Team Reasoning from an Evolutionary Perspective: Categorization and Fitness

Ade, Leyla; Roy, Oliver

Team Reasoning from an Evolutionary Perspective: Categorization and Fitness

Leyla Ade and Olivier Roy *

Abstract: *»Team Reasoning aus evolutionärer Perspektive: Kategorisierung und Fitness«.* The question of the evolutionary stability of team reasoning has been answered in multiple, even opposing ways. We provide a general, conceptual categorization of these existing answers along four dimensions: (1) the unit of selection, (2) the notion of fitness for team reasoners, (3) the stage of decision-making, and (4) the ludic ecology. Beyond affording a better assessment of the different modeling choices underlying the existing results, the categorization highlights important conceptual questions for the evolutionary foundations of team reasoning. We illustrate this by looking in more detail into what should count as fitness for team reasoners.

Keywords: Team reasoning, evolutionary game theory, fitness, cooperation, group agency.

1. Introduction and Motivation

Team reasoning (Sugden 1993, 2003; Bacharach 1999, 2006) is an extension of classical game theory that aims at explaining the selection of the Pareto-optimal equilibria in coordination games like the Stag Hunt or the HiLo game, as well as the cooperative outcome in social dilemmas and common good problems.[1]

* Leyla Ade, Department of Philosophy, University of Bayreuth, Bayreuth, Germany; leyla.ade@uni-bayreuth.de. ORCID iD: 0000-0002-2708-6399.
Olivier Roy, Department of Philosophy, University of Bayreuth, Bayreuth, Germany; olivier.roy@uni-bayreuth.de. ORCID iD: 0000-0002-9085-5701.
Acknowledgements: Extensive comments and discussions with Maximilian Noichl and Johannes Marx have greatly contributed to developing the ideas presented in this paper. We also thank the participants at the Kickoff Meeting of the SCoLA project in February 2022 and the reviewers and participants at the LOFT 2022 conference in July 2022 for helpful inputs on an earlier draft of this paper. Finally, financial support of the German Research Foundation (DFG, project MA 4716/4-1) is gratefully acknowledged.

1 In what follows we only sketch the motivation for the development of the theory. See Colman and Gold (2018) and Gold (2017) for general presentations and discussion.

Table 1 Three Games of Interest: The Prisoners' Dilemma (PD), the Stag Hunt (SH), and the Hi-Lo game (HiLo)

PD	C	D
C	(3,3)	(-3,4)
D	(4,-3)	(1,1)

SH	S	R
S	(2,2)	(-1,1)
R	(1,-1)	(1,1)

HiL	H	L
H	(3,3)	(0,0)
L	(0,0)	(1,1)

The main motivation for developing the theory of team reasoning is the observation that classical equilibrium concepts fail to account for these phenomena in a *unified* matter. To see this, let us consider first the Prisoners' Dilemma (Figure 1, top left). Despite the importance of this game, e.g., the classical work of Ostrom (1990), the fact that mutual defection is the only Nash equilibrium of that one-shot, simultaneous interaction strikes many as counter-intuitive.[2] Several solutions for explaining cooperative behavior have been proposed, for instance by redescribing the game as a repeated one (Axelrod and Hamilton 1981) or by introducing assertive matching in evolutionary dynamics (Alexander 2007; Alger and Weibull 2013).

Unlike the Prisoners' Dilemma, the Stag Hunt game (Figure 1, top right) has two Nash equilibria in pure strategies: (*S,S*), the payoff-dominant equilibrium, and (*H,H*), the risk-dominant equilibrium. Authors like Skyrms (2014, 2004) have argued at length for the importance of this game for the theory of social conventions, and in particular for seeing the choice of the payoff-dominant equilibrium as a proxy for understanding the emergence of moral behavior. Yet, viewed as a one-shot interaction, the players face an equilibrium selection problem in the Stag Hunt, and so the concept of a Nash equilibrium alone cannot explain why players would or should choose *S* instead of *H*. This problem has been addressed successfully, again, by redescribing it in evolutionary terms where convergence to the payoff-dominant equilibrium occurs under much weaker conditions than, for instance, convergence to the cooperative outcome in the Prisoners' Dilemma, e.g., again Skyrms (2014, 2004).

Finally, the HiLo game (Figure 1, bottom) is a coordination game with a Pareto-dominant Nash equilibrium. The game has attracted attention in the early literature on the equilibrium selection problem (Schelling 1960; Harsanyi and Selten 1988) and is one of the main motivations for developing the theory of team reasoning. Indeed, Bacharach (2006) points out that there is

[2] See, e.g., (Kuhn 2019) for an overview of the discussion.

solid empirical evidence that players easily solve the equilibrium selection problem in HiLo by overwhelmingly choosing (*H*,*H*), and he argues that there are widespread and stable intuitions that this profile is the only rational outcome of the game. Canonical explanations of these empirical observations and normative intuitions, however, again go by a redescription of the equilibrium selection problem in evolutionary terms or by introducing additional choice principles like salience or payoff dominance (Schelling 1960; Harsanyi and Selten 1988).

Two points thus stand out regarding these solution to the equilibrium selection problems in the HiLo and the Stag Hunt game, and of the cooperative outcome in the Prisoner's Dilemma. First is their heterogeneity. Conditions like assortative matching, which appear central to an evolutionary explanation of cooperation in the Prisoners' Dilemma, are notably unnecessary for explaining the coordination on (*H*,*H*) in Hi-Lo. Similarly, supplementing the decision rules with principles like payoff dominance is of no help in explaining cooperation in the Prisoners' Dilemma, since mutual defection is the only Nash equilibrium of that game. The second point is that most of these explanations employ what Bacharach (2006) calls a strategy of redescription. By addressing the equilibrium selection problems in evolutionary terms, for instance, one moves away from the one-shot interpretation of the HiLo and the Stag Hunt game. The empirical observations and the normative intuitions mentioned above seem, however, to hold there as well: players overwhelmingly choose the Pareto-dominant equilibrium in HiLo even in experiments where care is taken to emphasize the one-shot character of the interaction, and our intuitions that this outcome is the only rational one do not seem to depend on repetition or long-term convergence.

The theory of team reasoning has been developed as an attempt to address these two points. It aims at providing a unified explanation of coordination and cooperation, valid in a broad "ludic ecology" (Bacharach 2006), i.e., across many games, be they one-shot or repeated. It does so by postulating that players can reason not only as individuals but also as team members. In a nutshell, team reasoners solve these games by asking "What should *we* do?" instead of the classical "What should *I* do?" This switch from the individual to the team perspective is fleshed out in terms of a transformation of *preferences* and of the *unit of agency* for team players.

Preference transformation is a change of preferences over the possible outcomes of the game. Assuming that a meaningful concept of team or group preference is given, team reasoners are taken to adopt or internalize these group preferences as their own and rank the possible outcomes of the game accordingly. Various proposals have been made for meaningful notions of group preferences, e.g., in terms of mutual advantage (Sugden 2015; Karpus and Radzvilas 2018; Duijf 2021). Many presentations of the theory, however, use simple aggregations of individual preferences, for instance taking the sum (Lempert 2018) or the average of individual payoffs (Bacharach 2006). Table 2 illustrates this for the Prisoner's Dilemma. In that case, preference transformation turns the social dilemma faced by individual reasoners into a HiLo game for team reasoners, i.e., a coordination game with (*C*,*C*) as the Pareto-dominant equilibrium.

Table 2 Prisoner's Dilemma Augmented with Team Preferences, the Third Entry in Each Cell, Here Taken as the Average Payoff of Both Players

	C	D
C	(3, 3, 3)	(-3, 4, 0,5)
D	(4, -3, 0,5)	(1, 1, 1)

Agency transformation bears on the objects over which team reasoners deliberate. Classically, individual reasoners deliberate over, and choose, individual strategies. Team reasoners take instead the perspective of an external team planner and compare different *combinations* of strategies, one for each team member. In the case where the team consists of the grand coalition of all players, this boils down to so-called profile-based, instead of the classical strategy-based, reasoning (Bacharach 2006). Each team member selects or identifies a profile that maximizes the team's preferences, i.e., asks herself what the team should do. The team members then individually do "profile projection," i.e., inferring their part in the team's action. Coming back to the Prisoner's Dilemma (Table 2), team reasoners would identify (*C*,*C*) as the best outcome for the team. Each team reasoner would then choose her action to realize this profile, i.e., each would individually choose *C*.[3] The same holds, *mutatis mutandis* for HiLo, where team reasoners would coordinate on (*H,H*), and the Stag Hunt, where both players would choose S.

Arguably, preference and agency transformations are substantial departures from classical game theory,[4] which immediately raises the question of their empirical plausibility. This question has been answered in two broad ways. On the one hand, a relatively large number of studies have been focused on the psychological basis of team reasoning, and in particular at pinpointing the relative importance of group identification in comparison with known decision heuristics like focal point reasoning or Strong Stackelberg reasoning (Bardsley et al. 2010; Colman, Pulford, and Lawrence 2014; Pulford, Colman, and Lawrence 2014; Faillo, Smerilli, and Sugden 2016).

This paper focuses on the second strand of answers to this question, which has been given comparatively less attention, namely the evolutionary or long-term stability of team reasoning. Roughly put, the question here is whether team reasoners could emerge, survive, or even thrive under diverse

[3] This holds for so-called basic team reasoning, where there is no uncertainty about the type of others. See Section 2.3.

[4] C.f. Hakli, Miller, and Tuomela (2010); Duijf (2021); but also Radzvilas and Karpus (2021) for a more compatibilist view.

circumstances, in long-term interactions with both like-minded and individual reasoners. Already in 1999, Bacharach claimed that "team reasoning [...] would have been adaptive in a world of interactions with [...] a modest frequency of team reasoners" (1999, 143). Since then, however, a more nuanced picture has emerged. In the case of the Prisoner's Dilemma, for instance, Amadae and Lempert (2015) report that team reasoners would be driven to extinction by individual reasoners.[5] Newton (2017) on the other hand identifies conditions under which team reasoners might be evolutionarily stable even in the Prisoner's Dilemma. Work in agent-based models also point in that direction (Angus and Newton 2015; Elsenbroich and Payette 2020). These apparently contradictory results rest on different modeling choices, choices that deserve to be highlighted more prominently.

This paper makes a twofold contribution. First, in Section 2 we extend the work of Lempert (2018) to provide a general, conceptual categorization of the modeling choices that we see as most relevant for the evolutionary analysis of team reasoning. The categorization rests on four dimensions: (1) the unit of selection, (2) the notion of fitness for team reasoners, (3) the uncertainty of types, and (4) the ludic ecology. Our second contribution, in Section 3, is to discuss in more detail the question of what should count as fitness for team reasoners. We present, in particular, arguments for the non-orthodox view that team preferences can be meaningfully understood as fitness for team reasoners, without invoking group selection.

The paper makes a conceptual contribution. The basic mathematical building blocks of the theory are thus presented only briefly in the Appendix. We, furthermore, mostly focus on Bacharach's version of the theory of team reasoning (1999, 2006). The main alternative is the theory developed by Sugden (1993, 2003), which we discuss briefly in Section 3 and the Conclusion.

2. Categorization

A distinguishing feature of Bacharach's version of the theory of team reasoning is that the switch between individual and team perspectives is not intentional. The players do not decide whether to team reason or not. It is something that happens to them, similarly to the assignment of types by "nature" in Bayesian games (Harsanyi 1967).

Bacharach's theory also takes into account the fact that not all players might be team reasoners, or that some might not be reliably expected to team reason. In the most general case, so-called *circumspect* team reasoning, each player team reasons with a certain, common prior probability, and this

[5] As Kantian equilibrium can be seen as a special case of team reasoning (Istrate 2021), Roemer (2019) also provides indirect evidence for that fact.

probability is common knowledge. Bacharach has proposed to solve such games using the notion of an *unreliable team equilibrium*, in which individual and team reasoners play a mutual best response, taking into account the given probabilities that the players will be of either type (see Appendix A.1). The cases in which all or only proper subsets of the players team reason, but where the identity of team reasoners is common knowledge, have been respectively labeled "basic" and "restricted" team reasoning.

The framework of circumspect team reasoning is sufficiently formalized to be congenial to evolutionary analysis, but not all aspects of the latter address conceptually important components of the former. Although consequential for the results, modeling choices regarding the nature of time (discrete or continuous), the specific evolutionary dynamics, or the assortativity of matching do not directly bear on important aspects of team reasoning. For that reason, they will be bracketed here. For simplicity of exposition, we will assume basic versions of the replicator dynamics (c.f. again the Appendix A.2).

Four aspects of evolutionary modeling appear most relevant to address the question "Is team reasoning evolutionarily stable?": (1) unit of selection, (2) fitness, (3) uncertainty of types, and (4) ludic ecology. The last two have been already identified explicitly by Lempert (2018), who also points toward the first, although less explicitly so. Our contribution here is to discuss their philosophical importance for the theory of team reasoning in more detail. The second category, i.e., what counts as fitness for a team reasoner, has only been fleetingly addressed in the literature. Here we briefly explain why, and in the next section we show that it raises important issues on its own. Lempert also argues for the importance of a fifth category: the mechanism of inheritance. We come back to this point in the conclusion.

2.1 Unit of Selection

The first dimension is the unit of selection, i.e., the object or property for which fitness is calculated and compared, and which is transmitted from one generation to the next (Lewontin 1970). In the case of team reasoning, one can make two broad choices: the strategies or the reasoning mode itself.[6]

Using strategies as the unit of selection can constitute a useful shortcut, but this modeling choice is of limited scope for team reasoning. For basic team reasoning in the Prisoner's Dilemma (Amadae and Lempert 2015), for instance, team reasoners can be distinguished behaviorally from individual reasoners. In this case, it is meaningful to use strategies as the unit of selection. This furthermore explains Amadae and Lempert's findings that

[6] These correspond respectively to what Bacharach (2006, chap. 3) calls trait and mechanism selection. That terminology is useful but not used consistently in the literature, so we leave it aside here.

cooperators, i.e., team reasoners, would be driven to extinction in the Prisoner's Dilemma. Note, however, that this approach only studies team reasoning indirectly, as one possible but not unique mechanism underlying the choice of certain strategies. As such, it falls under the extensive literature devoted to the evolution of cooperation in social dilemmas or equilibrium selection in coordination games, e.g., Axelrod and Hamilton (1981) and Weibull (1997). Furthermore, beyond the basic case, team reasoners generally cease to be behaviorally distinguishable from individual reasoners, making strategies less well-suited to studying the evolution of team reasoning.

Most existing studies thus take the mode of reasoning, i.e., individual or team reasoning, as the unit of selection (Bacharach 2006; Amadae and Lempert 2015; Newton 2017). Since this involves both preference and agency transformation, one can further subdivide the results in this category according to whether they rather focus on the first or the second.

The results of Lecouteux (2015), Newton (2017), and Rusch (2019) can be broadly understood as focusing on the evolution of agency transformation. Newton (2017), for instance, explicitly models profile-based reasoning where there is no uncertainty regarding the players' types. This, in turn, allows for defining conditional cooperation strategies,[7] through which team reasoning can be shown to be evolutionarily stable even in an environment consisting only of Prisoner's Dilemmas. Rusch (2019) extends this analysis to include uncertainty about the types of others and shows that as long as the cooperative solution has a high benefit, team reasoning can still be adaptive. Similar results can also be found in Lecouteux (2015), although the generality of his approach in terms of decision heuristics can be argued to cover not only agency but also preference transformation.

Although, to the best of our knowledge, this has not been done explicitly, indirect evolutionary approaches (Gueth and Kliemt 1998) provide fertile ground for studying the evolution of preference transformation. Indeed, the question of the evolution of altruistic preferences has been central in that literature, e.g., Alger and Weibull (2013). To the extent that preference transformation might lead to altruistic behavior, the results obtained there could help understand whether *both* preference and agency transformations are necessary to explain cooperation and coordination in a unified manner. The question has indeed been recently debated, with classical arguments to that effect, e.g., in Bacharach (2006), being restated and expanded (Colman 2022) in reaction to the claim that preference transformation might be sufficient for team reasoning (Duijf 2021). Comparing the potential stability of the preference transformation that Duijf suggests with results that instead focus on agency transformation would fruitfully complement this debate.

---

[7] As such, his analysis is arguably close to Sugden's (2000) theory of team reasoning.

2.2 Fitness for Team Reasoners

The second dimension along which it is fruitful to categorize existing work is the notion of fitness for team reasoners. We present it here only briefly and come back to this question in Section 3. Assuming that individual reasoners are standard game-theoretic agents, transferred to a classical, i.e., not indirect (see the previous section) evolutionary setting, it is natural to take the utility values representing their preferences as fitness. For team reasoners, however, two options are available.

The first option is to take fitness for team reasoners to be their utilities as individual reasoners. Notably, *all* the existing results surveyed here take that route. This modeling decision is rarely addressed explicitly, but a number of arguments can be given in its favor. First, both strategy choices and reasoning modes, the two units of selection that we considered above, are properties of individuals. This naturally suggests taking individual preferences as fitness even for team reasoners.[8] Second, doing so might put pressure against the selection of team reasoners, for instance, if they cooperate in social dilemmas. If team reasoning can nonetheless be shown to be evolutionarily stable in such contexts, this provides a stronger argument for its long-term viability (Amadae and Lempert 2015). Third, taking individual preferences as fitness for team reasoners does not preclude that team preferences play a role at a different level of selection (McElreath and Boyd 2008). Bacharach's (2006) suggestion that group selection can explain the evolutionary stability of team reasoning could be read along those lines.

Despite the so far universal adoption of the first option, a second one is possible: taking the utilities representing the team preferences directly as fitness for team reasoners. We will, in fact, argue below that this is a meaningful option, roughly because it does justice to the idea that each team reasoner *individually* undergoes preference and agency transformation.

2.3 Uncertainty of Types

Circumspect team reasoning introduces uncertainty about the players' types (individual or team reasoners), which suggests considering the three stages of decision making familiar to Bayesian games or epistemic game theory (Pacuit and Roy 2017): *ex ante*, where none of the types are known; *ex interim*, where the players know only their own types but not the others'; and *ex post*, where everyone's type is known.

Bacharach's version of circumspect team reasoning combines the *ex ante* and the *ex interim* perspective.[9] On the one hand, the team's best response is computed from the perspective of an external planner, who takes into

---

[8] We thank Andrew Colman for stressing this point to us.

[9] We thank Robert Sugden for stressing this point to us.

account the fact that *any* player only team reasons with a certain probability. In other words, the team's best response is computed *ex ante*. When doing profile projection, however, team reasoners know their types, i.e., they operate *ex interim*. Amadae and Lempert's (2015) analysis of the evolutionary stability of circumspect team reasoning falls in that category.

The second avenue, followed by Rusch (2019), is to consider that each player computes the team's best response *ex interim*. This route is conceptually important since it coincides with the reasoning of what has been called benefactors (Bacharach 2006) or prosocial I-mode reasoners (Hakli, Miller, and Tuomela 2010; Paternotte 2018), which is otherwise known to lead to different equilibria rather than *ex ante* circumspect team reasoning (Bacharach 1999).

Lastly, agents could be assumed to make their decisions *ex post*, knowing not only their own types but also the types of others. Lecouteux (2015) and Newton (2017) arguably take that perspective. This coincides with restricted team reasoning and allows the players to condition their choices on the type of their opponents. This, in turn, opens the door to study the evolution of Sugden's (2003) theory of team reasoning, which crucially rests on mutual assurance between the team members. This is an important point that distinguishes Sugden's theory from Bacharach's. In the latter, no mutual assurance is required, to the point that individual players can even team reason unilaterally.

2.4 Ludic Ecology

The ludic ecology describes the set of games that are being played. As already mentioned in Section 2.1, one of the main motivations behind Bacharach's theory of team reasoning is to provide a unified explanation of cooperation and coordination across many games. He claims that "if there is a relatively simple mechanism [...] that supports several traits, this may confer an evolutionary advantage [...] over other less versatile mechanisms" (Bacharach 2006, chap. 3).

This dimension has been explored quite broadly in existing studies. Amadae and Lempert (2015) consider an ecology consisting of the Prisoner's Dilemma and HiLo. This minimal diversity turns out to be crucial to their result: team reasoners can compensate in HiLo for the negative evolutionary pressure they face in the Prisoner's Dilemma. Rusch (2019) studies 12 symmetric 2x2 games that cover all possible strict ordinal preferences on the profiles, showing that a decision rule supporting collaboration can be viable under some conditions on the ludic ecology or as he calls it the *niche*. Rusch builds his work on Newton (2017), who introduces a general framework for the evolutionary analysis of collaboration. Newton identifies conditions on games under which collaborations are stable. He mentions some games which raise the problem of cooperation or coordination and fulfill these conditions, such as the threshold public good game, the multiplayer Prisoner's Dilemma, or the trust game.

One could, of course, consider even more diverse ecologies, all the way to Bacharach's (2006, chap. 3) "life game." Evolutionary analyses of choice mechanisms over large sets of games generated at random exist, e.g., Galeazzi and Franke (2017), and could naturally be used as the basis for a general approach to studying team reasoning in high ludic diversity.

Considering such potentially very high degrees of ludic diversity, however, raises the question of the intended scope of the theory of team reasoning itself. It has been explicitly developed with a focus on social dilemmas and coordination problems with Pareto- or payoff-dominant outcomes. As such, the fact that team reasoning cannot solve, for instance, games like Bach or Stravinsky[10] does not appear as a shortcoming. To the extent that this intuition is correct, the current literature has covered much of the relevant ground regarding ludic ecology.

3. Illustration: Team Preferences as Individual Fitness

As an illustration of the conceptual fruitfulness of the categorization that we just presented, we now go into more detail regarding what should be taken as fitness for team reasoners in evolutionary models. In section 2.3 we mentioned that all studies reviewed here take the fitness of team reasoners to be their individual utilities, and we sketched a number of arguments supporting this modeling decision. Here we want to explore a number of arguments in favor of the second option, i.e., taking the utilities representing the team preferences as fitness for team reasoners.

At the outset, one should emphasize again that models of group selection (Wilson 1975; Sober and Wilson 1998; Okasha 2006) could allow holding the view that fitness is individual preferences while keeping some explanatory role for team preferences. This is in fact suggested by Bacharach (2006, chap. 3). These models show conditions under which inter-group comparisons could give evolutionary advantages to certain traits, even though they might be individually disadvantageous at the intra-group level. Group selection is, however, not uncontroversial as an evolutionary model (Smith 1964; Okasha 2006), which motivates at least exploring the possibility of taking team preference as fitness for team reasoners while staying within the standard, individualist interpretations of fitness and the replicator dynamics.

[10] Taking team preference as average in this game, for example, turns Bach or Stravinsky into a pure coordination game, which team reasoning cannot solve either.

Whether the utilities representing the team preferences can be meaningfully interpreted as individual fitness depends more generally on the interpretation of the main component of the evolutionary models. Here we focus on biological and cultural interpretations, broadly understood. According to the first, fitness is interpreted either directly as the number of offspring, or more generally as reproductive success. The replicator dynamics, on that interpretation, capture the idea that variations in frequencies of particular traits in a population reflect the relative reproductive successes endowed upon agents with these traits (Weibull 1997).

There are numerous interpretations and extensions of standard evolutionary models in terms of cultural evolution, but as an illustration here we follow the presentation of Gruene-Yanoff (2011), which focuses on the so-called imitation (Weibull 1997, 152-61) and experimentation interpretations (Bjoernerstedt and Weibull 1996). Both go back to the more standard interpretation of preferences as represented by subjective utilities. In the imitation interpretation, players sample the payoffs of players of other types and switch to that type with a probability corresponding to a normalized difference with their own payoffs, if the former is higher. In the experimentation interpretation, players who get lower payoffs than the average for their type review and change their trait at a given rate, which is at least monotone in the difference between their own and the average payoffs of the other types.

Under the biological interpretation, taking the utilities representing the team preferences as fitness for team reasoners would boil down to having a population consisting of two phenotypes – individual and team reasoners – with their respective payoffs interpreted as their respective reproductive successes. In other words, team reasoners simply constitute a different type *of individuals* in the population. The fact that their reproductive successes differ from those of individual reasoners is merely a feature of their phenotype. It does not have to rest on any assumption about sharing mechanisms between team members.

This interpretation captures two key aspects of Bacharach's version of the theory of team reasoning. First, it does justice to the idea that preference and agency transformation are individual transformations, i.e., that individuals internalize the team's preference as their own. The biological interpretation gives a perhaps extreme reading of this internalization, as a hard-coded phenotype that determines reproductive success. Second, in the biological interpretation, the agents do not decide whether they team reason or not. This is also in line with Bacharach's theory. Care should be taken here, however, since the theory explicitly assumes that the propensity to team reason might vary for individual players depending on specific features of the game.[11]

[11] See Bacharach's remarks on the so-called "interdependence hypothesis" (Bacharach 2006, chap. 2).

This interpretation has two main conceptual downsides. On the one hand, it leaves out an explicit notion of "switch" between individual and team reasoning. By reducing team preferences to just another phenotype in the population, this interpretation also lacks an explanation of what is special about the fact that these preferences are the team's, and why this might yield different reproductive successes. In other words, it appears to bracket the "team" in "team preference." Note, however, that Bacharach's theory is not explicit on that point either. Team preferences are a given of the theory. Furthermore, team reasoning involves not only preference but also agency transformation. So, one could still meaningfully see the players as team reasoners even though the team preferences are abstractly given.

Let us now turn to cultural interpretations. Both the imitation and experimentation interpretations explicitly incorporate the idea that the players are switching between reasoning modes. They indeed move away from the view that the population is either entirely or in part reset at each round. Individuals stay from one round to the next, but as in the evolutionary models of Bayesian games (Ely and Sandholm 2005), the players can be seen as switching between types, here corresponding to individual and team preferences.

Under a cultural interpretation of the evolutionary process, taking the utilities representing team preferences as fitness for team reasoners arguably reintroduces intentionality in switching between types. In the imitation interpretation, although the actual switch is probabilistic, the sampling and the payoff comparisons are actions of the players. The same holds for the experimentation interpretation, where the reviewing and change are naturally seen as intentional processes. As such, they depart from the orthodoxy of Bacharach's version of team reasoning which, recall, views the switch as something that happens to the players, not something they do.

Both cultural interpretations that we consider here raise the question of interpersonal comparisons of utility (Gruene-Yanoff 2011), which in the present case boils down to asking whether individual and team utilities can be meaningfully compared. It is well known that in some contexts, interpersonal comparisons of utilities can be legitimate (Fleurbaey and Hammond 2004). For the sake of the argument, however, we can assume that the comparison between individual and team utilities falls outside this set of known cases. Bacharach's theory is explicitly presented as an extension of classical game theory, in which individual preferences are assumed to be represented by von Neumann-Morgenstern (vNM) utilities. The models of cultural evolution that we are considering now inherit this interpretation of individual utilities. Whether the team preferences can also be seen as being represented by vNM utilities is a question that goes beyond the scope of this paper, but again for the sake of the argument, we can assume that they are.

Even under these assumptions, one might argue that comparing individual and team utilities is legitimate because it is not completely interpersonal.

First, one should emphasize that theories of team reasoning assuming players to actively decide to team reason, e.g., Sugden's (1993), *de facto* suppose that individual and team utilities are comparable. More generally, however, taking seriously the idea that a player undergoes the preference transformation means that both the individual and the team's preferences are personal. Under the classical interpretation of vNM utilities, these represent two possible hypothetical choice behaviors of the same player.[12] If that player can switch between the two perspectives, comparing them appears less problematic than in the classical, interpersonal case. If, however, one insists that the switch is completely unintentional and constitutes a transition between two fully "compartmentalized" selves of a player, then the classical worries about interpersonal comparisons still apply.

The plausibility of taking the utilities representing team preferences as fitness for team reasoners thus rests on the idea that they internalize the team preferences as their own, and that these two perspectives are transparent to one another for individual players. The biological interpretation entails a strong commitment to the first idea. The team preference represents, on that interpretation, the personal reproductive successes of a particular phenotype, viz. the team reasoners. However, this is not to say that team reasoners are defined exclusively by their preferences under the biological interpretation. As we argued above, agency transformation can still play a key role in the evolutionary process, as well as uncertainty about the types of the others. The two cultural interpretations that we considered also assume that team reasoners endorse the team's preference as their own. Still, they raise the question whether, while team reasoning, a player can still contemplate how she would think, and especially evaluate the outcomes, where she is an individual reasoner (and vice-versa for individual reasoners). If the players are capable of doing so, this might alleviate traditional worries regarding the need to compare individual and team utilities. However, these worries remain if these two perspectives are viewed as completely separate states of mind for the players. Whether this is so is ultimately an empirical question, reaching out from the evolutionary to the psychological foundation of team reasoning.

4. Conclusion

The categorization that we proposed in this paper expands on the one proposed by Lempert (2018), but here we have bracketed the inheritance or transmission mechanism, which he argues is also fundamental. His

[12] Colman's (2022) argument, to the effect that payoff transformation is not sufficient for capturing team reasoning, rests in part on the assumption that individual preferences are assumed to be represented by vNM utilities. This might shed critical light on the present interpretation.

argument rests on the observation that, at least in the cultural interpretations of the evolutionary models (Section 3), team reasoning is a psychological process that is not simply copied from parent to offspring. As different inheritance mechanisms are otherwise known to yield qualitatively different results, they should be taken explicitly into account in studying the evolutionary foundation of team reasoning. We fully agree with this view. It was left out of the present discussion because we focused on dimensions that directly bear on specific aspects of team reasoning. The inheritance mechanism is important for studying the evolution of team reasoning, not because of the specifics of that theory but as a model of the evolution of psychological processes.

All in all, the general categorization we propose in this paper allows not only to situate existing contributions according to four fundamental modeling decisions, but also to point towards less explored but still promising paths towards a better understanding of the evolutionary stability of team reasoning. Revisiting existing results by taking the utilities representing the team preferences as fitness for team reasoners is the one most prominently emerging from that discussion. Others include extending the ludic ecology in Amadae and Lempert's (2015) model, looking at the evolution of agency transformation (Newton 2017; Rusch 2019) when the team's best response is calculated *ex ante*, and using indirect evolutionary approaches to study the evolution of preference transformation and comparing it to the results surveyed here. As we illustrated through the question of fitness for team reasoners, exploring these avenues is not only important from an evolutionary perspective. It touches on fundamental aspects of the theory as a whole, both in its philosophical and even psychological underpinnings.

Appendix

A.1 Circumspect Team Reasoning

An *unreliable team interaction* (UTI) includes teams that are taken to be subsets of the set of agents $\{1,\ldots,n\}$. The special case of a singleton, such as $\{i\}$, are considered to represent individuals, and hence often written simply as i (Bacharach 1999).

Definition 1 *(Bacharach 1999) An unreliable team interaction (UTI)*[13] *is a tuple* *(M,A,U,T,Ω) with*

- M = $\{M^1, \ldots, M^k\}$, *where for l∈{1,…,k},* $M^1 \subseteq \{1, \ldots, n\}$ *is a team*
- $A = A_1 \times \cdots \times A_n$*, where* A_i *is the set of possible actions for agent i∈{1,…,n}*
- $U = \{U^1, \ldots, U^k\}$*, where* $U^1: A \rightarrow R$ *is the utility function of team* $M^1 \in M$
- $T = T_1 \times \cdots \times T_n$*, where* $T_i \subseteq \{1, \ldots, k\}$ *is the set of participation states of agent i*
- $\Omega: T \rightarrow [0,1]$ *is a probability measure on T*

A profile in a UTI consists of an action for each individual player and each proper team. For example, if there are two players and one team, a profile would take the form (a_1, a_2, a_3) for a_1 and a_2 being the individual action of player $M^1 = \{1\}$ and $M^2 = \{2\}$, respectively, and a_3 the action of the team[14] M3 = {1, 2}. The expected utility of a team *l* (denoted by u_1) for a certain profile $(a_1, \ldots, a_k)$ is the sum over all possible participation states, given their probability and respective utility for this team.

$$u_1(a_1, \ldots, a_k) = \sum_{t=(t_1,\ldots,t_n)\in T} \Omega(t) U^1(a_{t_1}, \ldots, a_{t_n})$$

Whereas in general an agent could have multiple teams she potentially is a member of, in a two-player version it comes down to each player being either an individual or team reasoner. Let us again consider the Prisoner's Dilemma in Table 2 as an example. Assume the probability of team reasoning to be ω, hence the probability of an agent to reason individually is 1−ω. The expected utilities for the two reasoning modes (denoted by $u_{team}, u_{individual},$) given the profile (D,D,CC) are calculated as follows.

$$u_{team}(D,D,CC) = 3\omega^2 + 0.5\omega(1-\omega) + 0.5(1-\omega)\omega + 1$$
$$u_{individual}(D,D,CC) = 3\omega^2 + 4\omega(1-\omega) - 3(1-\omega)\omega + 1$$

Bacharach calls a profile a UTI equilibria in case any team maximizes their expected utility. He could show that in the Prisoner's Dilemma team reasoners will cooperate as long as the probability of team reasoning is at least $\frac{1}{3}$. If the probability is lower ($\omega < \frac{1}{3}$), the team action (*DD*) is a better response to the individuals' best responses (*D,D*), hence (*D,D,DD*) is the unique UTI equilibria in this case.

The UTI is different to a Bayesian game. In order to present team reasoning in a Bayesian game, one could take agents to have individual or team

13 Note that we are considering a special case of the original UTI. The original version includes outside signals from a set S_i for each agent *i*, which could contain any information about her own or other agent's membership in a team. However, we simplify to what Bacharach calls a *blind UTI*, in which case $S_i = \emptyset$ for all *i* ∈{1,…,*n*}.

14 Note that what we call *team actions* coincides with the profiles if the team consists of all agents. In order to differentiate the profiles and team actions, the latter will be written as (CC), hence without comma, whereas a profile is denoted by (C,C).

preferences depending on their own state, choosing an action for each state, respectively. The expected utility of a given profile would then depend on the probability of the other agents to be in a certain state. For example, in the Prisoner's Dilemma, the profile in which both agents play C if a team reasoner and D if individually reasoning, an agent who finds herself in the team reasoning state would have the expected utility of $3\omega+0.5(1-\omega)$. Bacharach has proven the UTI equilibria to be a proper subset of Bayesian equilibria (Bacharach 1999, Theorem 2).

A.2 Notions of Stability in Evolutionary Game Theory

Weibull (1997) gives an introduction to evolutionary game theory. The following section presents an overview of the most important notions, namely the replicator dynamics and evolutionary stable strategy (ESS).

Take $x = (x_1, \dots, x_m)$ to be a population state, s.t. $i \in \{1,\dots,m\}$ is a pure strategy, and $x_i \in [0,1]$ is the fraction of agents in the population playing pure strategy i. The state of agents only playing strategy i is denoted by e^i. The expected utility for a (mixed) strategy x played against another (mixed) strategy y is denoted by $u(x,y)$. In the replicator dynamics the fraction of agents playing a certain pure strategy gets adapted according to their expected utilities compared to the average expected utility in the population state. Formally this is defined as in Definition 2, where $\dot{x}_i$ denotes the new share of agents playing i.

Definition 2 (Replicator Dynamics)

$$\dot{x}_i = \left(u\left(e^i, x\right) - u(x,x)\right) x_i$$

"In other words, the growth rate $\frac{\dot{x}_i}{x_i}$ of the population share using strategy I equals the difference between the strategy's current payoff (fitness) and the current average payoff (fitness) in the population" (Weibull 1997, 73). A state is stationary under the replicator dynamics precisely if and only if all present strategies earn exactly the same payoff. If this is the case, then the according factor in the replicator dynamics is zero and no changes happen from the current to the next state.

A more advanced notion of stability is the evolutionary stability of strategies. A strategy is called evolutionarily stable if and only if it is immune to any sufficiently small mutant strategy invading a population state. Following Weibull's (1997) notation and taking the set of all mixed strategies to be Δ, ESS are defined as in Definition 3.

Definition 3 (Evolutionary Stable Strategy (ESS)) A strategy $x \in \Delta$ is evolutionary stable, if the inequality

$$u(x, \varepsilon y+(1-\varepsilon)x) > u(y, \varepsilon y+(1-\varepsilon)x)$$

holds for all mutant states y≠x and all small fractions of invasions $\varepsilon \leq \varepsilon_y$ *for some* $\varepsilon_y \epsilon (0,1)$.

An ESS can be characterized as a symmetric Nash equilibria in which the current strategy is doing better as the mutant. Formally this is, the strategy *s* is an ESS if and only if

1. (s,s) is a NE, i.e., $u(s,s) \geq u(s',s)$ for all s'
2. if $u(s,s) = u(s',s)$, then $u(s,s') > u(s',s')$ for all $s' \neq s$

A weaker notion of stability is that of *Neutrally Stable Strategies* (NSS) which describe that mutants might occur but cannot invade the population. The difference to the ESS defined in Definition 3 is that the inequality is only required to be weak.

Definition 4 (Neutrally Stable Strategy (NSS)) A strategy x∈Δ is neutrally stable, if the inequality

$$u(x,\varepsilon y+(1-\varepsilon)x) \geq u(y,\varepsilon y+(1-\varepsilon)x)$$

holds for all mutant states y≠x and all small fractions of invasions $\varepsilon \leq \varepsilon_y$ *for some* $\varepsilon_y \epsilon (0,1)$.

Note that any evolutionary stable strategy is also neutrally stable, and any neutrally stable strategy is also a symmetric Nash equilibria and as such stable under the replicator dynamics. For more details on these or further notions of stability, we refer the reader to Weibull (1997).

References

Alexander, J. McKenzie. 2007. *The structural evolution of morality*. New York: Cambridge University Press.

Alger, Ingela, and Joergen W. Weibull. 2013. Homo Moralis – Preference Evolution Under Incomplete Information and Assortative Matching. *Econometrica* 81 (6): 2269-302.

Amadae, S. M., and Daniel Lempert. 2015. The long-term viability of team reasoning. *Journal of Economic Methodology* 22 (4): 462-78.

Angus, Simon D., and Jonathan Newton. 2015. Emergence of Shared Intentionality is Coupled to the Advance of Cumulative Culture. *PLoS Computational Biology* 11 (10): e1004587.

Axelrod, Robert, and William D. Hamilton. 1981. The Evolution of Cooperation. *Science* 211 (4489): 1390-6.

Bacharach, Michael. 1999. Interactive team reasoning: A contribution to the theory of co-operation. *Research in Economics* 53 (2): 117-47.

Bacharach, Michael. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton and Oxford: Princeton University Press.

Bardsley, Nicholas, Judith Mehta, Chris Starmer, and Robert Sugden. 2010. Explaining Focal Points: Cognitive Hierarchy Theory versus Team Reasoning. *The Economic Journal* 120 (543): 40-79.

Bjoernerstedt, Jonas, and Joergen W. Weibull. 1996. Nash equilibrium and evolution by imitation. In *The rational foundations of economic behavior*, ed.

Kenneth Joseph Arrow, Enrico Colombatto, and Christian Schmidt. Proceedings of the IEA conference held in Turin, Italy, 155-81. New York: St. Martins Press.
Colman, Andrew. M. 2022. Team reasoning cannot be viewed as a payoff transformation. *Economics & Philosophy*: 1-11.
Colman, Andrew M., and Natalie Gold. 2018. Team reasoning: Solving the puzzle of coordination. *Psychonomic Bulletin & Review* 25 (5): 1770-83.
Colman, Andrew M., Briony. D. Pulford, and Catherine. L. Lawrence. 2014. Explaining strategic coordination: Cognitive hierarchy theory, strong stackelberg reasoning, and team reasoning. *Decision* 1 (1): 35-58.
Duijf, Hein. 2021. Cooperation, fairness and team reasoning. *Economics & Philosophy* 37 (3): 413-40.
Elsenbroich, Corinna, and Nicolas Payette. 2020. Choosing to cooperate: Modelling public goods games with team reasoning. *Journal of Choice Modelling* 34: 100203.
Ely, Jeffrey. C., and William H. Sandholm. 2005. Evolution in Bayesian games I: theory. *Games and Economic Behavior* 53 (1): 83-109.
Faillo, Marco, Alessandra Smerilli, and Robert Sugden. 2016. Can a single theory explain coordination? An experiment on alternative modes of reasoning and the conditions under which they are used. CBES [Centre for Behavioural and Experimental Social Science] Working paper.
Fleurbaey, Marc, and Peter J. Hammond. 2004. Interpersonally Comparable Utility. In *Handbook of Utility Theory*, ed. Salvador Barberà, Peter J. Hammond, and Christian Seidl, 1179-285. Boston: Springer.
Galeazzi, Paolo, and Michael Franke. 2017. Smart Representations: Rationality and Evolution in a Richer Environment. *Philosophy of Science* 84 (3): 544-73.
Gold, Natalie. 2017. Team Reasoning: Controversies and Open Research Questions. In *The Routledge Handbook of Collective Intentionality*, ed. Marija Jankovic and Kirk Ludwig, 221-32. New York: Routledge.
Gruene-Yanoff, Till. 2011. Evolutionary game theory, interpersonal comparisons and natural selection: a dilemma. *Biology & Philosophy* 26 (5): 637-54.
Gueth, Werner, and Hartmut Kliemt. 1998. The indirect evolutionary approach: bridging the gap between rationality and adaption. *Rationality and Society* 10 (3): 377-99.
Hakli, Raul, Kaarlo Miller, and Raimo Tuomela. 2010. Two kinds of we-reasoning. *Economics & Philosophy* 26 (3): 291-320.
Harsanyi, John C. 1967. Games with incomplete information played by “Bayesian” players, I–III Part I. The basic model. *Management Science* 14 (3): 159-82.
Harsanyi, John. C., and Reinhard Selten. 1988. *A general theory of equilibrium selection in games*. MIT Press Books 1.
Istrate, Gabrial. 2021. Game-theoretic Models of Moral and Other-Regarding Agents. Proceedings of the 18th Conference of Theoretical Aspects of Rationality and Knowledge (TARK) 2021.
Karpus, Jurgis, and Mantas Radzvilas. 2018. Team reasoning and a measure of mutual advantage in games. *Economics & Philosophy* 34 (1): 1-30.
Kuhn, Steven. 2019. Prisoner’s Dilemma. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford: Metaphysics Research Lab, Stanford University.

Lecouteux, Guilhelm. 2015. *Reconciling Normative and Behavioural Economics.* Theses: Ecole Polytechnique.

Lempert, Daniel. 2018. On Evolutionary Game Theory and Team Reasoning. *Revue d'économie politique* 128 (3): 423-46.

Lewontin, Richard C. 1970. The units of selection. *Annual Review of Ecology and Systematics* 1 (1): 1-18.

McElreath, Richard, and Robert Boyd. 2008. *Mathematical models of social evolution: A guide for the perplexed*. Chicago: University of Chicago Press.

Newton, Jonathan. 2017. Shared intentions: The evolution of collaboration. *Games and Economic Behavior* 104: 517-34.

Okasha, Samir. 2006. *Evolution and the Levels of Selection*. Oxford University Press.

Ostrom, Elinor. 1990. *Governing the commons: The evolution of institutions for collective action*. Cambridge: Cambridge University Press.

Pacuit, Eric, and Olivier Roy. 2017. Epistemic Foundations of Game Theory. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford: Metaphysics Research Lab, Stanford University.

Paternotte, Cédric. 2018. The Efficiency of Team Reasoning. *Revue d'economie politique* 128 (3): 447-68.

Pulford, Briony D., Andrew M. Colman, and Catherine L. Lawrence. 2014. Strong Stackelberg reasoning in symmetric games: An experimental replication and extension. *PeerJ* 2: e263.

Radzvilas, Mantas, and Jurgis Karpus. 2021. Team reasoning without a hive mind. *Research in Economics* 75 (4): 345-53.

Roemer, John E. 2019. *How We Cooperate*. New Haven and London: Yale University Press.

Rusch, Hannes. 2019. The evolution of collaboration in symmetric 2x2-games with imperfect recognition of types. *Games and Economic Behavior* 114: 118-27.

Schelling, Thomas. C. 1960. *The Strategy of Conflict*. Cambridge: Harvard University press.

Skyrms, Brian. 2004. *The stag hunt and the evolution of social structure*. Cambridge: Cambridge University Press.

Skyrms, Brian. 2014. *Evolution of the social contract*. Cambridge: Cambridge University Press.

Smith, J. Maynard. 1964. Group Selection and Kin Selection. *Nature* 201: 1145-7.

Sober, Elliott, and David Sloan Wilson. 1998. *Unto others: The evolution and psychology of unselfish behavior*. Cambridge and London: Harvard University Press.

Sugden, Robert. 1993. Thinking as a Team: Towards an Explanation of Nonselfish Behavior. *Social Philosophy & Policy* 10 (1): 69-89.

Sugden, Robert. 2000. Team preferences. *Economics & Philosophy* 16 (2): 175-204.

Sugden, Robert. 2003. The Logic of Team Reasoning. *Philosophical Explorations* 6 (3): 165-81.

Sugden, Robert. 2015. Team Reasoning and Intentional Cooperation for Mutual Benefit. *Journal of Social Ontology* 1(1): 143-66.

Weibull, Joergen W. 1997. *Evolutionary game theory*. Cambridge and London: MIT press.

Wilson, David Sloan. 1975. A theory of group selection. *Proceedings of the National Academy of Sciences* 72 (1): 143-6.

HSR

Historical Social Research

All articles published in HSR Special Issue 48 (2023) 3: The Emergence and Effects of Non-hierarchical Collective Agency

Historical Social Research

All articles published in HSR Special Issue 48 (2023) 3:
The Emergence and Effects of Non-hierarchical Collective Agency