TEAM AGENCY AND CONDITIONAL GAMES

ANDRE HOFMEYR
University of Cape Town
andre.hofmeyr@uct.ac.za

DON ROSS
University of Cape Town
Georgia State University
don.ross931@gmail.com

Abstract

We consider motivations for acknowledging that people participate in multiple levels of economic agency. One of these levels is characterized in terms of subjective utility to the individual; another, frequently observed, level is characterized in terms of utility to social groups with which people (temporarily) identify. Following Bacharach (2006), we describe such groups as 'teams'. We review Bacharach's theory of such identification in his account of 'team reasoning'. While this conceptualization is useful, it applies only to processes supported by deliberation. As this is only one of a range of causal mechanisms underlying behaviour by humans and other strategic agents, a more general account is desirable. We then argue that Stirling's (2012) account of 'conditional games' achieves the desired generalization.

JEL codes: B41, C72, C79, D03, Z13

1. *Introduction*

All of economics is concerned with how some *agent* could do something better or best, in response to choices of other agents, resource constraints, and incentives defined as such by reference to the agent's goals. It is appropriate to refer to 'agents' rather than to 'people', because in many economic models the agents are firms, or households, or governments, or teams. Indeed, the overwhelming majority of economic applications concern aggregated responses (Ross 2014, Chapter 5). Notwithstanding this fact, economics is frequently associated, both by its critics and by many of its leading practitioners and textbook authors, with individualism, the view that individual people are in some sense the fundamental sites of agency on which others are dependent.

Economists typically do not give individualism an ontological interpretation, that is, as reflecting a metaphysical doctrine to the effect that all properties of non-individual agents must decompose into, or be functions of, individual human (or other animal) agents. Even if some economists, when they dabble in philosophy, adhere to such social atomism, 'official' individualism is usually held to be 'methodological', and might be expressed as the following constraint on economic model building: a sound economic model should not require any individual human agent to choose an action that is sub-optimal for her, given the choices of the other agents, without some (good) explanation (Ragot 2012).

Stated this way, methodological individualism as expressed in game-theoretic applications is the assumption that the solutions of all models involving individual human agents, either explicitly or implicitly, should be compatible with a non-cooperative Nash equilibrium of a game, that also models the interaction[1], and in which the individual people in question are the players. Binmore (1994) provides an explicit defense of this methodological principle.

Following Ross (2014), we distinguish two variants of substantive (i.e., not merely methodological) individualism. *Normative* individualism refers to the Enlightenment conviction that individuals, not groups, are the centres of human dignity and valuation that most *deserve* valorisation. In modern democracies this is a premise that liberals and conservatives generally share. It is typically assumed in welfare economics. *Descriptive* individualism, by contrast, refers to the view that people acquire their preferences asocially. Descriptive individualism is, in general, false: most human preferences, and almost all of the most important ones, are copied from other people or shaped under their guidance and tutelage. Individual human distinctiveness merits valorisation *because* its cultivation and maintenance is an *achievement* for members of a social species given to high levels of suggestiveness and conformity. Thus, far from being in tension with one another, normative individualism and descriptive anti-individualism make a naturally complementary pair.

The fact that people tend naturally to identify with social groups to which they belong, but simultaneously strive to operate and optimise individual utility functions, is a phenomenon that a fully adequate economic modeling apparatus should be able to represent. This is one of the aims of the *team reasoning* idea promoted by Martin Hollis (1998), Robert Sugden (1993, 2000, 2003) and Michael Bacharach (1999, 2006). In Bacharch's (2006) unfinished[2] treatise *Beyond Individual Choice: Teams and Frames in Game Theory*, he and his scholarly executors emphasize that most people are experienced in executing gestalt switches between individual and group agency, sometimes choosing in such a way as to maximise an individual utility function and sometimes choosing in such a way as to maximise the utility of a team with which they identify. Furthermore, people are often aware of this gestalt duality and can and do compare and weigh the alternatives suggested by each gestalt in specific circumstances.

Ross (2014) argues that this phenomenon is better characterised as team *agency* rather than team *reasoning,* because like most economic responses it only sometimes involves deliberate reflection. This is not to say that when people reflexively optimise the utility of a group rather than themselves this doesn't amount to a choice. There is generally *some* hypothetical incentive that could move a person to try, in a specific interaction, exclusively to optimise her self-interest. The point, then, is that some

[1] We assume that because no model ever completely describes an economic interaction or situation, interactions and situations can have multiple models that should be compatible with one another where their applications overlap. In the kind of example emphasised by Binmore (1994), bargaining scenarios are modeled as both cooperative and non-cooperative games for different purposes; but the cooperative solution must correspond to one of the Nash equilibria of the non-cooperative model.

[2] Bacharach was approximately halfway through composing the manuscript when he passed away unexpectedly in 2002. Sugden and Natalie Gold, one of Bacharach's PhD students at the time, edited his work and wrote introductory and concluding chapters so that it could be published post mortem.

chosen identifications do not result from reasoning, even though by definition all choice is motivated. But Bacharach (2006) and his executors use the phrase 'team reasoning' because they link the modeling problem to the rational solution of equilibrium selection problems in game theory.

In the chapter to follow, we will first summarise the team reasoning idea as Bacharach (2006) conceives it. However, we will then show that the effect of team reasoning on equilibrium selection in games is generalised, both conceptually and technically, by Wynn Stirling's (2012) modeling framework for *conditional games*. As with other games, conditional games might or might not be explicitly represented by their players; sometimes they might be selected and stabilised by processes of biological, but in humans more typically social and institutional, evolution. If Stirling generalises Bacharach where game theoretic representation is concerned, this can be seen as supporting Ross's (2014) suggestion that team reasoning is at best one special mechanism that supports team agency. If team reasoning sometimes goes on, discovery of the mechanisms that implement it falls within the domain of psychology rather than economics. What economists need to be able to model is team agency; and thanks to Stirling they now can.

2. *Equilibium selection and team reasoning*

Equilibrium selection problems in game theory arise from the fact that many games have multiple Nash equilibria (NE), but often some NE seem more 'sensible' and people in fact converge on them, even though the formal theory of choice that is built into game theory[3] includes no axioms or principles that recommend it. This property of NE, taken as a problem, motivated the *refinement* literature of the 1970s and 1980s (Kreps 1990), which sought to add restrictive axioms to solution concepts and thereby rule out 'inferior' NE as solutions. This approach threatened to degenerate into a programme for rationalising every distinct situation as a *sui generis* game, thus eviscerating the explanatory and predictive power of NE, and so was largely abandoned in the 1990s in favour of evolutionary and behavioural approaches to equilibrium selection. Behavioural models tend to restrict solutions by motivating bounds on people's rationality, whereas evolutionary models hardwire agents with strategies and rely on evolutionary dynamics to provide estimates of the likelihoods with which various equilibria will be played in a particular population.

In contrast to these approaches, Bacharach (2006) argues that when 'fully rational' individual people reason as members of teams, some equilibrium selection problems dissolve. He focuses on three types of game to illustrate and defend his general proposal.

The first type is the pure coordination game, for which the strategic form is presented in Table I. Players 1 and 2 simultaneously choose "Heads" or "Tails." If the labels match (i.e., Heads and Heads or Tails and Tails) the players each receive their highest-valued outcome. If the labels do not match, each player receives an outcome with lower utility. The game has two pure strategy NE, (Heads, Heads) and (Tails, Tails). In experiments literally involving coins, people tend to converge on (Heads,

[3] See Binmore (2009).

Heads) (Mehta et al, 1994). This suggests that (Heads, Heads) tends to be salient in the sense of Schelling (1960). Salience, famously, operates exogenously and is not captured by NE as a technical solution concept.

		Player 2	
		Heads	Tails
Player 1	Heads	1, 1	0, 0
	Tails	0, 0	1, 1

Table I

The second game Bacharach considers is the Prisoners' Dilemma (PD), as shown in strategic form in Table II. The PD's unique NE is (Defect, Defect), but the outcome when both players adopt their dominant strategies, (2, 2), is worse for both players than what they each would have obtained, (3, 3), had they adopted their dominated strategies instead.

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	3, 3	1, 4
	Defect	4, 1	2, 2

Table II

Bacharach argues that the one-shot PD presents a problem for applied game theory because in experiments many pairs of human players arrive at the Pareto superior outcome.

Bacharach argues that the third game he considers, the Hi-Lo game of Table III, provides the most representative frame for understanding the general class of equilibrium selection puzzles for which pure coordination and PD games furnish special cases. Hi-Lo has two pure strategy NE, (High, High) and (Low, Low), where the former Pareto dominates the latter.

		Player 2	
		High	Low
Player 1	High	2, 2	0, 0
	Low	0, 0	1, 1

Table III

Hi-Lo raises the same kind of equilibrium selection problem, according to Bacharach, as a pure coordination game because NE as a solution concept does not prescribe play of one pair of equilibrium strategies over the other. But the indeterminacy in Hi-Lo seems particularly troubling because in actual applications people have no problem at all in coordinating on the (High, High) equilibrium.

If we are willing to incorporate bounds on people's rationality then it is easy to explain the selection of (High, High): if each player assumes that the other assigns equal probability to both strategies then High is a mutual best response. But given the simplicity of the game this approach is unconvincing. If a style of unbounded reasoning that would prescribe the choice of High in this game can be identified, then Bacharach argues that it might also account for the solution principles apparently used

by many or most human players of pure coordination games and one-shot PDs. Bacharach's (2006) theory of *team reasoning* is intended to identify such a general solution.

In motivating his proposal, Bacharach also directs attention to non-toy examples such as the "offside trap" in football (soccer) where defenders simultaneously run forward so that the other team is caught offside when it tries to attack the goal. Each defender has two strategies: she can try to steal the ball from the attackers and block the goal, or she can rush forward. If all defenders choose the second strategy, they might catch the attackers offside. The first strategy, Bacharach argues, is akin to playing Low in Hi-Lo while the latter is equivalent to playing High, because when everyone adopts the offside trap defence, the likelihood of success is greater. Such play is routinely observed in experienced teams. Can game theory be used to play any role in explaining this achievement?

Team reasoning, according to Bacharach, provides the answer. Players using such reasoning find the strategy profile that yields the highest possible payoff for the team, and then the players adopt the strategies which, in combination, produce the profile.

To develop the idea, we begin with what Bacharch refers as a *simple coordination context*. Assume that there is a set T of n agents, with a set of feasible profiles of options O, and a shared ranking of these profiles as embodied in the payoff function U[4]. Thus, a simple coordination context is the triple (T, O, U). Bacharach argues that many non-toy situations have the properties of a simple coordination context, and directs attention to hypothesised causal processes or *choice mechanisms* which determine the actions of agents in these contexts. One such choice mechanism, which Bacharach calls *simple direction*, has the following features. Let o^* be the profile o that yields the highest value of U. Under simple direction, we assume that a $(n + 1)^{st}$ agent, the *director*, works out o^*, identifies the agents in control of the constituent components of o^*, tells each agent i to execute her component o_i^*, and the agents then perform the directed actions. If all the members of T are influenced by simple direction then o^* is implemented and U is maximised.

Team reasoning, Bacharach (2006, p. 123) argues, is "do-it-yourself direction." Agents in a simple coordination context team reason about choice problems as follows: each computes the optimal profile o^*; each identifies their component o_i^*; and each reasons that she should perform o_i^* because that is the component of the optimal profile over which she has control. Clearly if everyone in T team reasons then the optimal profile o^* is implemented and team welfare, as embodied in U, is maximised.

Team reasoning is thus a two-step process. The first step involves reasoning at the *group level* so as to identify the optimal profile o^*. The second step involves reasoning at the *individual level* so as to select and implement o_i^*, the individual's component of the optimal profile o^*. When the agents in T execute the profile o^*,

[4] Following Bacharach, we use U to refer to the shared or group payoff function and u_i to refer to an individual payoff function.

Bacharach refers to the mechanism by which they do so as a *team mechanism* and the members of *T* as a *team*.

If we apply the logic of team reasoning to Hi-Lo we see that the equilibrium selection problem dissolves. If the players of Hi-Lo team reason, they will identify (High, High) as the optimal profile (step one) and they will then each execute High, as this is each player's action from the optimal profile under her control (step two).

What should we expect when *T* includes team reasoners and non-team reasoners? And what happens when one cannot determine with certainty who is a team reasoner and who is not a team reasoner? These cases lead Bacharach to the notions of a *restricted coordination context* and *unreliable coordination context*, respectively.

A restricted coordination context occurs when there is common knowledge among the agents as to who team reasons and who does not team reason in a particular group. The latter agents are referred to as the *remainder* and they are assumed to adopt a fixed sub-profile *f* of actions, which is known to the team reasoners. The team reasoners in the group apply *restricted team reasoning*, the goal of which is to maximise *U* subject to the constraint that the non-team reasoners will adopt *f*. A restricted coordination context is arguably more realistic than a simple coordination context, but a generalisation of both of these interactions is an unreliable coordination context.

The scope of a restricted coordination context is limited by two assumptions. The first is that there is common knowledge concerning those agents who comprise the remainder. In most coordination contexts there is likely to be uncertainty about who team reasons and who does not. The second assumption is that the agents in the remainder adopt a fixed sub-profile *f*. It may be the case, however, that members of the remainder have strategic inclinations of their own which produces a noncooperative game between the team reasoners and the remainder. Bacharach refers to this case as the *strategic remainder* problem; it is discussed in detail in Bacharach (1999). We will focus on the limitations engendered by the first assumption, which Bacharach terms the *unknown remainder* problem.

Assume that membership of the remainder is determined by a random process. That is, let *M* be a mechanism governing team choice, and assume that every agent functions under *M* with probability ω, which is common knowledge among the group. Assume further that if agent *i* turns out to be in the remainder, she adopts option f_i, where f_i is referred to as her *default choice*. We can describe this *unreliable coordination context* by the collection (S, ω, O, U, f), where S represents the set of n agents, ω is the probability of functioning under M, O is the set of feasible profiles of options, U is the shared payoff function, and f is the profile of default options. T now represents the subset of agents from S that function under M, and $R = S - T$ denotes the remainder.

In an unreliable coordination context the crucial issue is how a team defines an optimal profile given the probability ω of functioning under *M*. The first-best profile o^* is unlikely to be attained because that is only possible when all agents function under *M*. The optimal profile in this context will be the one which maximises the *expected value* of *U* and thereby takes into account the probabilities of functioning

and failing under M. This particular profile is labelled o^{**} and it is to be understood as the profile that maximises the expected value of U given that each agent i will choose o_i with probability ω and f_i with probability $1 - \omega$.

In an unreliable coordination context, the interpretation of a *team mechanism* is one where the agents adopt o_i^{**} if they function under the mechanism, such that the mechanism delivers the profile o^{**}. A *team* is therefore defined as those agents in S who function under the mechanism M, which implies that T is a random set of agents. The definition of team reasoning in this context follows naturally from the definition given earlier: each agent i in T (i.e., an agent that functions under M) determines o^{**}, and then identifies and implements o_i^{**}. Bacharach refers to team reasoning in this unreliable coordination context as *circumspect team reasoning*. This mode of reasoning is efficient, in the sense of maximising the expected value of U, even when there is uncertainty about which agents will function and fail under the mechanism.

A reader who finds this framework for analysis useful is bound to wonder what conditions, in general, tend to generate team reasoning. Bacharach argues that *group identification* primes team reasoning and he refers to this as the *reasoning effect* of group identification.

This brings us squarely to the question of how we identify human game players with (economic) agents. When we attribute agency to an individual person, it is natural to think about the person's options and his or her ranking of alternatives. In this case, one asks questions of the form, "What should *he* or *she* do?" But when we attribute agency to a group of people then the focus can shift to the profiles which the group can enact, and to the group's ranking of the outcomes. Then the question of interest might change to, "What should *they* do?" Note that this latter question exemplifies the first step of a director's reasoning. Bacharach's core intuition is that as the focus shifts from the options that an individual can choose to the profiles which groups can implement, the answers to these should-do questions change from being indeterminate (as in the equilibrium selection problem) to determinate.

Suppose that instead of simply attributing agency to a group from the outside (i.e., the case of the director), members of the group come to self-identify with the team. In this case, the relevant question changes from, "What should *they* do?" to, "What should *we* do?" Bacharach argues that when people start to ask these questions they undergo a two-part transformation: they not only experience a *payoff transformation* (seeking to promote U rather than u_i) but also an *agency transformation* - each person thinks of herself as a component part of the team's agency. Then just as the director engages in the first step of director reasoning so the team member engages in the first step of team reasoning, identifying the profile that maximises U.

Bacharach argues that the likelihood of group identification is a function of a range of factors, some of which may be identifiable characteristics of a strategic interaction. One such characteristic is *strong interdependence*: each player realises that she will do well from framing her decision in terms of team agency only to the extent that she can be assured that her similarly motivated partner takes a particular action, and there is uncertainty as to whether the partner will take the action in question. In coordination games, PDs, and Hi-Lo games, solving for equilibria of the interactions

among players optimizing their individual preferences does not provide such assurance.

Bacharach refines this notion of strong interdependence to define the *interdependence hypothesis*. Consider the profiles S and S^*, where S is a solution to the game when players reason individualistically (e.g., (Defect, Defect) in a PD) and S^* is optimal for the group (e.g., (Cooperate, Cooperate) in a PD). Strong interdependence implies that, given S and S^*, the players have common interest in, and copower for, S^* over S, while recognising that S is a solution to the game that will obtain if the players reason individualistically. The interdependence hypothesis is that group identification is stimulated by perception of strong interdependence.

Bacharach then argues that the salience of strong interdependence, the lack of countervailing pressures to self-identify, and the degree of strong interdependence, are likely to affect the tendency to group identify, as implied by the interdependence hypothesis. Finally, he argues that when endogenous group identification occurs, the shared payoff function U will respect *unanimity*: if u_i and u_j, $i \neq j$, share the same ranking of profiles, then U will embody this ranking.

We are now in a position to state Bacharach's proposed resolution of the selection problem that seems to leave the game theorist unable to apply the Hi-Lo game analysis reasonably to applications with human players. In strategic interactions that fit the Hi-Lo specification, strong interdependence is highly salient and the payoff assignment indicates that there are no countervailing pressures to self-identify. As the players' preferences are in perfect alignment, u_i and u_j will share the same ranking of profiles and, by unanimity, the group payoff function U will embody this ranking. Consequently, the tendency for players to group identify will be strong, and, if this occurs, it sets in motion the process whereby players team reason, identify (High, High) as the best profile, and then implement the action that falls to them as part of the optimal profile, i.e., play High.

Bacharach argues that this theory has a far wider scope than coordination games. Specifically, strategic interactions, such as Stag Hunt, Battle of the Sexes and PD, which embody mixed motives, are likely to prime group identification and prompt team reasoning. However, these games differ in important ways from coordination games. In mixed motive games, there are countervailing pressures to self-identify, which, therefore, imply reduction in the salience of strong interdependence. Consequently, one expects team reasoning and attainment of the optimal profile for the team to be less prevalent in these interactions than in coordination games. People might waver between the gestalts of self-identification and team-identification.

In mixed motive games the link between the group payoff function U and the individual payoff functions u_i and u_j is more complex than in games where players' utilities are perfectly aligned. Bacharach argues that when endogenous group identification primes team reasoning the shared payoff function U will respect unanimity in u_i and u_j and symmetry between individual payoffs. That is, in the PD example below, we only need to specify u_F to account for the strategy profile when one player cooperates and the other defects, rather than use two variables to index the outcomes in which one player cooperates and one defects.

Consider Table IV below, which represents a generic PD, where C stands for Cooperate and D stands for Defect. The following inequalities must hold for the game to be dominance-solvable: $a > b > c > d$. In addition, for (C, C) to be Pareto optimal, $b > [(a + d) / 2]$. Now suppose that we want to find mechanisms which maximise U, where U is defined as a sum of agents' payoffs, and assume that anyone in the remainder plays D as her default choice. Assume further that the players interact in an unreliable coordination context and will engage, therefore, in circumspect team reasoning.

The first-best profile o^* is (C, C) and if $\omega = 1$, this profile will be enacted. But if $\omega < 1$ then matters are more complicated. Let $u_C = 2b$ represent the sum of individual payoffs when both players cooperate, $u_D = 2c$ represent the sum of individual payoffs if both players defect, and $u_F = a + d$ represent the sum of individual payoffs if one player cooperates while the other player defects; the subscript F refers to free-riding. Given the inequalities above, $u_C > u_D$ and $u_C > u_F$.

		Player 2	
		C	D
Player 1	C	b, b	d, a
	D	a, d	c, c

Table IV

Now consider the profile (C, C). With probability ω^2 both players will adopt the profile and u_C will result; with probability $2\omega(1 - \omega)$ one player will play C while the other plays D and u_F will result; and with probability $(1 - \omega)^2$ both players will play D and u_D will result. Thus, the expected value of U for the profile (C, C) is $EU(\text{C, C}) = \omega^2 u_C + 2\omega(1 - \omega)u_F + (1 - \omega)^2 u_D$.

Now consider the profile (D, D). As D is the default choice of both players, they will adopt the (D, D) profile with certainty and u_D will result. Finally, consider the two profiles (C, D) and (D, C). As one of the players is always playing D, the expected value of U from these profiles is: $\omega u_F + (1 - \omega)u_D$.

To determine o^{**}, the profile which maximises U in an unreliable coordination context, we must consider two cases. The first is where $u_F \geq u_D$. In this case, $EU(\text{C, C}) = \omega^2 u_C + 2\omega(1 - \omega)u_F + (1 - \omega)^2 u_D > \omega u_F + (1 - \omega)u_D = EU(\text{C, D}) = EU(\text{D, C})$ for all values of ω and (C, C) therefore defines the profile o^{**}.

The second case is where $u_D > u_F$. In this situation, the optimal profile o^{**} depends on the value of ω. To see this, normalise $u_C = 1$ and $u_D = 0$ and note that $u_F < 0$. Then, $EU(\text{C, C}) = \omega^2 + 2\omega(1 - \omega)u_F > 0 = EU(\text{D, D}) \Leftrightarrow \omega > [2u_F / (2u_F - 1)]$. The reverse holds if $\omega < [2u_F / (2u_F - 1)]$. In words, at high values of ω (that is, when the likelihood that agents function under M is high) the optimal profile o^{**} is (C, C) but when the value of ω is low, the optimal profile o^{**} is (D, D). Thus, the PD can be averted when the probability that agents team reason is relatively high.

In summary, Bacharach's theory structures models in which individuals may undergo payoff and agency transformations when strategic interactions are characterised by strong interdependence. Such interdependence prompts group identification and team

reasoning, which together entail identifying the optimal profile and then reasoning to the conclusion that each player should adopt her component of the optimal profile.

Bacharach's project reflects the assumption that to explain the outcome of an interaction by identifying it with the equilibrium of a game requires specifying a path of reasoning that would select the outcome in question. This leaves open the possibility that people sometimes reach the outcomes that team reasoners would by other means – say, emotional identification with symbols of fused agency. In such instances Bacharach's account encourages the judgment that game theory has nothing to contribute to the explanation. The agents in this kind of case don't decide that it is best to reason as a team, but simply do fuse their agency; if they participate in any processes usefully modelled as games, these will be interactions of their team with other agents (including, perhaps, other teams).

Following Coleman (1990), Ross (2014) argues for wider applicability of game theory. The mathematics of games is the basis for more than models of rational choice based on deliberation; it is also a technology for modelling social group formation and maintenance. Evolutionary game theory is one widespread approach to this project, but it abstracts from the context of choice altogether; individuals in evolutionary games simply express strategies selected by fitness competitions. If people can in fact switch between individually framed and team framed agency in the course of their strategic interactions, as Bacharach suggests and as observation supports, then it is natural to ask whether game theory can contribute anything to our understanding of this. If in fact it can, then a second question arises: might the strategic principles that govern team framing itself also help to explain the relative stability of the equilibria at which teams arrive?

The issue at hand transcends questions about the reach of game theory. Bacharach defends a deontological interpretation of team reasoning as a driver of behaviour. Once an individual has identified the optimal team profile and her component in it, he insists, she is rationally obliged to execute her component. Bacharach refers to this as the *projection feature* of profile-based reasoning, arguing that, "The underlying general principle is that I cannot coherently will something without willing what I know to be logically entailed by it" (2006, p. 136). It seems plausible that people sometimes reason in this way. However, we are sceptical of a claim to the effect that when people identify with teams and choose actions accordingly, they *typically* do so by means of reasoning or are much influenced by 'logical compulsion'. Game theory, like economics, is concerned with choices. If choice is *defined* in terms of outputs of reasoning processes, it follows that an account of team agency must be an account of reasoning. It might not necessarily be an account of actual deliberation in which people consciously engage, but rather an ex post rationalization of behaviour that serves as a 'stand-by' or 'back-up' to more common behaviour-generating processes, as per the account of Pettit (2001). However, in our view a general theory of an aspect of agency, particularly economic agency, should reflect the more deflationary account of choice that, as argued by Ross (2011), partly distinguishes economics from psychology, both methodologically and in terms of explanatory domains. According to this deflationary view, a behaviour is chosen just in case it is subject to influence by incentives, regardless of whether the causal channel that links incentives and behaviour involves deliberation. For example, if people spontaneously copy the behaviour of higher-status, kin-bonded, or apparently successful people without

thinking, this behaviour can still be regarded as chosen because counter-incentives could dampen it, even though by hypothesis it does not result from reasoning.

An empirical basis for doubting that team reasoning is the only, or even principal, basis for team agency among people is drawn from developmental psychology. In efforts to shed light on the evolutionary depth of human altruism, researchers have compared spontaneous prosocial behaviour in human and chimpanzee infants (Warneken and Tomasello 2006, 2007, 2009a, 2009b). Much of the focus has been on sharing, which does not necessarily implicate team agency. However, one of the primary alleged sites of difference between young humans and young chimpanzees has been based on observations of spontaneous assistance provided to adults who feign difficulty in completing tasks. The claim that young chimpanzees do not do this has been called into question (Horner *et al* 2011); but that humans as young as 14 months join the projects of others without direct inducement by contingent reward is well established. This is at least *prima facie* evidence that team agency in humans is a natural propensity, rather than behaviour that depends on deliberate reasoning. If young chimpanzees in fact show the same proclivity, at least under certain conditions, this would provide further grounds for seeking a more general theory.

Thus there are both theoretical and empirical motivations for seeking a more general game-theoretic account of team agency. A theory of team reasoning would then be a special application of this more general theory that could augment the relative stability of team solutions in groups of agents who are overwhelmingly motivated by rational deliberation. Wynn Stirling (2012)[5] has recently provided a formal theory that Ross (2014, Chapter 5) conjectured as filling just this role. In the next section, we first summarise Stirling's construction, and then confirm Ross's conjecture.

3. *Conditional game theory and social agency*

The avowed aim of Stirling (2012) is to develop a concept of group preference, which is not simply an exogenous aggregation of individual preferences, but which arises endogenously as social influences propagate through a group. Stirling's framework is a strict generalisation of orthodox, non-evolutionary game theory that incorporates the influence of social bonds through the technology of *conditional preferences*.

To illustrate the intuition we employ an example due to Ross (2014). Consider a Board of Directors that must decide whether to engage in a risky hostile takeover bid. There are at least two ways in which the views of the Board can be elicited. Under process (i), the Chair sends out a detailed risk analysis of the costs and benefits of the proposed takeover prior to the board meeting. Under process (ii), the Chair, citing security concerns, presents the same information to the Board but only after they have assembled in the boardroom. The question of interest is whether these two processes should be expected to yield the same outcome.

Process (i) encourages the Board members to form unconditional preferences prior to the meeting, which they might then defend against other members' arguments. Process (ii), by contrast, may induce members to monitor one another while they

[5] See also Stirling and Felin (2013).

decide which option is best and may lead them to modulate their preferences on the basis of the preferences of others. Under both processes, differing individual preferences are likely to be expressed through non-unanimous votes. But the distribution of these preferences might vary across the two scenarios because process (ii) encourages revelation of preferences that are influenced by information about the preferences of others, which thereby affords more opportunity for preference calibration.

The starting point of Stirling's analysis is the distinction between what he terms *categorical* and *conditional* preferences. Categorical preferences *unconditionally* define an agent's ranking of all possible outcomes, regardless of other agents' preferences, whereas conditional preferences are based on influence flows which propagate through a group and define agents' rankings of alternative outcomes as *conditional* on the preferences of others. This propagation of influence flows, which is modelled using graph theory, defines a social model that enables agents to jointly consider individual and group interests, as in Bacharach's framework, but without requiring us to leave the Nash constraint.

Building on the earlier example, but simplifying to the case of the Chair and one Board member, assume that each player has two actions[6]: support (S) the takeover bid or do not support (NS) the takeover bid. Thus, the outcome space for this game is: (S, S), (S, NS), (NS, S), (NS, NS), where the Chair's action is listed first and the Board member's action is listed second. Assume further that the Chair has categorical preferences over the action profiles but, as suggested earlier, the board member's preferences may be influenced by the Chair's. Specifically, suppose that if (S, S) is the Chair's optimal profile, the board member will define his ranking of the alternatives on the basis of this hypothesis. By contrast, if (NS, NS) is the Chair's optimal profile then the board member may define a different preference ordering. Given the four possible outcomes of this game, the Board member can define different preference orderings which are conditional on his conjecture concerning the preference ordering of the Chair.

Stirling's intuition is that as social influence propagates through a group and players modulate their preferences on the basis of other players' preferences, a complex notion of group preference may emerge. This notion may not directly provide the basis for action, but rather serve as a social model which incorporates all of the relationships and interdependencies that exist among the agents. Stirling refers to this concept as *concordance* and it captures the extent to which a conjectured[7] set of (categorical or conditional) preferences yield controversy within a group. Crucially,

---

[6] *Actions* properly refer to the alternatives available to a player at an information set in the extensive form of a game, whereas a strategy is a complete plan of action, specifying the move that a player will make at each information set where he or she may be called upon to act. Stirling confines his attention to finite strategic form games and employs the term 'action' interchangeably with 'strategy.' Despite some discomfort with this louche talk, we will follow his usage here.

[7] The notion of a conjecture is familiar from Bayesian games, where each player is assigned a distribution of expectations over the elements of other players' strategy sets. In Stirling's framework, this idea is generalized so that a conjecture is a belief about the strategy profile that will be instantiated by *all* players, including the player to whom the conjecture is assigned. As will become clear, it is the recursive nature of equilibrium determination in conditional game theory that allows for this.

concordance does not refer to the goals of a group nor to the goals of the individuals who comprise it, but rather to the level of discord that hypothetical propositions concerning players' preferences engender among members of the group.

For example, consider the following joint conjectures for the Chair and Board member: $\mathbf{a_1}$ = {(S, S), (S, NS)} and $\mathbf{a_2}$ = {(S, S), (NS, NS)}. Assume that under $\mathbf{a_1}$, the Chair's conjecture (S, S) is best for her and next-best for the Board member while the Board member's conjecture (S, NS) is best for him but next-best for the Chair. By contrast, assume that under $\mathbf{a_2}$, the Chair's conjecture (S, S) is, once again, best for her and next-best for the board member while the Board member's conjecture (NS, NS) is worst for both players. Which conjecture is likely to entail a greater level of controversy among the players? The joint conjecture $\mathbf{a_1}$ involves different conjectures by the two players but they do not include the players' worst outcome. The joint conjecture $\mathbf{a_2}$, by contrast, incorporates a conjecture (S, S) that might be satisfactory to either player but one (NS, NS) which is the worst for both players. Consequently, we might expect $\mathbf{a_2}$ to produce more severe dispute among the players than $\mathbf{a_1}$ and an ordering over these joint conjectures that is sensitive to these varying levels of controversy encodes the concept of concordance.

The level of concordance varies with the specific strategic interaction under study. In games where players' interests are perfectly aligned, the extent of controversy will be minimised when players conjecture identical action profiles. In zero-sum games, by contrast, a low degree of controversy is more likely when conjectures are diametrically opposed. In a penalty shootout in soccer, for example, success for the group (i.e., the two teams together) requires fierce competition and rivalry so if the goalkeeper were to favour a conjecture similar to the striker this would undermine competition and produce a high level of controversy. As Stirling (2012, p. 40) notes, "… even antagonists can behave concordantly."

While the concept of concordance may provide the basis for an emergent notion of group preference its value derives from the extent to which it is determined by the individuals who make up a group. In other words, concordance should not be imposed exogenously on a group from the outside but should instead be determined by the social linkages and influence flows among members of a group. Stirling refers to this principle as *endogeny*. It is among the building blocks of his *aggregation theorem*, which in turn provides a model of the social relationships and interdependencies of members of a group, and a device for simultaneously representing individual and group agency.

To develop a concordant ordering which respects the principles of conditioning (i.e., that players' preferences may be conditional on the preferences of others) and endogeny, Stirling employs the logic of multivariate probability theory in a *praxeological* context. He urges us to understand praxeology on the basis of an analogy with epistemology. Whereas epistemology is concerned with the nature and scope of knowledge and classifies propositions on the basis of their veracity, praxeology classifies propositions on the basis of their efficacy and efficiency.

In probability theory, given a set of two discrete random variables $\{X, Y\}$, the conditional probability mass function $p_{Y|X}(y \mid x) = P(Y = y \mid X = x)$ is a measure of the likelihood that the random variable $Y = y$ given that, or conditional on, the random

variable $X = x$. This conditional probability mass function is defined as the ratio of the joint probability of X and Y and the marginal probability of X or $p_{Y|X}(y \mid x) = p_{XY}(x, y) / p_X(x)$. Solving this expression for $p_{XY}(x, y)$ as the subject of the formula (i.e., $p_{XY}(x, y) = p_{Y|X}(y \mid x) \times p_X(x)$) it is clear that the joint probability of X and Y can be derived from the conditional probability of Y given X and the marginal probability of X. In other words, probability theory provides a framework for combining information from different sources – in this instance, the conditional probability of Y given X and the marginal probability of X – to determine the joint likelihood of an event.

In the praxeological framework, Stirling's goal is to derive a concordant ordering for the group which combines the conditional and categorical preferences of members of the group, in much the same way as the joint probability of an event is determined by conditional and marginal probabilities. Working directly with preference orderings quickly becomes cumbersome, so Stirling seeks to derive utility functions that represent the players' categorical and conditional preferences and the group's concordant preference ordering. The existence theorem for a utility function that represents categorical preferences is well known so we will focus on the derivation of a conditional utility function and the principles which must hold so as to permit aggregation of categorical and conditional preferences to derive a concordant utility function.

Let $\{X_1, \ldots, X_n\}$, $n \geq 2$, represent a set of n players, and let A_i denote a finite set of actions available to player i from which he or she must choose one element to instantiate. An action or strategy *profile* is an array $\mathbf{a} = (a_1, \ldots, a_n) \in A_1 \times \ldots \times A_n$. Under classical game theory, players have categorical utility or payoff functions defined over strategy profiles: $u_i : A_1 \times \ldots \times A_n \rightarrow \mathsf{R}$.

In the context of conditional preferences it is useful to define the parent set $pa(X_i) = \{X_{i1}, \ldots, X_{in}\}$ as the n_i-element subset of players whose preferences influence X_i's preferences. Assume that X_{ij}, the j^{th} parent of X_i, forms the hypothetical proposition that profile $\mathbf{a}_{ij}$ will occur. This hypothetical proposition is termed a *conjecture*. Thus, let $\mathbf{a}_i = \{\mathbf{a}_{i1}, \ldots, \mathbf{a}_{in}\}$ represent the *joint conjecture* of $pa(X_i)$. Then there exists a function which maps action profiles, *conditional* on the joint conjecture of $pa(X_i)$, to the real line R, which represents X_i's preferences: $u_{X_i \mid pa(X_i)}(\cdot \mid \mathbf{a}_i) : A_1 \times \ldots \times A_n \rightarrow \mathsf{R}$. Note that if $pa(X_i) = \varnothing$, then the conditional utility $u_{X_i \mid pa(X_i)}$ becomes the categorical utility u_i. Given the existence of a conditional utility function which represents players' conditional preferences, the collection $\{X_i, A_i, u_{X_i \mid pa(X_i)}, i = 1, \ldots, n\}$ constitutes a finite, normal form, noncooperative *conditional game*.

Returning to our example of the Chair (C) and the Board member (B), the conditional game consists of two players $\{X_C, X_B\}$, each with two actions $A_i = \{\text{S, NS}\}$, and the utility functions $u_C(\mathbf{a}_C)$ and $u_{B|C}(\mathbf{a}_B \mid \mathbf{a}_C)$, for the Chair and Board member, respectively.

Note that through appropriate normalisation one can ensure that all utilities (i.e., categorical and conditional) are non-negative and sum to unity, which implies that the utilities have all of the characteristics of probability mass functions. As discussed earlier, in an epistemological framework marginal and conditional probabilities can be combined to determine a joint probability: $p_{XY}(x, y) = p_{Y|X}(y \mid x) \times p_X(x)$. Consequently, if the praxeology-epistemology analogy is appropriate, it may be

possible to aggregate the conditional and categorical utilities to define a group utility function that incorporates the social linkages and interdependencies of members of a group and thereby represents the level of concordance of the group. The benefit of showing that this praxeology-epistemology analogy holds is that it will then be possible to apply concepts from multivariate probability theory, such as Bayes's rule and marginalisation, in a praxeological context and derive game-theoretic solution concepts that incorporate both individual and group interests.

Returning to our example, the goal is to combine the categorical preferences of the Chair with the conditional preferences of the Board member to produce an emergent preference ordering for the group. The requirement is to prove that the group or concordant utility $U_{CB}(\mathbf{a}_C, \mathbf{a}_B) = u_{B|C}(\mathbf{a}_B \mid \mathbf{a}_C) \times u_C(\mathbf{a}_C)$. In words, the concordant utility U is the product of the board member's conditional utility and the Chair's categorical utility.

In assembling the basis for such a proof, Stirling adopts three further assumptions or *principles*. The first is *acyclicity*, which means that no cycles can occur in the social influence relationships among players. In other words, if the Chair influences the Board member, then the Board member cannot influence the Chair. The problem with cyclical influence relationships is that they raise the possibility of indirect self-influence: the Chair influences the Board member, who in turn influences the Chair, which leads to a non-terminating cycle. Clearly this limits the generality of the model, and in so doing raises the stakes on its capacity to generalise the idea of team agency. As we shall see below, however, the restrictive power of acyclicity is countered elsewhere in the theory. An implication of acyclicity is that influence relationships are hierarchical and that at least one player in a strategic interaction must possess categorical preferences. Another implication is that social influence relationships can be represented using a directed acyclic graph (DAG).

The second principle is *exchangeability*, which Stirling and Felin (2013) refer to as framing invariance. This principle requires that if a strategic interaction can be framed in different ways but there is no loss of information under the different framings, then all framings must produce an identical concordant ordering. What this principle implies is that players must be willing to take into consideration the preferences of others when defining their own preferences, even if only to a small degree, and that the same information is available to the players under alternative framings.

In an epistemological context, framing invariance is a natural restriction because it implies that $p_{XY}(x, y) = p_{Y|X}(y \mid x) \times p_X(x) = p_Y(y) \times p_{X|Y}(x \mid y) = p_{YX}(y, x)$. For framing invariance to hold in a praxeological context, the concordant utility must satisfy the following conditions: $U_{CB}(\mathbf{a}_C, \mathbf{a}_B) = u_{B|C}(\mathbf{a}_B \mid \mathbf{a}_C) \times u_C(\mathbf{a}_C) = u_B(\mathbf{a}_B) \times u_{C|B}(\mathbf{a}_C \mid \mathbf{a}_B) = U_{BC}(\mathbf{a}_B, \mathbf{a}_C)$. In words, the concordant utility U_{CB}, which combines the conditional preferences of the Board member and categorical preferences of the Chair, must be the same as the concordant utility U_{BC}, which combines the categorical preferences of the Board member and the conditional preferences of the Chair. This principle mitigates the restrictive force of acyclicity with respect to the range of interactions we can use the theory to model.

The final principle required to derive a concordant utility function that has all of the characteristics of a joint probability mass function is *monotonicity*. This is a natural

restriction on the concordant utility function, which ensures that no individual's preferences will be arbitrarily subjugated by the group. Specifically, if an individual or subgroup prefers option A to B and the other players are indifferent among them, then the group must not prefer B to A. Thus, if the Chair prefers S to NS and the Board member is indifferent, the group must not prefer NS to S.

Stirling (2012, p. 59 – 60) proves that if the principles of *conditioning*, *endogeny*, *acyclicity*, *exchangeability* and *monotonicity* hold, then a concordant utility function exists that represents the social relationships of the group, and is derived from the conditional and categorical utility functions of its members. The most general form of the concordant utility function is:

$$U_{X1\ldots Xn}(\mathbf{a}_1, \ldots, \mathbf{a}_n) = \Pi^n_{i=1}\, u_{Xi \mid pa(Xi)}(\mathbf{a}_i \mid \mathbf{a}_i)$$

This expression shows that the concordant utility function, which combines information in a praxeological domain, shares exactly the same syntax as a joint probability mass function that combines information in an epistemological domain. Consequently, the full power of multivariate probability theory (particularly Bayes' rule and marginalisation) can be applied in a praxeological context to determine effective and efficient action when social influences propagate through a group.

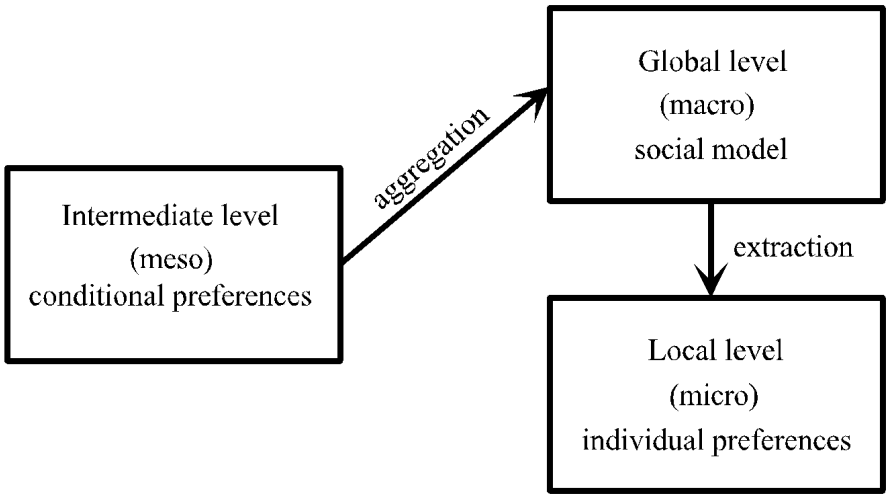
Marginalisation is an important operation in the praxeological domain because it allows the analyst to extract players' *ex post* preferences once social influence has permeated the group. A player's *ex post unconditional* preferences are extracted in the following manner:

$$u_{Xi}(\mathbf{a}_i) = \Sigma_{\sim \mathbf{a}i}\, U_{X1\ldots Xn}(\mathbf{a}_1, \ldots, \mathbf{a}_n),$$

where $\Sigma_{\sim \mathbf{a}i}$ means that the sum is taken over all arguments except $\mathbf{a}_i$. Note that these *ex post* categorical utilities represent the players' preferences after taking into account the social relationships and interdependencies that exist in the group. As the preferences are unconditional, standard solution concepts such as dominance and NE can be applied to them.

The preceding discussion is summarised in Figure I. As social influences propagate through a group, players define their conditional preferences. Through the process of aggregation these social linkages and interdependencies lead to an emergent notion of group preference: concordance. Finally, through the process of marginalisation, the analyst extracts the players' *ex post* categorical preferences.

Figure I: Conditioning, Aggregation and Extraction



Source: Stirling (2012, p. 19)

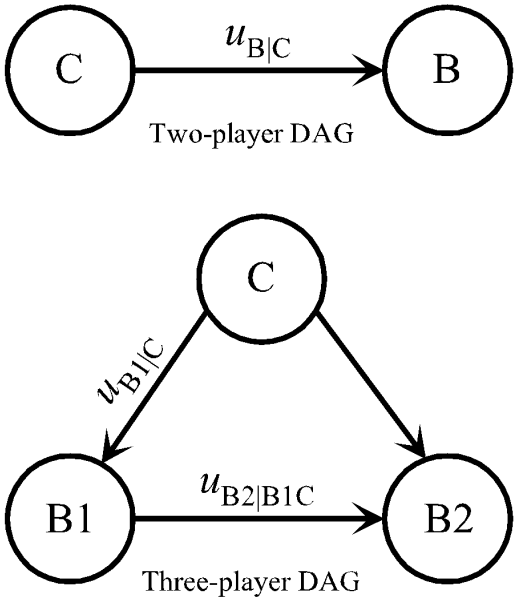
Acyclicity implies that social influence relationships in conditional games can be modelled using a DAG. A DAG is a *graph* made up of *vertices* or *nodes*, which in a praxeological context represents the players, and *directed edges* or *links*, which capture the influence relationships between the players. If one player, C, influences another player, B, we write C → B, where C is referred to as the *parent* of B and B as the *child* of C. The set of parents of B is denoted *pa*(B) and the set of children of B is denoted *ch*(B). If a vertex has no parents *pa*(C) = ∅ then it is called a root vertex. Figure II shows the DAG for the Chair and board member example and the case where there is a Chair and two Board members.

In the two-player DAG, the Chair influences the Board member but, given acyclicity, the board member does not influence the Chair. The Board member's conditional utility $u_{B|C}$ is represented by the edge between the nodes C and B. In the three-player DAG, the Chair influences Board member B1 and Board member B2, and Board member B1 influences Board member B2. The influence flow between C and B1 is captured by the conditional utility $u_{B1|C}$ and the influence flows between C and B1 toward B2 are captured by the conditional utility $u_{B2|B1C}$. The associated concordant utility for the three-player DAG is:

$$U_{CB1B2}(\mathbf{c}, \mathbf{b}_1, \mathbf{b}_2) = u_C(\mathbf{c}) \times u_{B1|C}(\mathbf{b}_1, \mathbf{c}) \times u_{B2|B1C}(\mathbf{b}_2 \mid \mathbf{c}, \mathbf{b}_1)$$

This expression combines information from the categorical and conditional utilities to define the concordant utility in much the same way that a Bayesian network, which can also be represented in a DAG, combines information from marginal and conditional probabilities to determine a joint probability. Thus, a DAG provides a graphical method to represent the influence flows, and associated conditional utilities, of a conditional game.

Figure II: Directed Acyclic Graphs



The three-player DAG in Figure II shows that B2 does not directly influence B1 and that neither B1 nor B2 directly influence C. However, this does not imply that B1 and B2 have no influence on C whatsoever. Recall that the exchangeability constraint means that a social model should be invariant to the way in which the information about linkages and influence flows is aggregated. This implies that once the concordant utility has been defined, we can apply Bayes's rule to extract reciprocal influence relationships. Specifically, suppose that B1 conjectures $\mathbf{b}_1$ and we want to determine the influence of this conjecture on the Chair's preference for $\mathbf{c}$: $u_{C|B1}(\mathbf{c} \mid \mathbf{b}_1)$. The answer follows directly from Bayes' rule:

$$u_{C|B1}(\mathbf{c} \mid \mathbf{b}_1) = [u_{B1|C}(\mathbf{b}_1 \mid \mathbf{c}) \times u_C(\mathbf{c})] / u_{B1}(\mathbf{b}_1),$$

where $u_{B1}(\mathbf{b}_1)$ is derived by marginalising the concordant utility.

We can also determine the influence that B1 and B2 exert on C and the influence that B2 exerts on B1 by computing the appropriate conditional and categorical utilities using Bayes's rule and marginalisation. The crucial idea here is that once the concordant utility has been defined, exchangeability implies that many hierarchical structures are compatible with the social model of the group. In other words, the social model is framing invariant.

Stirling then extends – as opposed to refines – the standard solution concepts of dominance and NE, to apply over group-level preference orderings. His approach is to extract a marginal utility for the group in much the same way as a marginal utility for each player was extracted from the concordant utility. A crucial assumption behind the procedure is that, given that players can only control their own actions, each player will make conjectures over her own action sets and not those of other players.

Thus, let a_{ij} denote the j^{th} element of $\mathbf{a}_i$, where $\mathbf{a}_i$ is X_i's conjecture profile. Now form the action profile $(a_{11}, \ldots, a_{nn})$ by taking the i^{th} element of each X_i's conjecture profile. Finally, sum the concordant utility over all elements of each $\mathbf{a}_i$ except a_{ii} to form the group utility or welfare function:

$$V_{X1\ldots Xn}(a_{11}, \ldots, a_{nn}) = \Sigma_{\sim a11} \ldots \Sigma_{\sim ann}\, U_{X1\ldots Xn}(\mathbf{a}_1, \ldots, \mathbf{a}_n)$$

As Stirling notes, the group does not act as a single entity and it cannot, therefore, instantiate its own preferred alternative, but the group utility provides a metric by which individual players determine the impact of their choices on the group. In much the same way as players can extract their marginal utilities from the concordant utility function, they can extract their own individual marginal welfare functions from the group utility. Specifically, the marginal individual welfare function v_{Xi} of X_i is the i^{th} marginal of $V_{X1\ldots Xn}$:

$$v_{Xi}(a_i) = \Sigma_{\sim ai}\, V_{X1\ldots Xn}(a_1, \ldots, a_n)$$

The existence of group and individual welfare functions allows Stirling to derive a solution concept that allows us to formally integrate consideration of the interests of the group with consideration of the interests of the individual players. This solution concept relies on the maximum individual and group welfare solutions.

The maximum group welfare solution is:

$$\mathbf{a}^* = \arg\max_{\mathbf{a} \in A1 \times \ldots \times An} V_{X1\ldots Xn}(\mathbf{a})$$

The maximum individual welfare solution is:

$$a_i^{\S} = \arg\max_{ai \in Ai} v_{Xi}(a_i)$$

If $a_i^{\S} = a_i^*$ for all $i \in \{1, \ldots, n\}$, the action profile is a *consensus* choice, meaning that group and individual welfare is maximised when **a** is instantiated. As Stirling notes, a consensus choice will often not exist, in which case players might be motivated to enter into negotation to reach compromise. In a noncooperative game setting, the outcomes of such negotiations would need to be protected by commitment devices. This would signal a failure of team agency to form, though repeated interaction with the resulting new institutions might ultimately incentivise players to identify with them, and thereby create conditions for team agency later. For present purposes, however, it suffices to show that conditional game theory generalises team agency in cases where consensus choice applies, because Bacharach's unanimity condition is a special case of it.

To show this, we begin with the PD. As Bacharach recognises, one cannot obtain cooperation in a PD – in his framework, the conditions for team reasoning are not present – if no player cares about the welfare of the group at all. Thus, as established by Binmore (1994), if the preference structure of the PD describes *all* of the relevant preference information pertinent to the interaction, then general defection is the only outcome that a game theoretic model of it can predict. Binmore further insists that if the model does *not* incorporate all such information in the specification of preferences, then the game should not be characterised as a PD in the first place. However, admitting the mere *possibility* of team agency allows us to admit that more than one game structure might be relevant to modeling an empirical interaction. This situation is hardly unprecedented in economics. We are used to the idea, for example, that a plurality of models are useful for foregrounding different aspects of

international trade, oligopoly, national production, and other phenomena. Stirling (2012, p. 80) draws a distinction between simple reciprocal altruism that transforms PDs into coordination games, and background representations of interaction structures in which players' models of their own and others' preferences are consistent with the PD structure but they are also aware of preferences they *would* have *conditional* on the implementation of some degree of socially mediated agency. This is indeed the basis on which Stirling's general framework is given its name.

The machinery by which Stirling represents genuine PD structure simultaneously with scope for team agency representation are *cooperation and exploitation indices*. Specifically, Stirling endows each player X_i with a cooperation index $\alpha_i \in [0, 1]$ and an exploitation index $\beta_i \in [0, 1]$, where α_i represents the extent to which a player is conditionally willing to cooperate, and β_i represents the extent to which a player is conditionally willing to exploit his or her partner. Because these tolerances are conditional on the same model transformation, we impose a minimal consistency requirement by assuming that $\alpha + \beta < 1$. To respect acyclicity, assume further that X_1 has categorical preferences and that X_2's conditional preferences are conditional on X_1's.

Given the cooperation and exploitation indices, X_1's categorical utility is defined as follows:

$$u_{X1}(C, C) = \alpha_1 \qquad u_{X1}(C, D) = 0$$
$$u_{X1}(D, C) = \beta_1 \qquad u_{X1}(D, D) = 1 - \alpha_1 - \beta_1$$

In the PD representation of the interaction, $\beta_1 > \alpha_1 > 1 - \alpha_1 - \beta_1 > 0$, and X_2 has a categorical utility function such that $u_{X2}(C, D) > u_{X2}(C, C) > u_{X2}(D, D) > u_{X2}(D, C)$.

For the conditional representation, we calculate $u_{X2|X1}(a_{21}, a_{22} \mid a_{11}, a_{12})$ by computing utilities for every possible conjecture that player X_1 can make. Assume that if X_1 conjectures either (C, C) or (D, D) then X_2 will place all of her conditional utility mass on the same action profile. In other words, if X_1 conjectures cooperation then X_2 finds it optimal to cooperate but if X_1 conjectures defection then X_2 finds it optimal to defect. If X_1 conjectures (C, D), then X_2's utility mass will be apportioned according to her cooperation and exploitation indices. Specifically, X_2 will assign α_2 to (C, C), β_2 to (C, D), $1 - \alpha_2 - \beta_2$ to (D, D) and zero utility mass to (D, C) because this is the worst possible outcome for X_2. Finally, if X_1 conjectures (D, C), the worst outcome for X_2, X_2 should place zero utility mass on (D, C), α_2 and (C, C), β_2 on (C, D) and $1 - \alpha_2 - \beta_2$ on (D, D). The conditional utilities associated with each conjecture of X_1, represented in the columns, and every action profile which can be instantiated by the two players, represented in the rows, are given in Table V.

	(a_{11}, a_{12})			
(a_{21}, a_{22})	(C, C)	(C, D)	(D, C)	(D, D)
(C, C)	1	α_2	α_2	0
(C, D)	0	β_2	β_2	0
(D, C)	0	0	0	0
(D, D)	0	$1 - \alpha_2 - \beta_2$	$1 - \alpha_2 - \beta_2$	1

Table V

To compute the concordant utility we combine X_1's categorical utility with X_2's conditional utility: $U_{X1X2}(\mathbf{a}_1, \mathbf{a}_2) = u_{X2|X1}(a_{21}, a_{22} \mid a_{11}, a_{12}) \times u_{X1}(a_{11}, a_{12})$. The result is shown in Table VI where the rows index X_1's conjecture and the columns index X_2's conjecture.

	(a_{21}, a_{22})			
(a_{11}, a_{12})	(C, C)	(C, D)	(D, C)	(D, D)
(C, C)	α_1	0	0	0
(C, D)	0	0	0	0
(D, C)	$\alpha_2\beta_1$	$\beta_1\beta_2$	0	$\beta_1 - \alpha_2\,\beta_1 - \beta_1\beta_2$
(D, D)	0	0	0	$1 - \alpha_1 - \beta_1$

Table VI

The concordant utility can now be used to extract the *ex post* marginal utilities, the group welfare function, and the individual welfare function. X_1's *ex post* utilities are equivalent to her categorical utilities, whereas X_2's *ex post* utilities must be derived through marginalisation: $u_{X2}(\mathbf{a}_2) = \Sigma_{\sim \mathbf{a}2}\, U_{X1X2}(\mathbf{a}_1, \mathbf{a}_2)$. For example, $u_{X2}(\text{C, C}) = \alpha_1 + 0 + \alpha_2\beta_1 + 0 = \alpha_1 + \alpha_2\beta_1$. The *ex post* payoff matrix for the PD is shown in Table VII.

		X_2	
		C	D
X_1	C	$\alpha_1, \alpha_1 + \alpha_2\beta_1$	$0, \beta_1\beta_2$
	D	$\beta_1, 0$	$1 - \alpha_1 - \beta_1,\ 1 - \alpha_1 - \alpha_2\beta_1 - \beta_1\beta_2$

Table VII

The group welfare function for this two-player game is derived using the following expression: $V_{X1X2}(a_{11}, a_{22}) = \Sigma_{\sim a11}\, \Sigma_{\sim a22}\, U_{X1X2}(\mathbf{a}_1, \mathbf{a}_2)$. For example, $V_{X1X2}(\text{D, D}) = \beta_1\beta_2 + 0 + \beta_1 - \alpha_2\,\beta_1 - \beta_1\beta_2 + 1 - \alpha_1 - \beta_1 = 1 - \alpha_1 - \alpha_2\beta_1$. The full group welfare function is:

$$V_{X1X2}(\text{C, C}) = \alpha_1$$
$$V_{X1X2}(\text{C, D}) = 0$$
$$V_{X1X2}(\text{D, C}) = \alpha_2\beta_1$$
$$V_{X1X2}(\text{D, D}) = 1 - \alpha_1 - \alpha_2\beta_1$$

Finally, the individual welfare functions are extracted from the group welfare function using marginalisation: $v_{Xi}(a_i) = \Sigma_{\sim ai}\, V_{X1X2}(a_1, a_2)$. For example, $v_{X2}(\text{C}) = \alpha_1 + \alpha_2\beta_1$. Thus, the individual welfare functions are:

$$v_{X1}(\text{C}) = \alpha_1 \qquad v_{X1}(\text{D}) = 1 - \alpha_1$$
$$v_{X2}(\text{C}) = \alpha_1 + \alpha_2\beta_1 \qquad v_{X2}(\text{D}) = 1 - \alpha_1 - \alpha_2\beta_1$$

To find the NE of this game after incorporating the social influence flows between X_1 and X_2, we work directly with the conditional and categorical utilities (Stirling refers to the equilibria identified using this method as *conditioned NE*) or the *ex post* marginal utilities (Stirling refers to the equilibria identified using this method as *ex post NE*). The two approaches yield identical solutions. Table VII shows that (D, D) is a NE for all admissible values of α_i and β_i. Unlike the unconditional PD, (C, C) is a NE when $\alpha_i > \beta_i$. Furthermore, when $\alpha_i > \beta_i$, (C, C) is a consensus choice because it maximises both group and individual welfare. In an unconditional representation of the play that will in fact be observed, this would be reflected in altered payoff

rankings, making the empirically correct unconditional game an Assurance Game rather than a PD.

It is intuitive that if both players prefer cooperation to exploitation then (C, C) will be a conditioned or *ex post* NE but this result fails to highlight the role that social influences can play in this game. To see this, assume that $\alpha_1 = 0.6$ and $\beta_1 = 0.3$ and that $\alpha_2 = 0.3$ and $\beta_2 = 0.6$. Thus, X_1's cooperation index is twice as large as her exploitation index but X_2's cooperation index is half as large as her exploitation index. So, in the absence of influence flows, X_1 is a cooperator and X_2 is an exploiter. But after X_2 takes into account X_1's preferences, X_2's penchant for exploitation is tempered by X_1's desire for cooperation and (C, C) is a conditioned NE.

While explaining cooperation in an empirical interaction that might be mis-predicted if we attend only to its unconditional model as a one-shot PD is an important accomplishment, we must keep in mind Bacharach's argument that the litmus test for an effort to represent team agency is that it furnish an explanation for High play in Hi-Lo. We now show that conditional game theory passes this test. In the discussion below, H stands for High and L stands for Low.

To allow social influences to affect the analysis of Hi-Lo, we endow each player X_i with a *High play index* $\alpha_i \in [0, 1]$ and a *Low play index* $\beta_i \in [0, 1]$, where $\alpha_i + \beta_i = 1$, because the players will assign zero utility mass to mis-matches (i.e., (H, L) and (L, H)). Assume again that X_1's preferences are categorical and that X_2's conditional preferences are conditional on X_1's.

Given the High play and Low play indices, X_1's categorical utility is defined as follows:

$$u_{X1}(\text{H, H}) = \alpha_1 \qquad u_{X1}(\text{H, L}) = 0$$
$$u_{X1}(\text{L, H}) = 0 \qquad u_{X1}(\text{L, L}) = \beta_1$$

To calculate $u_{X2|X1}(a_{21}, a_{22} \mid a_{11}, a_{12})$ it is necessary to compute utilities for every possible conjecture of player X_1. Assume that if X_1 conjectures either (H, H) or (L, L) then X_2 will place all of her conditional utility mass on the same action profile. That is, if X_1 conjectures High then X_2 finds it optimal to play High but if X_1 conjectures Low then X_2 finds it optimal to play Low. If X_1 conjectures (H, L) or (L, H), then X_2's utility mass will be apportioned according to her High play and Low play indices. Specifically, X_2 will assign α_2 to (H, H) and β_2 to (L, L), and zero utility mass to (H, L) and (L, H) because these are the worst outcomes for X_2. The conditional utilities associated with each conjecture of X_1, represented in the columns, and every action profile which can be instantiated by the two players, represented in the rows, are given in Table VIII.

	(a_{11}, a_{12})			
(a_{21}, a_{22})	(H, H)	(H, L)	(L, H)	(L, L)
(H, H)	1	α_2	α_2	0
(H, L)	0	0	0	0
(L, H)	0	0	0	0
(L, L)	0	β_2	β_2	1

Table VIII

To compute the concordant utility we combine X_1's categorical utility with X_2's conditional utility: $U_{X1X2}(\mathbf{a}_1, \mathbf{a}_2) = u_{X2|X1}(a_{21}, a_{22} \mid a_{11}, a_{12}) \times u_{X1}(a_{11}, a_{12})$. The result is shown in Table IX where the rows index X_1's conjecture and the columns index X_2's conjecture.

	(a_{21}, a_{22})			
(a_{11}, a_{12})	(H, H)	(H, L)	(L, H)	(L, L)
(H, H)	α_1	0	0	0
(H, L)	0	0	0	0
(L, H)	0	0	0	0
(L, L)	0	0	0	β_1

Table IX

The concordant utility can now be used to extract the *ex post* marginal utilities, the group welfare function, and the individual welfare function. As X_1's preferences remain categorical, her *ex post* utilities are her categorical utilities whereas X_2's *ex post* utilities must be derived through marginalisation: $u_{X2}(\mathbf{a}_2) = \Sigma_{\sim \mathbf{a}2} U_{X1X2}(\mathbf{a}_1, \mathbf{a}_2)$. For example, $u_{X2}(\text{H, H}) = \alpha_1 + 0 + 0 + 0 = \alpha_1$. The *ex post* payoff matrix for Hi-Lo is shown in Table X.

		X_2	
		H	L
X_1	H	α_1, α_1	0, 0
	L	0, 0	β_1, β_1

Table X

The group welfare function for this two-player game is derived using the following expression: $V_{X1X2}(a_{11}, a_{22}) = \Sigma_{\sim a11} \Sigma_{\sim a22} U_{X1X2}(\mathbf{a}_1, \mathbf{a}_2)$. For example, $V_{X1X2}(\text{H, H}) = \alpha_1 + 0 + 0 + 0 = \alpha_1$. The full group welfare function is:

$$V_{X1X2}(\text{H, H}) = \alpha_1$$
$$V_{X1X2}(\text{H, L}) = 0$$
$$V_{X1X2}(\text{L, H}) = 0$$
$$V_{X1X2}(\text{L, L}) = \beta_1$$

Finally, the individual welfare functions are extracted from the group welfare function using marginalisation: $v_{Xi}(a_i) = \Sigma_{\sim ai} V_{X1X2}(a_1, a_2)$. For example, $v_{X2}(\text{H}) = \alpha_1$. Thus, the individual welfare functions are:

$$v_{X1}(\text{H}) = \alpha_1 \qquad v_{X1}(\text{L}) = \beta_1$$
$$v_{X2}(\text{H}) = \alpha_1 \qquad v_{X2}(\text{L}) = \beta_1$$

To find the NE after incorporating the social influence flows between X_1 and X_2, we work directly with the conditional and categorical utilities to identify the conditioned NE, or with the *ex post* marginal utilities to identify the *ex post* NE. As desired, the two approaches yield identical solutions. Table X shows that (H, H) and (L, L) are NE for all admissible values of α_1 and β_1.

When one focuses on the group and individual welfare functions we see that group and individual welfare is maximised through the profile (H, H) when $\alpha_1 > \beta_1$. As this is the assumption in Hi-Lo, the profile that caters for the interests of the individuals and the group is (H, H) and this is a consensus choice. Consequently, we would expect this profile to be instantiated when players take into account their own individual interests and the interests of the group, as encoded in the social linkages among the players and expressed through the group welfare function.

4. *Conclusion*

Conditional game theory has full power to represent team agency using only resources that can be defined within standard game-theoretic formalism, and which can be represented using only standard solution concepts. It does not presuppose that players explicitly reason their way to solutions based on identification with teams, but it captures conditionalisation of games by that mechanism, among others.

A conditional game-theoretic specification is also compatible with the hypothesis that people experience the sorts of gestalt switches between individual and team agency that Bacharach conjectures. Psychologists can contribute to our unified understanding of social behaviour by investigating the frequency of such switches, in both directions, in different sorts of circumstances, along with general kinds of conditions that encourage or interfere with them. It might be the case that, in most interactions, people either simply assume group-level agency and stick to it, or play their unconditioned best responses without reflection. (These tendencies might likely be both statistical and context dependent). It might even be typically *best* – because of the importance of stability of strategic expectations – if gestalt switches are relatively unusual.

The strategic life of a social being is complicated, and one of the leading sources of this complication is multiple scales of agency. Game theory is up to the job of representing this multiplicity. The philosopher's task of assessing it through its many normative angles and shadows is much less likely to find straightforward resolution, but can benefit from the existence of a general technical framework in which to describe its structure.

REFERENCES

BACHARACH, M. (1999): "Interactive Team Reasoning: A Contribution to the Theory of Cooperation," Research in Economics, 53.

BACHARACH, M. (2006): Beyond Individual Choice: Teams and Frames in Game Theory. Princeton, NJ: Princeton University Press.

BINMORE, K. (1994): Game Theory and the Social Contract, Volume 1: Playing Fair. Cambridge, MA: MIT Press.

BINMORE, K. (2009): Rational Decisions. Princeton, NJ: Princeton University Press.

COLEMAN, J. (1990): Foundations of Social Theory. Cambridge, MA: Harvard University Press.

HOLLIS, M. (1998): Trust within Reason. Cambridge: Cambridge University Press.

HORNER, V., J. D. CARTER, M. SUCHAK, AND F. B. DE WAAL (2011): "Spontaneous Prosocial Choice by Chimpanzees," Proceedings of the National Academy of Sciences, 108, 13847-51.

KREPS, D. (1990): Game Theory and Economic Modeling. Oxford: Oxford University Press.

MEHTA, J., C. STARMER, AND R. SUGDEN (1994): "The Nature of Salience: An Experimental Investigation of Pure Coordination Games," American Economic Review, 84, 658-673.

PETTIT, P. (2001): "The Virtual Reality of Homo Economicus," in The Economic World View, ed. by U. Mäki. Cambridge: Cambridge University Press, 75-97.

RAGOT, X. (2012): "The Economics of the Laboratory Mouse: Where Do We Go from Here?," in What's Right with Macroeconomics?, ed. by R. Solow, and J.-P. Touffut. Cheltenham: Edward Elgar, 181-194.

ROSS, D. (2014). Philosophy of Economics. Houndmills, Basingstoke: Palgrave Macmillan.

SCHELLING, T. C. (1960): The Strategy of Conflict. Cambridge, MA: Harvard University Press.

STIRLING, W. C. (2012): Theory of Conditional Games. New York, NY: Cambridge University Press.

STIRLING, W. C., AND T. FELIN (2013): "Game Theory, Conditional Preferences, and Social Influence," PLoS ONE, 8, e56751. doi: 10.1371/journal.pone.0056751.

SUGDEN, R. (1993): "Thinking as a Team: Towards an Explanation of Nonselfish 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.

WARNEKEN, F., AND M. TOMASELLO (2006): "Altruistic Helping in Human Infants and Young Chimpanzees," Science, 311, 1301-3.

WARNEKEN, F., AND M. TOMASELLO (2007): "Helping and Cooperation at 14 Months of Age," Infancy, 11, 271-294.

WARNEKEN, F., AND M. TOMASELLO (2009a): "The Roots of Human Altruism," British Journal of Psychology, 100, 455-71.

WARNEKEN, F., AND M. TOMASELLO (2009b): "Varieties of Altruism in Children and Chimpanzees," Trends in Cognitive Sciences, 13, 397-402.