,7	
4.	8,8	6,14	Chicken
	14,6	4,4	
5.	6,6	0,15	Prisoner’s Dilemma
	15,0	5,5	
6.	14,14	3,8	Tender Trap
	8,3	11,11	
7.	14,14	2,18	Prisoner’s Dilemma
	18,2	5,5	
8.	12,12	4,5	Tender Trap
	5,4	7,7	
9.	10,10	6,14	Chicken
	14,6	4,4	

Table 1 (continued)

10.	9,9	0,7	Tender Trap
	7,0	8,8	
11.	12,12	0,16	Prisoner’s Dilemma
	16,0	3,3	
12.	10,10	0,9	Stag Hunt
	9,0	8,8	
13.	6,6	0,12	PD/NC Game
	12,0	6,6	
14.	12,12	8,16	Chicken
	16,8	6,6	
15.	8,8	0,10	Prisoner’s Dilemma
	10,0	5,5	
16.	12,12	0,10	Stag Hunt
	10,0	8,8	
17.	8,8	2,11	Prisoner’s Dilemma
	11,2	6,6	
18.	12,12	1,6	Tender Trap
	6,1	9,9	
19.	11,11	7,12	Chicken
	12,7	0,0	

Table 1 (continued)

20.	7,7	5,16	Chicken
	16,5	4,4	
21.	8,8	0,10	Prisoner’s Dilemma
	10,0	5,5	
22.	12,12	0,7	SH/TT Game
	7,0	7,7	
23.	10,10	6,14	Chicken
	14,6	4,4	
24.	10,10	0,16	Prisoner’s Dilemma
	16,0	4,4	
25.	10,10	0,9	Stag Hunt
	9,0	8,8	

Table 3: Predictions and Results for Experiment 1

Game	c*(p)	We-Frame (c*(p)) Predicted %A	Me-Frame (p*) Predicted %A	Overall Predicted %A	Experiment 1 Results %A
1.	$\frac{7-9p}{12-9p}$	70.8%	22.2%	70.8%	60.5%
2.	$\frac{4+3p}{6+3p}$	27.0%	0%	13.5%	11.1%
3.	$\frac{5-12p}{10-12p}$	87.2%	58.4%	87.2%	86.1%
4.	$\frac{8p-2}{8p+2}$	70.9%	25%	47.9%	43.2%
5.	$\frac{4p+5}{4p+6}$	12.8%	0%	6.4%	12.6%
6.	$\frac{8-14p}{11-14p}$	70.7%	42.8%	70.7%	73.1%
7.	$\frac{3+p}{12+p}$	72.0%	0%	36.0%	39.5%
8.	$\frac{3-10p}{8-10p}$	93.5%	70%	93.5%	91.2%
9.	$\frac{6p-2}{6p+4}$	84.4%	33.3%	58.8%	65.8%
10.	$\frac{8-10p}{9-10p}$	42.0%	20%	42.0%	50.6%
11.	$\frac{3+p}{12+p}$	72.0%	0%	36.0%	35.8%
12.	$\frac{8-9p}{10-9p}$	46.9%	11.1%	46.9%	50.0%
13.	1	0	0%	0.0%	8.8%

14.	$\frac{6p-2}{6p+4}$	84.4%	33.3%	58.8%	54.3%
15.	$\frac{5-3p}{8-3p}$	47.0%	0%	23.5%	22.2%
16.	$\frac{8-10p}{12-10p}$	63.9%	20%	63.9%	50.6%
17.	$\frac{4-p}{6-p}$	36.5%	0%	18.2%	17.3%
18.	$\frac{8-14p}{11-14p}$	70.7%	42.8%	70.7%	60.5%
19.	$\frac{8p-7}{8p+4}$	99.5%	87.5%	93.5%	90.1%
20.	$\frac{10p-1}{10p+2}$	51.6%	10%	30.8%	22.5%
21.	$\frac{5-3p}{8-3p}$	47.0%	0%	23.5%	23.7%
22.	$\frac{7-12p}{12-12p}$	78.1%	41.6%	78.1%	56.8%
23.	$\frac{6p-2}{6p+4}$	84.4%	33.3%	58.8%	57.5%
24.	$\frac{2p+4}{2p+10}$	54.7%	0%	27.3%	21.0%
25.	$\frac{8-9p}{10-9p}$	46.9%	11.1%	46.9%	43.2%

Table 4: Predictions and Results for Experiment 2

Game	c*(p)	We-Frame (c*(p)) Predicted %A	Me-Frame (p*) Predicted %A	Overall Predicted %A	Experiment 2 Results %A
1.	$\frac{1}{2}$	50%	0%	25%	34%
2.	$\frac{11p-2}{11p+1}$	56%	18.2%	37.1%	34%
3.	$\frac{2}{3}$	33.4%	0%	16.7%	35%
4.	$\frac{8p-4}{8p+4}$	90.5%	50%	70.2%	76.7%
5.	$\frac{4p+5}{2p+7}$	13%	0%	6.5%	15.5%
6.	$\frac{6p-4}{6p+6}$	97.5%	66.6%	82%	82.5%
7.	$\frac{4p+5}{4p+6}$	12.8%	0%	6.4%	15.5%
8.	$\frac{11p-2}{11p+5}$	70.8%	18.2%	44.5%	67%
9.	$\frac{4p+5}{4p+6}$	12.8%	0%	6.4%	13.6%
10.	$\frac{8p-2}{8p+2}$	70.8%	25%	47.9%	48.5%
11.	$\frac{1}{3}$	66.6%	0%	33.3%	33%
12.	$\frac{11p-2}{11p+1}$	56%	18.2%	37.1%	35%
13.	$\frac{1}{10}$	90%	0%	45%	55.3%

14.	$\frac{2p-1}{2p+9}$	97.7%	50%	73.8%	88.3%
15.	$\frac{1}{2}$	50%	0%	25%	22.3%
16.	$\frac{14p-7}{14p+1}$	85.9%	50%	67.9%	72.8%
17.	$\frac{2}{3}$	33.3%	0%	16.7%	10.7%
18.	$\frac{8p-7}{8p+1}$	99.3%	87.5%	93.4%	89.3%
19.	$\frac{1}{4}$	75%	0%	37.5%	29.1%
20.	$\frac{2p-1}{2p+9}$	97.7%	50%	73.8%	77.7%
21.	$\frac{2}{3}$	33.3%	0%	16.7%	17.5%
22.	$\frac{12p-1}{12p+1}$	39.5%	8.3%	28%	15.5%
23.	$\frac{4p+5}{4p+7}$	22.6%	0%	11.3%	6.8%
24.	$\frac{4p-2}{4p+6}$	94.6%	50%	72.3%	82.5%
25.	$\frac{1}{4}$	75%	0%	37.5%	37.9%
26.	$\frac{4p-2}{4p+6}$	94.6%	50%	72.3%	78.6%
27.	$\frac{1}{2}$	50%	0%	25%	19.4%
28.	$\frac{8p-4}{8p+4}$	90.5%	50%	70.2%	75.7%
29.	$\frac{1}{3}$	66.6%	0%	33.3%	26.2%

30.	$\frac{6p-4}{6p+2}$	95.4%	66.6%	81%	79.6%
31.	$\frac{1}{10}$	90%	0%	45%	49.5%
32.	$\frac{4p-2}{4p+6}$	94.6%	50%	72.3%	77.7%
33.	$\frac{1}{2}$	50%	0%	25%	13.6%
34.	$\frac{8p-7}{8p+1}$	99.3%	87.5%	93.4%	88.3%
35.	$\frac{1}{3}$	66.6%	0%	33.3%	33%
36.	$\frac{11p-2}{11p+1}$	56%	18.2%	37.1%	30.1%
37.	$\frac{1}{3}$	66.6%	0%	33.3%	24.3%
38.	$\frac{14p-7}{14p+1}$	85.9%	50%	67.9%	74.8%
39.	$\frac{2}{3}$	33.3%	0%	16.7%	15.5%
40.	$\frac{6p-4}{6p+2}$	95.4%	66.6%	81%	84.5%

Figure 3: The standard model

		Player 2		
		A		**B**
Player 1	EV(**A**):	pR	+	$(1-p)$S
	EV(**B**):	pT	+	$(1-p)$P

Figure 4: The circumspect we-thinking model

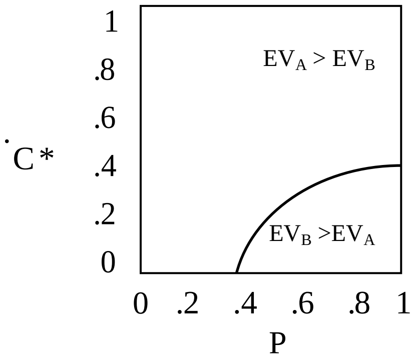
		Player 2		
		A		**B**
Player 1	EV(**A**):	$[p(1-c)+c]$R	+	$[1-p(1-c)-c]$S
	EV(**B**):	$[p(1-c)]$T	+	$[1-p(1-c)]$P

Figure 5: **Unit Diagrams in [*p*, *c*] Space**

Example 1:

	A	**B**
A	10, 10	6, 14
B	14, 6	4, 4

Therefore $C^* = \dfrac{6p - 2}{6p + 4}$ for equal EV



Example 2:

	A	**B**
A	10, 10	0, 12
B	12, 0	4, 4

Therefore $C^* = \dfrac{4 - 2p}{10 - 2p} =$ for equal EV

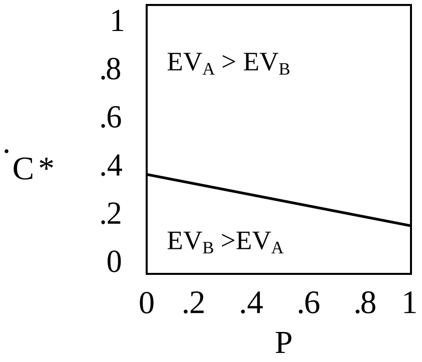


Figure 6

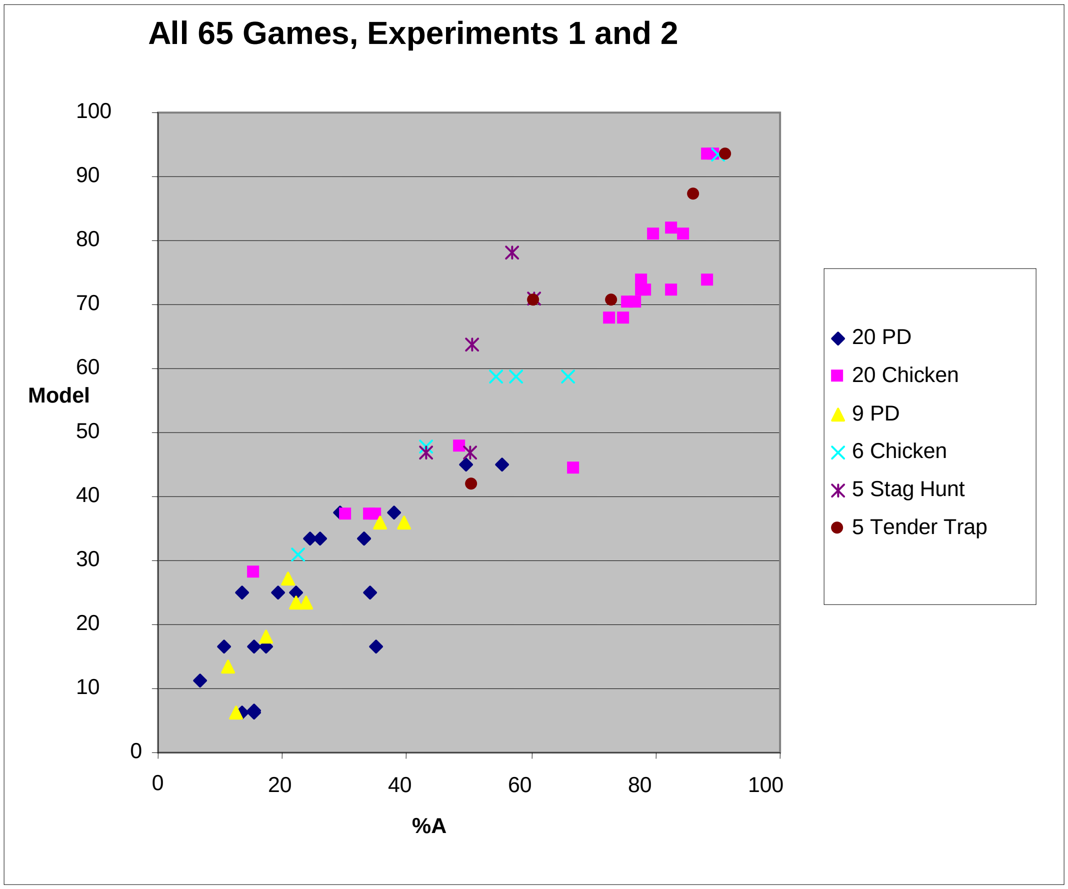

All 65 Games, Experiments 1 and 2
Model
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
%A
◆ 20 PD
■ 20 Chicken
▲ 9 PD
× 6 Chicken
✳ 5 Stag Hunt
● 5 Tender Trap

Table 2: Games from Experiment 2

Prisoner's Dilemma Games			**Chicken Games**		
1.	4, 4	0, 6	**2.**	5, 5	4, 14
	6, 0	2, 2		14, 4	2, 2
3.	6, 6	0, 10	**4**.	10, 10	6, 14
	10, 0	4, 4		14, 6	2, 2
5.	7, 7	0, 16	**6**.	10, 10	4, 12
	16, 0	5, 5		12, 4	0, 0
7.	8, 8	2, 17	**8**.	7, 7	2, 16
	17, 2	7, 7		16, 2	0, 0
9.	6, 6	0, 15	**10**.	8, 8	6, 14
	15, 0	5, 5		14, 6	4, 4
11.	6, 6	0, 8	**12.**	7, 7	6, 16
	8, 0	2, 2		16, 6	4, 4
13.	10, 10	0, 11	**14**.	10, 10	1, 11
	11, 0	1, 1		11, 1	0, 0
15.	8, 8	0, 12	**16.**	8, 8	7, 15
	12, 0	4, 4		15, 7	0, 0
17.	12, 12	0, 20	**18.**	8, 8	7, 9
	20, 0	8, 8		9, 7	0, 0
19.	8, 8	0, 10	**20.**	12, 12	3, 13
	10, 0	2, 2		13, 3	2, 2

Prisoner’s Dilemma Games

21.	10, 10	4, 14
	14, 4	8, 8
23.	9, 9	2, 18
	18, 2	7, 7
25.	10, 10	2, 12
	12, 2	4, 4
27.	8, 8	4, 10
	10, 4	6, 6
29.	10, 10	4, 12
	12, 4	6, 6
31.	12, 12	2, 13
	13, 2	3, 3
33.	8, 8	0, 12
	12, 0	4, 4
35.	12, 12	0, 16
	16, 0	4, 4
37.	6, 6	0, 8
	8, 0	2, 2
39.	10, 10	4, 14
	14, 4	8, 8

Chicken Games

22.	4, 4	3, 15
	15, 3	2, 2
24.	8, 8	2, 10
	10, 2	0, 0
26.	10, 10	4, 12
	12, 4	2, 2
28.	8, 8	4, 12
	12, 4	0, 0
30.	8, 8	6, 10
	10, 6	2, 2
32.	8, 8	2, 10
	10, 2	0, 0
34.	8, 8	7, 9
	9, 7	0, 0
36.	5, 5	4, 14
	14, 4	2, 2
38.	10, 10	9, 17
	17, 9	2, 2
40.	10, 10	8, 12
	12, 8	4, 4