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

By David J. Butler[#]

Economics Program,

UWA Business School,

University of Western Australia, Crawley, WA 6009

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

JEL classification numbers: C70, C91, D81

Abstract

This paper draws upon Bacharach's (2006) dual levels of agency to develop a model of how the circumspect aspect of we-reasoning functions for relevant games. The aim is to predict the extent of cooperation and coordination across a variety of game types. The model we propose yields a cost/benefit ratio threshold for players to choose cooperatively. Two experiments are run to provide data on some relevant hypotheses. The results of these experiments offer strong support both to the presence of circumspect we-reasoning and to our proposed model.

david.butler@uwa.edu.au Economics Program, UWA Business School, M251, 35 Stirling Highway, Perth, Australia.

*Financial support from the Business School's precursor at UWA is appreciated, and to a UWA-administered Australian Research Council small grant. Thanks also to Paul Miller for comments and to Dahai Fu for statistical assistance.

1. Introduction

It used to be said that while the laws of aerodynamics show the bumblebee is too heavy to fly, the bumblebee, being unaware of these laws, just carries on flying. A similar situation appears to exist in some strategic settings for game-theoretic logic versus successful human coordination and cooperation. Perhaps the classic example is a game called 'Dodo' (Binmore, 1992) or 'Hi-Lo' (Bacharach, 2006); I will refer to it as Hi-Lo but perhaps the most appropriate name is 'Bumblebee'.

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 simple game has been so widely debated is that while game theory can not provide a coherent reason for either player to choose **A** over **B** (see Bacharach 2006, Gold and Sugden 2007) human players find it almost self-evident to choose '**A**'. Indeed, Gold and Sugden (2007) go further:

> "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".

The problem does indeed go well beyond the simple Hi-Lo game. The set of games that interested Bacharach in his *magnum opus* (2006) are those this paper will address: the one-shot, simultaneous-move, symmetric, two-player games of *Chicken*, *Prisoner's Dilemma*, *Hi-Lo, Stag Hunt* and an alternate form of Stag Hunt sometimes called *Tender Trap*. The most famous of these, the one-shot Prisoner's Dilemma (PD) game, is of course a barren landscape for cooperative acts to take root; for that reason it is also an important benchmark case in which the rationality of cooperation has been vigorously debated. For example in 'Morals by Agreement' (1986) Gauthier develops a controversial model he calls 'constrained maximization' which promotes the view that cooperation in that game is the uniquely rational choice. A staunch opponent of this view, Binmore's (1993 p.133) comments on Gauthier's book include:

> "Gauthier's notion of constrained maximization requires departing from the most fundamental principle of non-cooperative game theory-namely, that players will not use a strongly dominated strategy."

Nor are obviously dominated options often chosen in other domains. Numerous studies of choice over lottery pairs have shown that despite the many systematic violations of Expected Utility 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 model of imprecise preferences, as does Blavatskyy (2007).

In a later comment, Binmore adds (1994, p.181):

> "…it makes as much sense to propose an alternative theory of rationality that justifies cooperation in the PD as it would to propose a replacement for traditional arithmetic in which 2+2=5".

The debate has involved numerous other authors; referring to the one-shot PD, Anatol Rapoport (1989 p.203) comments:

> "The different prescriptions of decisions based on individual and collective rationality in some conflict situations cast doubt on the very meaningfulness of the facile definition of 'rationality' as effective maximization of one's own expected gains".

But in the same volume Aumann (1989 p.23), although not addressing Rapoport, stridently opposes any effort to justify a cooperative choice in the one-shot PD game:

> "Worse than just nonsense, this is actually vicious, since it suggests that the prisoner's dilemma does not represent a real social problem that must be dealt with".

What, if anything, are game theorists such as Binmore and Aumann missing?[1] Game theory permits only the question 'what should *I* do' not one stemming from an alternate frame 'what should *we* do' (Sugden 2003, p.167). If I use the 'me-thinking' frame, my preferences, as represented by the payoffs, can be summarised just as they are written in the normal form matrix and will govern my choice in the standard way. If the payoffs in the matrix are dollars, other sources of utility such as fairness or altruism can transform these individual-level dollar payoffs, possibly to the extent that the game is no longer a true Prisoner's Dilemma. A desire to cooperate in this new game would not, of course, justify cooperation in a PD.

But suppose my response to the specific configuration of *individual* dollar payoffs in the game is to transform the level of agency I bring to the decision, by seeing it as a problem for *us*? Agency transformation is quite different to payoff transformation. There are profound conceptual difficulties involved in asserting we-motives can be included by transforming the payoffs such that a PD is for example really a Stag Hunt game. This point has been made previously, for example Hollis and Sugden write (1993, p.28):

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

And Nozick (1993, p.55):

[1] These authors are chosen simply because of their standing and clarity; they are much more representative of mainstream opinion than the contrary authors.

> "But if the *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."

Therefore, if it is my *attitude towards* the particular configuration of individual payoffs that is relevant, e.g., what my cooperation might *mean* to me in that situation, I can use we-thinking for my decision but *not* payoff-transformation[2]. Perhaps Gauthier had a similar idea in mind (Gauthier and Sugden 1993, p.186) when in his reply to Binmore 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 shares with Gauthier a view that standard game theory is not the only way to reason; an equally valid alternative mode (We-thinking) can also exist (Sugden 2003, Bacharach 2006). Unlike Gauthier, Bacharach does not say a cooperative disposition is *uniquely* rational, just that we-thinking and the choices it leads to are equally rational. 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

[2] Payoff transformation can not be a general solution anyway as it can not be used to explain the paradox of successful coordination in Hi-Lo.

dominance of 1-2% seen in choices over lottery pairs which also use dollar payoffs. The difference between these examples is so dramatic it suggests that maybe half of our subjects do *not* view cooperation in the context of a one-shot PD as a transparently dominated option, implying use of the We-frame. After establishing the underlying principle, Bacharach modifies we-reasoning to recognise that people won't always choose cooperatively even under a 'we'-frame because they lack assurance; this broader perspective he calls circumspect we-thinking. I shall assume hereon that circumspect we-thinking is the missing motivation that can be incorporated into these games only by transforming one's perception of the decision to a choice for 'us', and not by payoff-transformation.

2. Adaptive Origins

Bacharach (2006) refers to the enigma of human cooperation as a 'curate's egg', for good reason. Any cooperative instincts need to be nuanced and finely balanced against selfish concerns if they are to have survived the discipline of natural selection. Bacharach argues that we have to recognize the sheer diversity of encounters that early humans faced such that no one game will serve as an adequate model for all kinds and contexts of possible cooperation. As the evolution of a single, facultative, mechanism for generating cooperation in related game situations is more probable than selection for many game-specific special-purpose mechanisms, economies of scope in the application of this instinct are likely to span the set of games Bacharach identified.

Although the PD game is the only 2x2 game to present the individual/collective conflict in its purest form, this tension also exists in other games. Colman (1995) for

example argues the Chicken game is in many ways *more* suited to investigating cooperative versus competitive tensions than is the more famous PD game and Thaler and Camerer (2003 p.164) make a similar point. More generally still, Bacharach (2006, p. 111) claims:

> "I suggest that group identification is the key proximate mechanism in sustaining cooperative behaviour in man. More fully, I conjecture this: 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. The main virtue of this hypothesis over that of altruism and other contenders is that group identification is a more powerful *explanans* of the *diversity* of cooperative behaviours we see. 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".

Bacharach's perspective fits neatly within the increasingly popular 'social brain hypothesis' for why the human brain increased so dramatically in size (Dunbar and Shultz, 2007). This hypothesis notes that early man's ecological challenges were solved socially, requiring both coordination and flexibility in interactions with other group members. Similarly, Tomasello *et al* (2005 p.690) argues that underlying human cognition is "an adaptation for participating in collaborative activities involving shared

intentionality"; this shared intentionality they argue is unique to humans and it is probably this that makes we-thinking not simply possible, but natural.

Although similar affective attitudes likely underlie observed cooperation and coordination in our class of games, dispositions for altruism and equality may well draw upon subtly different emotions and instincts. For instance, Price *et al* (2002) argue that any mental module adapted for reciprocity would not also generalize to 'fairness' games. Similarly, inequity-aversion may perform best for games where social norms of equality and generosity, rather than group action, are suggested (e.g. the Dictator Game: Guala and Mittone 2009). 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. In other work (under review) by the author and colleagues we found that a player's contributions in one-shot Dictator or Ultimatum Games had almost no correlation with the extent of their cooperation in a set of one-shot PD and Chicken Games. Therefore like Bacharach, we will not be concerned in this paper with games that primarily invoke altruism or inequity-aversion rather than we-thinking.

How might an adaptation that promotes we-thinking actually work? Although referring to his own theory of reciprocal altruism (RA), Trivers (1985) argued that RA in humans evolved by "moulding our emotional responses to the cost/benefit calculus of social exchange". A possible way this could work is through Damasio's (1994) 'Somatic Marker Hypothesis' (SMH). Damasio's SMH maintains that listening to the "beacons and alarm bells" of the emotional responses of the body (soma) to choices is an integral part of rational decision-making under uncertainty. These affective attitudes likely draw upon the

dopamine system which is associated with anticipatory desire and the signalling of reward and danger (Tobler, Fiorillo and Schultz 2005). Damasio argues (p.128) that:

> "Nature appears to have built the apparatus of rationality [the cerebral cortex] not just on top of the apparatus of biological regulation [the limbic system], but also *from* it and *with* it".

But reciprocal altruism is not the only theory that may lead to a cost/benefit calculus by drawing upon the somatic marker pathway. If the feelings behind our intuitively we-thinking stance reflect adaptations embodying information accumulated over many thousands of generations of tribal life, rational choices will use and build upon the messages transmitted by these adaptive feelings. A study by Rilling et al (2002) using *f*MRI scans on subjects playing PD games found that individuals have "different patterns of (brain) activation depending on whether the playing partner was identified as a human or a computer". They also found increased activation in specific brain regions following mutual cooperation relative to either other outcome pairs or an equivalent (non-social) monetary reward. They conclude: "that [the relevant activation patterns] may relate specifically to cooperative social interactions with human partners". In summary, our somatic responses to specific individual-payoff combinations in these games could be nature's way of helping a self-interested species extract the long-run benefits of cooperation through the capacity for agency transformation.

3. The Game-Theoretic Framework

3a) The Scope

In Figure 1, let us define '**A**' as the cooperative and '**B**' as the defecting choice.

-----Figure 1 -----

The following payoff inequalities demarcate the scope of the theory for these games: **R**>**P** so that mutual cooperation exceeds mutual defection; **R**>**S** to ensure cooperating alone is risky; **T**>**S** so the off-diagonal cells penalise the cooperator. Of 24 possible rank orderings only five satisfy these inequalities. One of these, **R**>**T**>**S**>**P**, is trivial because cooperation is a dominant strategy for both me- and we-thinkers. This leaves four payoff rankings of 2x2 symmetric games where the relevant emotional states arise to trigger we-thinking. These are: *Prisoner's Dilemma* (**T**>**R**>**P**>**S**)[3], *Chicken* (**T**>**R**>**S**>**P**), *Stag Hunt* (**R**>**T**>**P**>**S**) and an alternate form of Stag Hunt sometimes (including here) called *Tender Trap* (**R**>**P**>**T**>**S**). Other games intermediate between these basic types, for example **T**>**R**>**P**=**S**, are also within the scope of the model. One such game, a limiting case of *Tender Trap*, is the *Hi-Lo* game from Section 1, where **R**>**P**>**T**=**S**=0.

Ideally, a single descriptive model of behaviour should encompass all of these games such that if the model predicts a 40% frequency of cooperation for some game, it

[3] Note the extra condition 2R > (T + S) is omitted, because our interests go well beyond the PD game. Even though the (asymmetric) 'group' payoff is not maximized through mutual cooperation if 2R < (T+S), 'we'-thinking is presumably still triggered for PD as in one-shot Chicken games, where often 2R < (T + S). See also footnote 11.

should not matter whether it was a PD, Stag Hunt, or some other game-type responsible for the prediction[4].

It is not essential to subscribe to the perspective developed in Section 2 to use the model we develop. Even if one were to view we-reasoning as a logical error, akin to a mistaken form of magical thinking, a model that predicts variations in the degree of such a 'bias' across this class of games would still be descriptively and practically useful. As an analogy, think of the way probabilities are transformed in decision theories such as cumulative prospect theory (Tversky and Kahneman 1992). Researchers differ on whether they view such transformations as rational, while still recognising the extra explanatory power those models have. Needless to say, we hope readers will be persuaded by all our arguments.

3b) The Models

In game theory we can assign a subjective probability to the strategy choice of the other player (e.g., Brandenburger, 1992). Let p denote her estimate of the chance the other player will choose option **A**, and $(1-p)$ for option **B**, where $p \in [0, 1]$ (see Figure 2).

- Figure 2 here -

The value of p that equates the expected values of options **A** and **B** so that she is indifferent between them, denoted by p^*, is the mixed-strategy Nash equilibrium in conjectures for the game:

[4] Hi-Lo was not tested in these experiments because the expectation of near universal cooperation has not been disputed.

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

If she believes either $p > p^*$ or $p < p^*$, then one or other pure strategy will offer a greater expected value. Let '**A**' be the cooperative choice. We can think of p as her estimate of the probability that a random other player will choose '**A**'. In these games a 'me'-thinker will use (1) to decide his choice. In a PD no matter how strong such a player's belief that others will play '**A**', it will always pay to defect and receive '**T**' in the 'Me' frame. If he is himself to cooperate in any one-shot PD game, a new model is needed, as no value of $p \in [0, 1]$ in (1) can ever make EV (**A**) > EV (**B**).

Now let us propose a simple adaptation of (1) to incorporate circumspect we-reasoning. For a player adopting this frame, the appeal of a cooperative choice in response to expected cooperation is increased, as is a defecting choice in response to expected defection. But how sure must one be that the other player will share one's 'we'-goal so as to justify cooperating? People are likely heterogeneous in this respect, differentially weighing the perceived risks and benefits at stake. In strongly conflicted games such as PD and Chicken, half of the players will be presumed to be me-thinkers, while the other half uses circumspect we-thinking. Whether this reflects the existence in evolutionary equilibrium of mixed 'types' (i.e., polymorphic phenotypes, see Sethi and Somanathan, 2003 or Kurzban and Houser, 2005), or perhaps just different personal life experiences is not clear, but also not central to this paper.

Let us assume a random we-thinker is drawn from a population represented by a continuum of circumspect we-thinkers. Denote this new dimension using a continuous

variable, c, and label it the 'We-assurance' dimension where $c \in [0, 1]$. We can then represent the expected value of choosing **A** and of choosing **B** for player 1 as follows.

----- Figure 3 -----

Our model should balance the We-frame payoffs (on the main diagonal) against the Me-frame (off-diagonal) payoffs. In such a model, if $c = 1$, either 'We' get **R** or 'We' get **P**, because then a player expects all other players to use the We-frame. Option **A** will be chosen as **R** > **P** in the games of interest in this paper. If $c = 0$ we are back to the standard model in (1). But the most important case is the contingent We-thinking player; as $0 < c < 1$, her circumspect approach leads her to balance her we-goals with one eye on self-protection, given her imperfect assurance regarding the other player's choice. In this case, 'c' increases the decision-weight she assigns to the diagonal 'We'-choice pairs, just as it lowers the decision-weight attached to the off-diagonal choice combinations, although not to zero. Her maximizing choice could then be either **A** or **B** depending on her beliefs and on the particular payoffs in the game.

The threshold value of c in a game, which we denote by c^*, is that for which EV (**A**) = EV (**B**). Using Figure 3 for player 1 and solving for c we get:

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

By symmetry (2) also holds for Player 2. When a We-thinking individual's $c > c^*$, we have EV (**A**) > EV (**B**), so her maximizing choice is to participate in bringing about [C,

C] over [D, D] by choosing option **A**. Similarly, if $c < c^*$, her assurance in that game is insufficient for her to cooperate so instead of playing her part in [C, C], she chooses option **B**. Ceteris paribus, the higher c^* is in any game, the smaller is the fraction of We-thinkers whose assurance satisfies $c > c^*$, and so a lower percentage of cooperative choices is predicted.[5]

Using equation (2) for specific payoffs of some game, we can simplify the resulting equation for threshold value, c^*. An important representation of these equations is seen by plotting $c^*(p)$ using a unit diagram in '*p*, *c*-space', see Figure 4. Here $c^*(p)$ shows the area above the equal EV line as the region where option **A** has the higher expected value to a player (as one's $c > c^*$), from below where option **B** would be preferred (as one's $c < c^*$). The simplest assumption we can make for prediction purposes is that when the 'we'-frame is used, c is distributed uniformly across the population.[6] If $c = 0$ the equal EV line intercepts the horizontal axis at the p^* derived in (1). For players using the 'Me'-frame, the optimal choice is given by comparing p with p^*. Again for simplicity of prediction, we will also assume a uniform distribution of p when the 'me'-frame is used. We then take the fraction of the unit diagram which lies *above* the EV line to derive the model's predicted cooperation rate for we-thinkers and the corresponding fraction of the horizontal axis for me-thinkers.

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

[5] Threshold value c^* is unique under a positive linear transformation of these payoffs.

[6] If the actual distribution is not uniform, deviations from the 45-degree line may occur, as they also may from general imprecision in choices.

Clearly, the equal EV line can have no horizontal intercept for PD games, as option **A** can never be optimal if a player's $c = 0$.[7] Interestingly, this model predicts a consistently high level of coordination in the Hi-Lo game (assuming all players use contingent We-reasoning). For Bacharach's example (2006, p.37) where **R**=5 and **P**=1, (**S**=**T**=0) our model predicts 98.2% will choose '**A**'. Holding the other payoffs constant but reducing **R** to 3 would lower predicted '**A**' choices to 95.3%, and if **R**=2 (as in Binmore 1992, p.4), to 89.8%. While we didn't test these predictions, they are not inconsistent with the coordination rates in excess of 90% suggested by Bacharach.

3c) Related Models

i) For the special case of the 'additive PD' game, (where **R** = β-α, **T** = β, **S** = -α and **P** = 0) we find (**T**-**R**) = (**P**-**S**) and (**T**-**P**) = (**R**-**S**). Sethi and Somanathan (2003, p.3) derive a cost/benefit ratio of α/β to allow cooperation in this special subset of PD games. Interestingly, using the model in (2), c^* yields not only precisely this same ratio for that narrow class of PD games, but it also applies to the non-additive PD games, as well as to all the relevant non-PD games Bacharach identified. The model in (2) can however be thought of analogously as player 1's loss from selecting the cooperative row as the numerator and player 1's gain from player 2 choosing the cooperative column as the denominator. Ceteris paribus, as the numerator falls (denominator rises), the costs of cooperation decline relative to the benefits, lowering the c^* threshold and raising the expected percentage of '**A**' choices. The model in (2) thus produces a threshold level for

[7] The new model would apply to 'me'-thinkers provided c% of players were not humans but computers programmed to simultaneously mirror back the choices of their selfish human co-players. But then the weights from Figure 3 would also be probabilities, given that causation between one's own choice and that of the other player would exist for this c% of the population.

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. A model which yields a cost to benefit threshold interpretation sits easily with the underlying logic of both evolutionary biology and economics.

ii) The model in (2) proposed for circumspect we-thinkers can also be connected 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', so (*1-x*) is 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". Notice that his *p(x)* is analogous to the decision weight for payoff **R** in Figure 3, as *q(x)* is for payoff **T**. The analogous terms before payoffs **S** and **P** follow straightforwardly from the requirement that the decision-weights for a choice sum to 1. 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 will be given by:

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

But if we set $\delta(x) = 0$, thereby equating the expected values of cooperation and defection and then rearrange (3) to solve for $a(x)$, we derive a threshold value for his matching index for each game, based upon that game's specific payoffs, **R**, **S**, **P** and **T**, and the proportion x. It turns out that after translating notations, the threshold value for $a(x)$ is identical to that which we found for c^*. So while the aims and applications of his research differ, the technology proposed in this paper to balance the conflicting motives facing circumspect team-thinkers, *might*[8] also have been derived from Bergstrom's evolutionarily hard-wired game-theoretic logic.

3d) Other Theories

i) Despite some formal links to Sethi and Somanathan (2003) and Bergstrom (2002) as discussed, the We-thinking motivation for our model is very different, as are the questions of interest. Bacharach's (2006) model however shares our goal of describing how the circumspect form of We-thinking might be implemented in practice; that is, can it accurately describe variations in the *extent* of observed cooperation that are at the root of the puzzle of successful cooperation and coordination?

Bacharach took circumspect we-reasoning to 'function' (i.e., the C, C equilibrium will prevail over the D, D equilibrium) in an 'unreliable coordination context' (as prevails in our experiments) 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:

[8] It wasn't; this was not Bergstrom's interest or purpose.

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

So, if the probability ω exceeds the threshold value ω* in (4), he predicts outcome [C, C] will prevail. The greater is ω*, the more likely is the unreliable coordination context to fail to function and so outcome [D, D] 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 our parameter *c* for prediction purposes. But when the numerator is zero or negative, he adds the claim that all we-thinkers will bring about [C, C].

However, while the present paper endorses and in general builds upon the essence of Bacharach's ideas, as for calculating the circumspect component of team-reasoning we propose the very different solution given in (3). What is more, Bacharach's model does not lead to a prediction capable of a cost-benefit interpretation. Indeed, for the important special case of additive PD games, Bacharach's denominator is always zero, leaving his ω* undefined. More generally, when $2P \leq S + T$[9], he predicts we-thinkers cooperate 100% in all additive PD games and many non-additive PD games. If we assume half the population we-reasons in these games, then the overall level of cooperation should be fairly stable at about 50% of players across a broad sweep of parameter values within the PD, a prediction strongly contradicted by the data.

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

Another problem for Bacharach is that his theory only allows a player to 'see' either the 'we'-frame or the me-frame for any particular problem; a player cannot "visualize switching frames" (Smerilli 2008). But as Smerilli argues, it is more likely that in many cases we *can* see a decision from both perspectives and we then weigh-up which perspective to employ.

ii) Dufwenberg and Kirchsteiger (2004) present a highly compelling model within the framework of psychological game theory, which was inspired by, and is a generalisation of, Rabin's (1993) classic paper. Their theory introduces the concept of 'sequential reciprocity equilibrium' (SRE), but they didn't design their model for, nor is it well-suited to, the task of explaining how observed cooperation and coordination rates vary with the parameters of these games.

For our set of games, Dufwenberg and Kirchsteiger's SRE concept can be rewritten in our notation and then solved for an individual's "sensitivity to reciprocity" parameter "Y_{12}". A player will then achieve the [C, C] SRE if her sensitivity to reciprocity parameter Y_{12} exceeds the threshold value:

$$Y_{12}^{*} = \frac{T - R}{2\left(\frac{R - S}{2}\right)^{2}} \qquad (5)$$

The greater is Y^{*}_{12} in some game, the fewer individuals should satisfy $Y_{12} > Y^{*}_{12}$ and so are less likely to achieve the [C, C] SRE. It transpires that the threshold value Y^{*}_{12} in (5) is not generally bounded between 0 and 1 for our set of games (e.g., Stag Hunt), making interpretation of the units and implications for prediction unclear. As with Bacharach's ω*, a

crude comparison of Y^*_{12} with observed %**A** for our set of games reveals little correlation; but as those authors made few claims regarding prediction it would be unfair to formally compare these models as they are currently expressed.

3d) Hypotheses:

For prediction purposes, we use the simple observation that in games with strongly conflicting pressures such as PD and Chicken, half the population will use the 'me'-frame while the other half will use circumspect we-thinking. In Stag Hunt and Tender Trap (also Hi-Lo) games, all subjects will be assumed to use circumspect we-thinking. These assumptions allow us to predict the *level* of cooperation in a given game. There is little to be gained by making minor subject-pool specific adjustments to these proportions. A minor increase in explanatory power would be outweighed by the loss in generality and simplicity. Therefore:

Claim 1: In PD and Chicken games we assume a 50:50 split between 'We' and 'Me' frames, but a 100:0 split for all other games.

Claim 2: Standard theory for players using 'me'-thinking requires $p \in [0, 1]$ and $c = 0$ and predicts *individual* maximizing choices according to $p<p^*$ and $p>p^*$

Claim 3: For players using circumspect we-thinking we require $p, c \in [0, 1]$ and compare the p, c pair to the equal EV line in the p, c unit diagram.

If one adaptive, facultative, mechanism underlies anomalous cooperation and coordination, we also expect no independent effect on the frequency of cooperative choices of the game-type within our set of games; only the prediction derived from the model matters.

Claim 4: Predictions should be independent of which game in the set of games is being played other than through the model.

While the assumptions of Claim 1 are robust for experimentalists' typical subject pools, they can easily be changed if factors known to affect the generation of emotions behind group identity, such as the contextual framing and 'social distance' an individual experiences in an experiment are substantially different. Also if anthropologists were to discover some population for whom we-thinking in PD and Chicken games is much more/less prevalent. These factors would affect the prevalence of we-thinking and therefore the level of cooperation/coordination achieved. Substantial alterations to the details of an experimental design or subject pool will then not detract from our model's predictive power, provided the proportion using we-reasoning is approximated.

As we-thinking is more likely to be prompted by symmetry of circumstance, only symmetric payoffs are used in the two experiments described in Section 4[10]. Players also had only intermediate social distance from each other. Our players also shared a common experience of being students seeking to win money from Professors. In these senses at least, they are a well-defined group sharing a common goal. These design choices were made because our model seeks to predict *variations* in the proportion of cooperative choices; if the setting were too sterile to trigger the feelings behind the emotions for we-thinking, e.g., double-blind, it is possible the proportions cooperating may have been too low.

[10] Evidence suggests that cooperation rates are lower in asymmetric PD games, and lower still as the asymmetry widens (Marwell and Schmitt, 1975). If anomalous cooperation is less common in society under asymmetric conditions, perhaps because fewer people then perceive the decision as a problem for 'us', it is not central to this paper.

4. The Experimental Design and Results

4a) The Experiments

Experiment 1 involved 81 subjects playing 25 2x2 one-shot games at the Economic Science Laboratory at the University of Arizona. The particular values of the payoffs in the games were chosen to reflect a variety of incentives to cooperate or defect, as measured by the model in this paper. The games used are listed in Table 1.

- Table 1 here-

Six sessions were run, each with 12-15 undergraduate subjects. No subject participated in more than one session. Subjects were seated at computer terminals in booths separated by screen walls. Each session began with a detailed introduction to the experiment, including an opportunity to ask questions. They worked through the introduction and sample games with the administrator, who projected his computer screen onto a large white screen visible to all. The same administrator ran all sessions, reading the instructions with the assembled subjects, to ensure common knowledge of the experimental conditions. An example of the experimental design can be seen in Appendix 1.

Each player was paired with another player in the lab at random 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. As the envelopes were not opened until all decisions were completed, the identity of the other player was concealed until after the completion of the experiment. In other words, there

were no opportunities for communication, feedback or learning, minimising any super-game effects.

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 drew a ticket from a box containing a number from 1-25. This denoted the number of the game to be played for real. Both players' responses to this game were retrieved, and they were paid according to the choice combination revealed. These 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.

After making their choice for each game, each player then has to answer four supplementary questions regarding that game. These supplementary questions are identical for all games and all players. The purpose of these was to gain some insight into a player's motives and beliefs, but we do not make formal use of that data in this paper. At the completion of the experiment we presented subjects with a list of several possible ways to best describe their play across the set of games and we report the results in Appendix 2.

Errors and/or incomplete preferences may occur in these experiments just as they do in tests of Expected Utility theory (e.g., Loomes and Sugden 1998). To observe any inherent variability in player's choices in experiment 1 three games were presented twice: one PD, one Stag Hunt and one Chicken game (in Table 1 compare games: 9 with 23, 12 with 25 and 15 with 21).

[11] If there were an odd number of players in a session, one random player 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.

Experiment 2 was run subsequently at the University of Western Australia. While the main focus of that experiment was to investigate individual differences in play which are not the topic of the current paper, we include the aggregate results from that set of games here to add to the weight of experimental evidence bearing upon the model developed here. Experiment 2 focused just on Chicken and PD games; there were 20 of each type of game, for a total of 40 games in all. 103 subjects played each of these 40 games. 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 close to crowded, this had little effect on privacy.

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

Six games were also repeated in experiment 2 to observe inherent variability in choices (in Table 4 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 1's games 4 and 5 were identical to experiment 2's games 10 and 9, so any difference in cooperation across subject pools can also be observed. Combining experiment 2's 40 games with the first experiment's 25 games, overall we have data from 65 games to investigate our claims.

4.2 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, taken from Table 1. The second column calculates the fraction of [p, c] space where the cooperative choice is optimal[12]. 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**) (always zero for PD games) and this is shown in column 3. In column 4 we then take half of column 2's fraction plus half of column 3's fraction as our prediction, 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 r = +0.966, significant at any level (giving an R^2 = 0.933 and adjusted R^2 of 0.930). Even more impressive is that not only the direction of change but the *level* of cooperation fits the predictions very well, given our assumption of uniformly distributed p and c in the population. As the model predicts a linear, one-to-one relationship with the data, a simple OLS regression can offer some insight. Regressing 'actual **A**' on the 'predicted **A**' gives:

$$\text{Actual } \%\mathbf{A} = 3.12 + 0.887x \qquad F=348.35$$

$$(t=1.47) \quad (t=18.66)$$

So as the predicted value (x) ranges from 0% to 93.5%, actual cooperation is predicted to vary from 3% to 86.1%, only slightly under-predicting the 8%-91% observed. As the dependent variable is a proportion, a generalised linear model (GLM) can take this restriction into account. 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

[12] All calculations of definite integrals were double-checked using Wolfram Alpha.

slope coefficient in OLS of 0.887. On an aggregate basis then, the new model performs very well in representing the balance of incentives to choose **A** or **B** within a game.

The three games played twice showed choice switching at an *individual* level at 28.2%, 12.5% and 8.7% of the time. These figures are typical for the choice instability rates found in most other tests of individual decision making (e.g., Loomes and Sugden 1998). But in aggregate, there are only minor differences of between 1.5 and 8.3 percentage points of cooperation suggesting predictability. As some preference imprecision or instability is unavoidable, the model appears to account for the extent of cooperation and coordination very well.

Appendix 2 lists several possible approaches to that experiment. Of our 81 players, 41 selected disposition (c) which comes closest to circumspect team-thinking, consistent with the ideas in this paper. There are perhaps two minor caveats to our assumptions (and/or our model). First, in the most severe PD games, achieved cooperation is consistently several points higher than our model predicts (also confirmed by the GLM estimate). There may be a few unconditional co-operators leading to this result, as is implied by the 5 players who selected disposition (b) (i.e., the distribution of c may have a small peak at $c = 1$ rather than being completely uniform).

Second, while only four of the 25 games have a cooperation level 10+ points different from our predictions, all four are SH or TT games and for all four we predicted a greater level of cooperation than actually occurred. The most likely reason is that unlike Hi-Lo, somewhat fewer than our assumption of 100% of people use we-thinking for every SH or TT game. While these caveats may well play a small but genuine part in the puzzle of human cooperation, for our purposes the loss in simplicity from incorporating them into our

assumptions outweighs the advantages of developing a parsimonious predictive model of cooperation and coordination.

Experiment 2

For the 40 PD and Chicken games in Experiment 2, we maintain the assumption that half the players use me-thinking (with a uniform distribution of p) and half we-thinking (with uniform distributions of p and c). The Pearson correlation coefficient across the 40 games is a very strong +0.961 (giving an R^2 =0.923 and adjusted R^2 of 0.921). A simple OLS regression of actual cooperation on predicted cooperation gives:

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

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

So as x varies from 6.4% to 93.4%, actual cooperation is predicted to go from 6.5% to 96.8%. Cooperation across these games actually varied 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.

In seven of the 40 games (three narrowly) actual cooperation was 10+ points different from the predicted level: 3 PD's and 4 Chicken games. While no obvious pattern is apparent in these games, one possibility is that contingent we-thinking is sometimes more and sometimes less common than our assumption of 50%. But once again, the advantages of simplicity, given the model's predictive power, obviate the need to pursue a more perfect but less parsimonious characterisation. The six games that were played twice also showed little aggregate variability: differences were again between 1 and 8 percentage points of cooperation, implying cooperation rates are broadly predictable, as in experiment 1. The

overwhelming impression however is of how closely actual cooperation follows the predicted level, notwithstanding some inevitable instability in preferences.

Experiments 1 and 2 Combined

Perhaps the sternest test is to combine the predictions and results from both of these experiments; if the model is to be truly useful it should have explanatory power both within *and* across broadly similar subject pools and experiments. The Pearson correlation coefficient of predicted and actual cooperation/coordination across the 65 games is a very high +0.957 (giving an $R^2 = 0.916$ and adjusted $R^2 = 0.915$). An OLS regression of actual on predicted cooperation now gives:

Actual %**A** = 1.39 + 0.974x F=690.23

(t=0.71) (t=26.27)

That is, as x varies from 0% to 93.5%, cooperation is predicted to increase from 1.4% to 92.5%. Actual cooperation in fact varied 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 partial effect rises above the OLS slope after cooperation of around 30% and falls back below it beyond about 70% cooperation.

The two games from experiment 1 that were included also in experiment 2 show very similar results: just a 1 point and a 5 point difference in %**A**, on a par with our games played twice *within* an experiment. In total therefore, 11 games were played twice, either within or across experiments. The average absolute value of differences in percentage cooperation is 4.7 percentage points, low enough to suggest predictable and

broadly replicable cooperation levels within and across subject pools in this class of games.

A scatter plot of %**A** against c* for all 65 games is shown in Figure 5. Figure 5, which separately identifies the experiments and game types of each game, offers support to Claim 4 and to the predictive success of the model across these types of games. The proportion of 'me'-thinkers is also consistent with previous research, as the highest cooperation rate we could achieve in a PD was 55%, implying close to half the players are best described by the 'me'-frame in PD games.

Discussion and Conclusion

The primary purpose of this paper has been to propose a model of the decision process players might use when balancing the individual and group goals that underlie variations in the extent of cooperation and coordination achieved. The model proposed here assumes in some games a player's feelings subconsciously triggered by such choices will frame the decision in a manner proposed by Bacharach's theory of circumspect we-reasoning. But as an explanation for observed choice patterns in these games, his specific model of when the [C, C] outcome will prevail seems inadequate. Our proposed alternative model is otherwise consistent with Bacharach's theory, as modified.

We also find strong support from two experiments for the claim that the underlying cause of excess cooperation spans the full set of games discussed herein, as %**A** directly follows our model's predictions and not the game-type that generated the prediction. This 'economy of scope' for we-reasoning in turn offers support to the proposition that the capacity for we-thinking has an adaptive origin.

In recent years scientists discovered that the bumblebee's wing stroke creates a vortex effect, giving it an extra lift that had gone unrecognised by physicists in earlier decades (Altshuler *et al*, 2005). Might game theory now be close to resolving its own long-standing paradox of successful cooperation and coordination via recognition of We-thinking? We hope this paper will be seen as a contribution to that process.

References

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.

Aumann, R. (1989), "Game Theory", in The New Palgrave: Game Theory, pp.1-53, Macmillan Press, New York.

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 Journal of Game Theory Review*, 5, pp.211-228.

Binmore, K., "Playing Fair", (1994) Vol. 1 of Game Theory and the Social Contract, (MIT Press).

Blavatskyy, P. (2007) Stochastic Expected Utility Theory, *Journal of Risk and Uncertainty*, 34, 259-286.

Brandenburger, A., (1992) "Knowledge and Equilibrium in Games", *Journal of Economic Perspectives*, 6, 83-101.

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. 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.

Damasio, A., (1994) Descarte's Error: Emotion, Reason and the Human Brain, (New York, Putnam).

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

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

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

Guala, F. and Mittone, L. (2009), "Paradigmatic Experiments: the Dictator Game", *Journal of Socio-Economics*, doi:10.1016/j.socec.2009.05.007, in press.

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

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

Kurzban, R. and D. Houser (2005) "An Experimental Investigation of Cooperative Types in Human Groups: A complement to Evolutionary Theory and Simulations", *Proceedings of the National Academy of Sciences*, 102, 1802-1807.

Loomes, G. and R. Sugden, (1998), "Testing Different Stochastic Specifications of Risky Choice", *Economica*, 65, 581-598.

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

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.

Rapoport, A. (1989), "Prisoner's Dilemma", in The New Palgrave: Game Theory, pp.199-204, Macmillan Press, New York.

Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G. And C. Kilts (2002) "A Neural Basis for Cooperation", *Neuron*, 35, 395-405.

Sethi, R. and E. Somanathan, (2003) "Understanding Reciprocity", *Journal of Economic Behavior and Organization*, 50, 1-27.

Shafir, E. and A. Tversky, (1992) "Thinking Through Uncertainty: Non-Consequential Reasoning and Choice", *Cognitive Psychology*, 24, 449-474.

Smerilli, A. (2008), "We-thinking and double-crossing: frames, reasoning and equilibria", *MPRA* paper no. 11545.

Sugden, R., (1993) "Thinking as a Team: Towards an Explanation of Non-Selfish Behaviour", *Social Philosophy and Policy*, 10, 69-89.

Sugden, R., (2000) "Team Preferences", *Economics and Philosophy*, 16, 175-204.

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

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

Tobler, P., Fiorillo, C. and W. Schultz, (2005), "Adaptive Coding of Reward Value by Dopamine Neurons", *Science*, 307, 1642-1645.

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, pp.675-691.

Trivers, R.L., (1971) "The Evolution of Reciprocal Altruism", *Quarterly Review of Biology*, 46, 35-57.

Trivers, R.L., (1985) Social Evolution, (Menlo Park: Benjamin/Cummings).

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

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 1: The 2x2 Game

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

Figure 2: The standard model

		A		**B**
Player 1	EV(**A**):	*p*R	+	(1-*p*)S
	EV(**B**):	*p*T	+	(1-*p*)P

Figure 3: The circumspect we-thinking model

		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

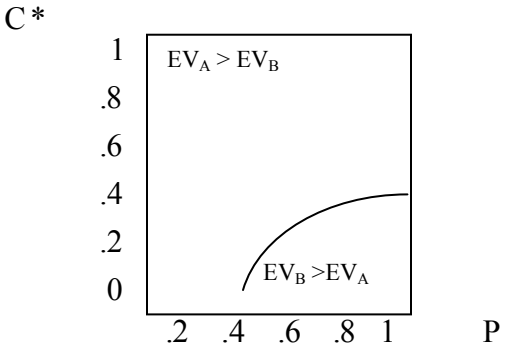
Figure 4: **Unit Diagrams in [*p, c*] Space**

Example 1:

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

$\therefore R = 10 \quad T = 14 \quad P = 4 \quad S = 6$

From $C^* = \dfrac{p(T-R)+(1-p)(P-S)}{p(T-P)+(1-p)(R-S)}$ $\quad\therefore C^* = \dfrac{6p-2}{6p+4}$ for equal EV

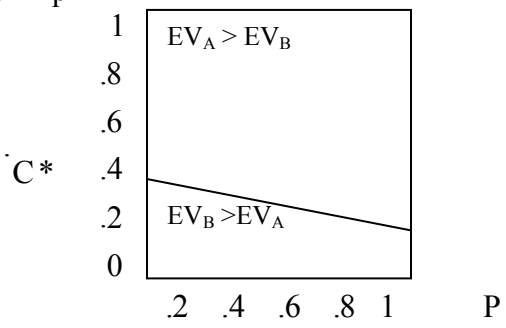


Example 2:

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

$\therefore R=10 \quad P=4 \quad T=12 \quad S=0$

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



APPENDIX 1

DECISION MAKING:
An Experiment using games

Introduction

In this experiment there are 25 problems. Each problem shows the payments, in francs (f), which result from various combinations of choices by two people, of whom you are one. An example follows:

OTHER / YOU	A	B
A	3,3	1,4
B	4,1	2,2

In this example, if you pick A and the other person also picks A, you both get f3. But if you pick B while the other person chooses A, then you get f4 and she/he gets f1. Similarly, if you choose A and the other person chooses B, you get f1 and she/he gets f4. But if both of you choose B, then you each get f2.

Instructions to Participants

You will not be informed of the other person's choices until the end of the experiment. Each problem has different payoffs from the others, so you should think carefully about the implications of your choice for each one. This is not an exam, and no answer is necessarily right or wrong, so don't feel pressured to force your choices to fit a pattern you do not feel expresses your own desires.

At the end of the experiment ONE problem will be selected at random and your decision played out for real. The payment you receive will then depend on your choice in that problem in conjunction with another person's choice in the same problem. That other person will be seeing exactly the same instructions and presentation of the question as you. The identity of the other person with whom you are paired is also determined randomly.

After you have made each of your choices, you will find a number of supplementary questions to answer. These are to help us interpret the way people make decisions. They are very important so please indicate your response as accurately as you can. Please ensure that you have included an answer to every question before you submit your response.

Check out the Sample Game.

DECISION MAKING:
An Experiment using games

Sample Game – Rules

OTHER YOU	A	B
A	10,10	6,12
B	12,6	4,4

If you both choose A, you each get f10, to be paid by the experimenter. If you both choose B, you each get f4. If you choose A when the other person chooses B, you will get f6 and she/he will get f12. Similarly, if you choose B when she/he chooses A, you will get f12 while she/he receives f6.

Game One – Answers

Your choice: A O or B O

Supplementaries:

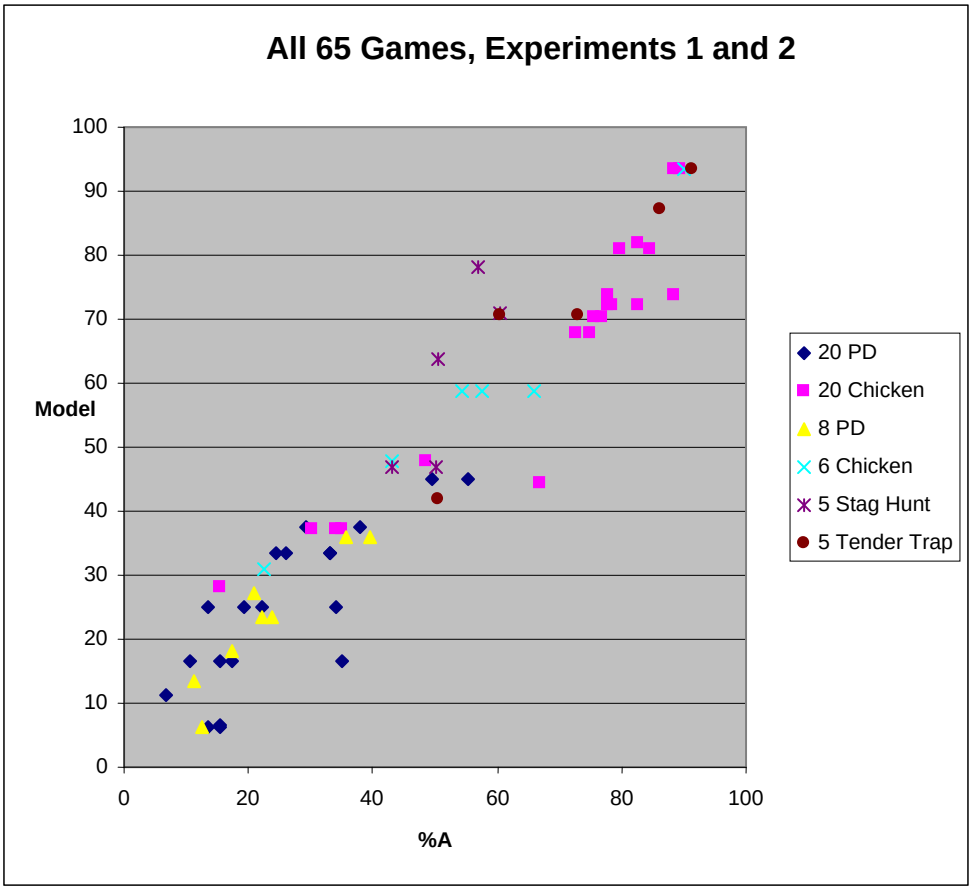
1. Suppose you had managed to see that the other person had chosen option 'A', please indicate which option you would now choose. Your hypothetical choice: A O or B O.

2. Suppose you had managed to see that the other person had chosen option 'B', please indicate which option you would now choose. Your hypothetical choice: A O or B O.

3. Indicate on the scale provided your actual belief about the choice of the other person. If you expect them to choose option 'A' with certainty, choose a '10'. If you expect them to choose option 'B' with certainty, choose a '0'. If you believe there is a 50:50 chance of them choosing 'A' or 'B', choose a '5'. Therefore, choose numbers greater than '5' if on balance you expect them to choose 'A' and numbers less than '5' if you expect them to choose 'B'. Please select your chosen number on the following scale:
   1 O 2 O 3 O 4 O 5 O 6 O 7 O 8 O 9 O 10 O

4. If you chose 'B' on this question, would you have still selected 'B' regardless of your beliefs about the other person's choice? If yes, click here: O. If no, or you chose 'A', indicate on the scale provided how confident you would have to be that the other person will choose option 'A' before you would have also chosen option 'A'. So, if you would select option 'A' even if you were very confident they had chosen option 'B', choose a '0'. If you would need to be very confident they would choose 'A' before you would also do so, choose a '10'. Similarly, if you would need to be at least 60% certain they had chosen 'A' before you would so, choose a '6', and so on.
   1 O 2 O 3 O 4 O 5 O 6 O 7 O 8 O 9 O 10 O

Figure 5


All 65 Games, Experiments 1 and 2
Model
%A
0
10
20
30
40
50
60
70
80
90
100
0
20
40
60
80
100
20 PD
20 Chicken
8 PD
6 Chicken
5 Stag Hunt
5 Tender Trap

Appendix 2

Choice of Dispositions in Experiment 1

Which of the following dispositions would you say best described your overall approach to these choice questions?

a) I unconditionally sought to maximise my own payoff, uninfluenced by the payoff the other would receive.
b) I always chose that option which was best for both of us.
c) I hoped to play my part in doing what was best for both of us. But whenever I was sufficiently unsure that the other person would share that view, I maximised my own individual payoff.
d) I wasn't trying to maximise my return in any of the above ways, I was happy to lower my own payoff in order to raise the payoff of the other person.
e) My chief aim was to win more than the other person, even if this meant we would both come out with less.
f) I consistently chose the option that did not provide me with the lowest of the four payoffs.
g) I rarely had any real preference in these games, I just wrote down something to get through the experiment.
h) Other, please specify below.

Results: a) 12; b) 5; c) 41; d) 1; e) 1; f) 11; g) 0; h) 10

Of the 10 players who chose h), 3 are similar to a), 3 are similar to c), 2 are similar to f) and 2 appear confused.

Table 2: Games from Experiment 2

Prisoner's Dilemma Games

1.	4, 4	0, 6
	6, 0	2, 2
3.	6, 6	0, 10
	10, 0	4, 4
5.	7, 7	0, 16
	0, 16	5, 5
7.	8, 8	2, 17
	17, 2	7, 7
9.	6, 6	0, 15
	15, 0	5, 5
11.	6, 6	0, 8
	8, 0	2, 2
13.	10, 10	0, 11
	11, 0	1, 1
15.	8, 8	0, 12
	12, 0	4, 4
17.	12, 12	0, 20
	20, 0	8, 8
19.	8, 8	0, 10
	10, 0	2, 2

Chicken Games

2.	5, 5	4, 14
	14, 4	2, 2
4.	10, 10	6, 14
	14, 6	2, 2
6.	10, 10	4, 12
	12, 4	0, 0
8.	7, 7	2, 16
	16, 2	0, 0
10.	8, 8	6, 14
	14, 6	4, 4
12.	7, 7	6, 16
	16, 6	4, 4
14.	10, 10	1, 11
	11, 1	0, 0
16.	8, 8	7, 15
	15, 7	0, 0
18.	8, 8	7, 9
	9, 7	0, 0
20.	12, 12	3, 13
	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