Economics & Philosophy (2021), **37**, 413–440
doi:10.1017/S0266267120000413

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Cooperation, fairness and team reasoning

Hein Duijf

VU Amsterdam, de Boelelaan 1105, 1081 HV Amsterdam, the Netherlands
Email: h.w.a.duijf@vu.nl

(Received 9 April 2020; revised 21 September 2020; accepted 28 September 2020; first published online 1 February 2021)

Abstract

This paper examines two strands of literature regarding economic models of cooperation. First, payoff transformation theories assume that people may not be exclusively motivated by self-interest, but also care about equality and fairness. Second, team reasoning theorists assume that people might reason from the perspective of the team, rather than an individualistic perspective. Can these two theories be unified? In contrast to the consensus among team reasoning theorists, I argue that team reasoning can be viewed as a particular type of payoff transformation. However, I also demonstrate that many payoff transformations yield actions that team reasoning rules out.

Keywords: Cooperation; team reasoning; fairness; other-regarding preferences; game theory

JEL classification: D63 (Equity; Justice; Inequality; and Other Normative Criteria and Measurement); C70 (Game Theory and Bargaining Theory); A13 (Relation of Economics to Social Value)

1. Introduction

Many economic models assume that people are exclusively pursuing their material self-interest, as opposed to pursuing social goals. Experimental economists have been gathering evidence that people often violate this assumption of self-interest (Güth *et al.* 1982; Fehr *et al.* 1993; Berg *et al.* 1995). The standard egoistic economic model can be adjusted to yield a model that fits these empirical findings by amending the assumption that people are exclusively pursuing their material self-interest to the assumption that people care about *fairness* considerations (Rabin 1993; Fehr and Schmidt 1999; Bolton and Ockenfels 2000; Charness and Rabin 2002). Simply stated, these theoretical models presuppose that people care about the distribution of the gains. They hence challenge the egoistic assumption in standard economic models. In effect, they stipulate a transformation of the payoffs, so let us refer to these as *payoff transformation theories*.

A conceptual clarification is in order. The scientific target of payoff transformation theories is to study *what* motivates people. To be more precise, these theories generally start with a material game, that is, a game that incorporates the material payoffs of the players, and then propose a transformation that turns this material game into a motivational game that takes into account some motivationally relevant factors.[1] (The egoistic assumption can then be characterized as positing that the only motivationally relevant factors are a player's own material payoffs.) There are, then, two challenges. First, we need to rigorously characterize the various motivationally relevant factors. Mathematical models have been introduced to characterize the nature of these factors and these models arguably provide an intelligible and systematic explication of the factors. Second, given such a rigorous theory, one can accurately formulate the behavioural predictions and, consequently, determine which experimental evidence would validate or falsify the proposition that a given set of motivational factors influence people's behaviour.

Another strand of literature emerged in ethics, more specifically, in utilitarian theory, which focused on the observation that a group may together fail to promote deontic utility even if each player performs an individual action that promotes deontic utility (Hodgson 1967; Regan 1980). Stated differently, it is perceivable that all the members of a group fulfil their individual moral obligation even though there is an alternative group action that would have benefited the group more. The problem arises when the members find themselves in a *coordination* problem. In recent years, these philosophical insights have been picked up and cultivated by some economists who developed the theory of *team reasoning* (Sugden 1993; Bacharach 2006). Furthermore, the theory of team reasoning has been shown to explain empirical results from a set of lab experiments regarding coordination games (Colman *et al.* 2008; Bardsley *et al.* 2010; Butler 2012; Bardsley and Ule 2017; Pulford *et al.* 2017). The unorthodox feature of these models is that they challenge the individualistic assumption in rational choice theory, that is, they allow for players to conceive of a decision problem as a problem *for the team* rather than for themselves. A team reasoner asks herself 'What should *we* do?' as opposed to asking herself 'What should *I* do?'. In effect, this induces an *agency transformation*.

Can these economic models be unified? A unification would help provide a rigorous model that supports both the experimental findings regarding fairness and those regarding coordination problems. However, the consensus among team reasoning theorists is that the action recommendations yielded by team reasoning cannot be explained by payoff transformations – at least, not in a credible way (Bacharach 1999; Colman 2003). I call this the *incompatibility claim*.

What is the importance of this incompatibility claim? First, if this incompatibility claim were true, then it would be *necessary* to include team reasoning theories in our theoretical toolbox in order to explain certain behavioural findings. The crucial problem arises in forms of coordination in which two or more individuals try to coordinate their actions in order to achieve a common goal. In particular, team reasoning theorists claim that payoff transformation theories cannot explain why people manage to jointly select an outcome that is best for everyone in a common-interest scenario.[2] This observation can be rephrased in two ways: the

[1] In contrast, it is common practice in game theory to take the utilities of a game to already include all motivationally relevant factors.

[2] Colman and Gold (2018: 1770) concur: "Orthodox game theory cannot explain such intuitively obvious forms of coordination as the selection of an outcome that is best for all in a common-interest game."

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

negative upshot is that payoff transformation theories are defective in that they cannot explain this finding; and, the positive upshot is an independent argument for endorsing team reasoning theories.

Second, one pivotal issue in the philosophy of the social sciences concerns whether collective intentionality and collective action can be explained in terms of standard individualistic intentional attitudes. This is an instance of the more general discussion on methodological individualism (Heath 2015). If the incompatibility claim were true, then this would generate a key argument for a negative answer to this issue (Tuomela 2013).[3] The latter conceptual distinction is important because it may have further consequences for the central debates about the tenability of methodological individualism.[4]

Alas, I argue *against* the incompatibility claim and demonstrate that team reasoning can be viewed as a kind of payoff transformation theory.[5] To be more precise, I show that there is a payoff transformation that yields the same behavioural predictions and action recommendations. I call this payoff transformation the theory of *participatory motivations*, because it indicates that people care about the group actions they participate in.

The second main goal is to compare and contrast team reasoning – and, by extension, participatory motivations – with existing payoff transformation theories. To illustrate this prospect, three well-known models from the payoff transformation literature will be investigated. The resulting insights yield three general impossibility results: a large class of payoff transformation theories recommends actions that team reasoning rules out. In other words, none of these payoff transformation theories can explain the action recommendations that are yielded by team reasoning. Hence, if payoff transformation theories wish to explain the experimental finding of coordination games, then a rather unorthodox payoff transformation is *required*.

The paper proceeds as follows. I start with some preliminaries regarding game theory and, in particular, explicate the notion of a material game (§2). Those familiar

[3] For instance, Tuomela (2013: 15) writes: “The social world can be adequately understood and rationally explained only with the help of we-mode concepts expressing full-blown collective intentionality and sociality in addition to I-mode concepts. We-mode thinking and reasoning is not conceptually reducible to I-mode reasoning; i.e., it is not definable by, or functionally constructible from, I-mode notions, nor does it seem fully explainable in terms of the I-mode framework. The central reason for this is that it employs a different reasoning mechanism that relies on groups (collective agents) as the basic agents of reasoning. These differences lead to functional differences.”

[4] Although the exact consequences for methodological individualism may not be trivial, Tuomela (2013: 213–214), for example, writes: “If we-mode reasoning and I-mode reasoning produce different rational behavior and if, as the discussed empirical evidence for we-reasoning suggests, human beings sometimes reason in the we-mode, methodological individualism is not sufficient as a foundation of the social sciences in either a prescriptive, rational, or descriptive sense. Theories that rely on the assumption of methodological individualism need revision or complementation.”

[5] It might be helpful to note that it is not my aim to render team reasoning superfluous. On the contrary, I believe that the explication of team reasoning in terms of participatory motivations will help bridge the two paradigms and facilitate a co-evolution of ideas. The theory of team reasoning could, for example, include the ideas of fairness and reciprocity in the explication of team preferences. I agree with Colman and Gold (2018: 1770) who draw “attention to exciting opportunities that appear to exist for incorporating team reasoning into social identity theory and also into the theory of cooperative social value orientation” and demonstrate “how certain psychological theories could be strengthened significantly by incorporating team reasoning”.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

		Responder	
		Reject	Accept
Proposer	Selfish	0 0	2 8
	Equal	0 0	5 5

Figure 1. Ultimatum game.

with game theory can decide to skip most of this section; however, for the purposes of this paper it is vital to become familiar with the distinction between material payoffs and personal motivation. An introduction to some theories and models of payoff transformation (§3) and team reasoning (§4) follows. The impossibility result and possibility result regarding the relation between payoff transformations and team reasoning are presented and discussed in §5. Finally, I conclude with a discussion of the empirical and theoretical ramifications of my findings.

2. Game theory and material games

Starting with the seminal work by von Neumann and Morgenstern (1944), the theory of games has been further developed and applied to study a wide range of phenomena and topics. The theory provides a useful framework for thinking about interdependent decision problems (Schelling 1960). I will begin by considering an example and its game-theoretical model. Then, I set the stage for the remainder of the paper by providing some definitions and conceptual clarifications.

Ultimatum game. Suppose a proposer and a responder bargain about the distribution of a cake. For simplicity's sake, let us suppose that there are two available distributions: the 80–20 split and the 50–50 split. The proposer can propose a particular distribution of the cake and the responder can either decide to accept the offer, or to reject it. If the responder accepts the offer, then each gets their allocation. If, however, the responder rejects the offer, both will get nothing; see Figure 1.[6]

This example highlights the two fundamental components of a game-theoretical model: the game form and the utilities. A *game form* involves a finite set N of individual agents. Each individual agent i in N has a non-empty and finite set A_i of available individual actions. The Cartesian product $\times_{i \in N} A_i$ of all the individual agents' sets of actions gives the full set A of action profiles.[7] The outcome function o selects for each action profile the resulting outcome $o(a)$

[6] I take this description from Forsythe *et al.* (1994). My presentation of the ultimatum game has been simplified in two ways: (1) the proposer cannot offer the distribution of the 20–80 split, and (2) the game is presented as a normal-form game where the agents act simultaneously instead of an extensive game where the proposer goes first. These simplifications help me avoid unnecessarily getting into the technical details of Rabin's model in §5 while retaining the conclusion that (equal, accept) could be a fairness equilibrium.

[7] I adopt the notational conventions of Osborne and Rubinstein (1994: sec. 1.7) and omit braces if the omission does not give rise to ambiguities.

from the set of possible outcomes X. These games are generally called *normal-form* games or strategic-form games.[8]

Definition 1 (Game Form). A *game form* S is a tuple $\langle N, (A_i), X, o\rangle$, where N is a finite set of individual agents, for each agent i in N it holds that A_i is a non-empty and finite set of actions available to agent i, X is a finite set of possible outcomes, and o is an outcome function that assigns to each action profile a an outcome $o(a) \in X$.

Let me mention some additional notational conventions and some derivative concepts. I use a_i and a_i' as variables for individual actions in the set A_i and I use a and a' as variables for action profiles in the set A. For each group $\mathcal{G} \subseteq N$ the set $A_{\mathcal{G}}$ of group actions that are available to group $\mathcal{G}$ is defined as the Cartesian product $\times_{i\in\mathcal{G}} A_i$ of all the individual group members' sets of actions. I use $a_{\mathcal{G}}$ and $a'_{\mathcal{G}}$ as variables for group actions in the set $A_{\mathcal{G}}$ $(= \times_{i\in\mathcal{G}} A_i)$. Moreover, if $a_{\mathcal{G}}$ is a group action of group $\mathcal{G}$ and if $\mathcal{F} \subseteq \mathcal{G}$, then $a_{\mathcal{F}}$ denotes the subgroup action that is $\mathcal{F}$'s component subgroup action of the group action $a_{\mathcal{G}}$. I let $-\mathcal{G}$ denote the relative complement $N - \mathcal{G}$. Finally, if $\mathcal{F} \cap \mathcal{G} = \emptyset$, then any two group actions $a_{\mathcal{F}}$ and $a_{\mathcal{G}}$ can be combined into a group action $(a_{\mathcal{F}}, a_{\mathcal{G}}) \in A_{\mathcal{F}\cup\mathcal{G}}$.

It is typically assumed that a *utility function* u_i of a given individual i assigns to each outcome x a value $u_i(x)$. However, we will assume that such a utility function u_i assigns to each action profile a a value $u_i(a)$. It is easy to see that the latter is a generalization of the former. Let us call a utility function u_i outcome-based if and only if for every $a, b \in A$ it holds that $u_i(a) = u_i(b)$ if $o(a) = o(b)$. The take-home message is that this generalization allows for non-outcome-based utility functions.

A utility function can be used to represent many different things. It is typically used by rational choice theorists to represent the preferences of an agent, or to represent the revealed preferences of an agent (Okasha (2016) provides a useful discussion on decision-theoretical interpretations of utility). But this is not the only available interpretation. Deontic logicians, for instance, use a (typically, binary) utility function to represent a single moral code (see Hilpinen (1971), or, more specifically, Føllesdal and Hilpinen (1971: 15–19)). Depending on the interpretation of the utility function, derived game-theoretical notions should be interpreted differently. The value that an agent i's utility function u_i assigns to an action profile is usually given by a real number, which straightforwardly induces a comparison between action profiles, viz. a yields more utility than b according to u_i if and only if $u_i(a) > u_i(b)$. Depending on the interpretation of the utility function this means that (i) agent i prefers a over b, (ii) agent i always chooses a over b, or (iii) a is deontically better than b.

My focus is on two different interpretations: the personal material payoff, denoted by m_i, and personal motivation, denoted by u_i. That is, $m_i(a) > m_i(b)$ means that agent i's material payoff that results from action profile a is higher than that associated with b;

[8] These normal-form games can be taken to represent a situation in which several agents act simultaneously. These are contrasted with extensive-form games, which drop this simultaneity assumption and can be taken to represent sequential moves. It is important to note that each extensive-form game can be transformed into a normal-form game, although this transformation will remove the temporal structure.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

and $u_i(a) > u_i(b)$ means that agent i cares more about action profile a than about b (the latter will be important in §3).

Definition 2 (Material Game). A *material game* S is a tuple $\langle N, (A_i), X, o, (m_i) \rangle$, where $\langle N, (A_i), X, o \rangle$ is a game form, and for each agent i in N it holds that m_i is a material payoff function that assigns to each action profile a in A a value $m_i(a) \in \mathbb{R}$.

Rational choice theory has produced many solution concepts. I will follow the dominant practice in the social sciences and focus on the *Nash equilibrium*, named after John Nash (1950, 1951). Stated simply, Ann and Bob are in a Nash equilibrium if Ann is making the best decision she can, given Bob's actual decision, and Bob is making the best decision he can, given Ann's actual decision. Likewise, a group of agents are in a Nash equilibrium if each agent is making the best decision she can, given the actual decisions of the others. A Nash equilibrium is typically taken to represent a state in which no one has an incentive to deviate, given the choices of the others.[9] To illustrate, the Nash equilibria in the discussed ultimatum game is (selfish, accept).

Definition 3 (Nash Equilibrium). Let $S = \langle N, (A_i), X, o, (u_i) \rangle$ be a game. Then an action profile a is a Nash equilibrium if and only if for each agent i in N and for every $b_i \in A_i$ it holds that $u_i(a) \geq u_i(b_i, a_{-i})$.

3. Payoff transformation theories

It is well known that standard economic models are unable to explain some trivial examples of human decision-making. For instance, the traditional theory of self-interested rational individuals cannot explain, at least not satisfactorily, why people vote, pay their taxes, or sacrifice their own prospects in favour of those of a peer. Following psychological research dating back to the 1950s, experimental economists started to gather further evidence in the 1980s and 1990s showing that people diverge from purely self-interest and these findings could be replicated. These observations include the fact that people are willing to sacrifice part of their own material payoff to the benefit of another.

To illustrate this conflict between empirical evidence and theoretical predictions, recall the ultimatum game (Figure 1). As noted before, the only Nash equilibrium is (selfish, accept). This means that standard egoistic rational choice theory predicts that proposers will choose the selfish distribution and that responders will accept this offer. The experimental evidence, in contrast, shows that people generally choose the fair distribution and that responders choose to reject selfish offers. Both empirical findings are at odds with the predictions of standard rational choice theory.

[9] Although the Nash equilibrium concept is widely accepted for descriptive purposes, its application in normative domains cannot be straightforwardly justified (Risse 2000). Attempts to justify Nash equilibria based on epistemic conditions gave rise to the field of epistemic game theory (see Perea 2012), originating from the work on rationalizability (Bernheim 1984; Pearce 1984).

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

To explain these empirical results, economists have proposed several theoretical models that take considerations of fairness and reciprocity into account.[10] For example, Ernst Fehr and Klaus Schmidt write:

> We model fairness as self-centered inequity aversion. Inequity aversion means that people resist inequitable outcomes; i.e., they are willing to give up some material payoff to move in the direction of more equitable outcomes. Inequity aversion is self-centered if people do not care per se about inequity that exists among other people but are only interested in the fairness of their own material payoff relative to the payoff of others. (Fehr and Schmidt 1999: 819)

These and similar models have been used to fruitfully explain various empirical findings (Rabin 1993; Bolton and Ockenfels 2000; Fehr and Schmidt 2006).[11] It is instructive to look a bit more closely at the theoretical models proposed by Fehr and Schmidt (1999). For my current purposes, it will be helpful to do so on the basis of, what I call, motivational games:

Definition 4 (Motivational Game). A *motivational game* S is a tuple $\langle N, (A_i), X, o, (m_i), (u_i)\rangle$, where $\langle N, (A_i), X, o, (m_i)\rangle$ is a material game, and for each agent i in N it holds that u_i is a personal motivation function that assigns to each action profile a in A a value $u_i(a) \in \mathbb{R}$.[12]

The models proposed by Fehr and Schmidt (1999) incorporate the intuition that people dislike inequitable outcomes. This experience of inequity works in both directions: people feel disadvantageous inequity if they are worse off than others; and they feel advantageous inequity if they are better off than others. These two experiences are modelled using two parameters: α_i represents agent i's disutility from disadvantageous inequality; β_i represents agent i's disutility from advantageous inequality. The motivation function is then given by the following equation:

$$u_i(x) = m_i(x) - \alpha_i \frac{1}{|N|-1} \Sigma_{j \neq i} \max[m_j(x) - m_i(x), 0]$$
$$- \beta_i \frac{1}{|N|-1} \Sigma_{j \neq i} \max[m_i(x) - m_j(x), 0]$$

[10] Of course, the proposition that people are not exclusively self-interested has long been recognized in social psychology and sociology. Social exchange theorists Van Lange and Balliet (2015: 68), for example, write: "Interaction situations may be subject to transformations by which an individual considers the consequences of his or her own (and other's) behavior in terms of outcomes for self and other and in terms of immediate and future consequences. Transformation is a psychological process that is guided by interaction goals, which may be accompanied and supported by affective, cognitive, and motivational processes."

[11] To illustrate the popularity of such payoff transformations, note that Bicchieri (2006: 3) studies social norms as a type of payoff transformation: "social norms, as I shall argue, *transform* mixed-motive games into coordination ones. This transformation, however, hinges on each individual expecting enough other people to follow the norm, too. If this expectation is violated, an individual will revert to playing the original game and to behaving 'selfishly.'"

[12] These models should not be confused with so-called psychological games (Geanakoplos *et al.* 1989). Psychological games explicitly model the beliefs and expectations of the agents, while these are absent in motivational games.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

where $|N|$ is the cardinality of the set of individual agents, and it is assumed that $\beta_i \leq \alpha_i$ and $0 \leq \beta_i < 1$. The resulting motivation function thus reflects a combination of personal material payoff and relative material payoff.

These payoff transformations are related to the socio-psychological literature on *social value orientations* (SVO) (Deutsch 1949; Messick and McClintock 1968; McClintock 1972, see Van Lange 1999; Murphy and Ackermann 2014 for more recent work), where different SVO types are represented by *linear* payoff transformations. Let us consider the simple two-player case involving agent i and j. It is common to distinguish four basic SVO types: agent i has an *individualistic* value orientation if $u_i = m_i$;[13] agent i has an *altruistic* value orientation if $u_i = m_j$; agent i has a *cooperative* value orientation if $u_i = m_i + m_j$; and, lastly, agent i has a *competitive* value orientation if $u_i = m_i - m_j$.[14]

Moreover, Van Lange (1999) has introduced the so-called egalitarian social value orientation, which readily corresponds to inequity aversion in the model by Fehr and Schmidt (1999). In its most general form, social value orientations can be modelled using motivation functions given by the following equation:

$$\begin{aligned} u_i(x) = & w_i \cdot m_i(x) + \Sigma_{j \neq i} w_j \cdot m_j(x) \\ & - \alpha_i \frac{1}{|N|-1} \Sigma_{j \neq i} \max[m_j(x) - m_i(x), 0] \\ & - \beta_i \frac{1}{|N|-1} \Sigma_{j \neq i} \max[m_i(x) - m_j(x), 0] \end{aligned}$$

where $|N|$ is the cardinality of the set of individual agents, it is assumed that $\beta_i \leq \alpha_i$ and $0 \leq \beta_i < 1$, and it is assumed that $w_i \in \mathbb{R}$ and $w_j \in \mathbb{R}$ for every $j \in N$. These motivation functions indicate that people may care about their own material payoff, the material payoff of others and that they may dislike inequitable outcomes. Let us call these *SVO motivation functions*.

It should be noted that these particular payoff transformation theories are outcome-based. That is, the postulated motivation functions only depend on the outcome, not on the way that outcome came about. There are other payoff transformation theories that involve motivation functions that are not outcome-based. In the literature on social preferences, these are commonly called intention-based, since these motivation functions are based on whether others' actions are performed with kind or unkind intentions.[15] Let us briefly go over the general components of the models introduced by Rabin (1993) (also see Dufwenberg and Kirchsteiger 2004; Falk and Fischbacher 2006; Segal and Sobel 2007).

Matthew Rabin writes that we need a model that incorporates three facts:

[13] To avoid confusion, note that this means that the *functions* u_i and m_i are identical. That is, $u_i = m_i$ means that for any profile a it holds that $u_i(a) = m_i(a)$. Similarly, $u_i = m_i + m_j$ means that for every profile a we have $u_i(a) = m_i(a) + m_j(a)$.

[14] I am grateful to Andrew M. Colman for pointing out the literature on social value orientations and for the recommendation to extend my general impossibility results to include these payoff transformations.

[15] Theories of psychological game theory involve the assumption that personal motivations might depend on the expectations or beliefs of the agents. For simplicity's sake, I refrain from discussing (higher-order) beliefs and expectations here.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

> (A) People are willing to sacrifice their own material well-being to help those who are being kind.
> (B) People are willing to sacrifice their own material well-being to punish those who are being unkind.
> (C) Both motivations (A) and (B) have a greater effect on behavior as the material cost of sacrificing becomes smaller. (Rabin 1993: 1282)

Without going into unnecessary technical details,[16] it is helpful to note that this payoff transformation relies on so-called *kindness* functions. Given an action profile, these kindness functions can be used to indicate how kind player i is being towards player j. To illustrate this notion of kindness, recall the ultimatum game, which is depicted in Figure 1. Let us consider the action profile (selfish, accept). The kindness of the proposer toward the responder is a function of the set of personal material payoffs of the responder that could result when the responder conforms to (selfish, accept). The idea is that, given the responder's choice to accept, the proposer effectively chooses between the selfish distribution of (8,2) and the equal distribution of (5,5). In particular, this means that the proposer decides to choose the distribution that yields a material payoff of 2 for the responder whereas she could have chosen a distribution that would have yielded a material payoff of 5 for the responder. This means that the proposer is being unkind to the responder.

Let us formalize some of these ideas to illustrate a few properties of these kindness functions. It suffices for our purposes to focus on two-player games. Consider a two-player material game S and a particular profile $a = (a_1, a_2)$. As indicated before, given the action profile a, the kindness of player 1 towards player 2 can be determined by considering the set of action profiles that are compatible with 2's choice, that is, $A^{conf}(a_2) := \{b \in A | b_2 = a_2\}$, and their associated material payoffs $M_2^{conf}(a_2) := \{m_2(b) | b \in A^{conf}(a_2)\}$. First, note that this means the kindness of player 1 towards player 2 at action profile a does not depend on action profiles in which all players deviate from a. That is, the kindness at a depends only on action profiles where some players conform to a. Second, in particular, if $m_2(a)$ is maximal in $M_2^{conf}(a_2)$, then player 1 is not being unkind to player 2. Third, if $m_2(a)$ is minimal in $M_2^{conf}(a_2)$, then player 1 is not being kind to player 2.[17] These observations will be crucial for the impossibility result and the accompanying discussion in §5.

The resulting personal motivation functions in Rabin's framework incorporate material payoffs and a notion of reciprocity in terms of kindness functions. These models indicate that players wish to treat others in kind, that is, player i wishes to treat player j kindly only if j treats her kindly and vice versa. This means that the personal motivation functions are not purely outcome-based. After all, their personal motivation depends on how the other agents treat them.

[16] For example, one complexity arises from the fact that his model involves the beliefs and expectations of the agents. The specification of a fairness equilibrium, however, does not require the specification of these beliefs and expectations, because it imposes the condition that "all higher-order beliefs match actual behavior" (Rabin 1993: 1287–1288).

[17] Note that this is slightly different from saying that player 1 is unkind to player 2. After all, an action could be kind, unkind, or neither.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

4. Team reasoning

The idea of team reasoning originates from ethical theories of utilitarianism.[18] The core idea of team reasoning is that a member of a group asks herself 'What should *we* do?' rather than 'What should *I* do?'. Team reasoning hence relies on a we-perspective. This means that a team reasoner first considers the group actions available to the group, assesses these group actions in terms of their consequences, finds the group action that best furthers their common or collective interests, and then chooses her component of that group action.

To explain team reasoning in more detail and to contrast it with traditional rational choice theory, let us consider the Hi-Lo game depicted in Figure 2. Team reasoning theorists claim that traditional rational choice theory does not adequately address the Hi-Lo game. It seems that (*high*, *high*) is the only rational solution, but the players have no reason for preferring one action over the other if they are guided by traditional rational choice theory. Let us see why. The Hi-Lo game contains two Nash equilibria: (*high*, *high*) and (*low*, *low*).[19] In particular, the action profile (*low*, *low*) represents a state of equilibrium in which no one has an incentive to deviate from performing her component individual action, given that everyone else performs their respective part.

As a response to this multiplicity, one may want to refine the Nash equilibria by appealing to the Pareto dominance of (*high*, *high*) over (*low*, *low*) in order to select the Nash equilibrium (*high*, *high*) as the only rational solution.[20] There are two problems with such a solution. First, what is the status of the Pareto principle in standard rational choice theory? Hollis and Sugden (1993: 13) argue that the Pareto principle "is a principle of rationality only to players who conceive of themselves as a team, but not for players who do not". It is therefore plausible that the rationalization of the Pareto principle requires a departure from the central assumption in rational choice theory that agency is only invested in individuals. Second, in any case, such a (Pareto-dominant) Nash equilibrium only captures a possible status-quo: *if* everyone expected the others to play their part in the Nash equilibrium, *then* they would have a reason to do the same. It hence gives only a conditional recommendation and triggers an infinite regress of reasons. I need not fully rehearse the problems for traditional rational choice theory here, but the gist of the paradox should be clear at this point: the theory does not rule out (*low*, *low*) and fails to select (*high*, *high*) as the unique solution to the Hi-Lo game.[21] This is not merely a theoretical glitch, but it is also at odds with experimental evidence that people generally choose *high*.[22] This

[18] Hodgson (1967) used it to demonstrate that rule and act utilitarianism rely on different modes of reasoning. Regan (1980) later expanded on this argument in his theory of 'cooperative utilitarianism'. In the nineties Sugden (1991, 1993) fruitfully introduced team reasoning to the field of game theory. Similar ideas have been proposed by Anderson (2001); Hurley (1989); and Gilbert (1989*a*).

[19] I will restrict my discussion to pure strategies. This restriction is harmless as mixed strategies do not succeed in addressing the present worries.

[20] Harsanyi and Selten (1988) argue for Pareto dominance as a principle of equilibrium selection. Later, Harsanyi (1995) gave risk dominance prominence over Pareto dominance.

[21] Hollis and Sugden (1993), Sugden (2000: 179–182) and Bacharach (2006: 35–68) provide more elaborate treatments of these objections to traditional rational choice theory. See also Colman (2003) and Gilbert (1989*b*).

[22] For experimental evidence for team reasoning, see Colman *et al.* (2008), Bardsley *et al.* (2010), Butler (2012), Bardsley and Ule (2017), and Pulford *et al.* (2017).

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

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

Figure 2. The Hi-Lo game.

inadequate response to the Hi-Lo game by traditional rational choice theory stands to be corrected, which is what team reasoning (as studied by Bacharach, Sugden, and Gold) has been designed for.[23]

For our current purposes, it suffices to investigate the question of whether payoff transformation theories can rule out (*low*, *low*) in the Hi-Lo game. This investigation will take centre stage in the next section. Notice that this question is more narrow than the question of whether rational choice theory can rule out (*low*, *low*).

Bacharach (2006: Ch. 1) and Sugden (2000: sec. 2, 3, 7 and 8) argue that traditional rational choice theory needs to be augmented with a collectivistic reasoning method to successfully address the Hi-Lo game. Team reasoning theorists appeal to the *reasoning process* by which an individual agent reasons about what to do. An individual agent engaged in team reasoning "works out the best feasible combination of actions for all the members of her team, then does her part in it" (Bacharach 2006: 121). In the Hi-Lo game, this reasoning goes as follows: the row player first identifies (*high*, *high*) as the best combination of individual actions that they can perform and then decides to perform her part in that combination, i.e. *high*. Similar reasoning prescribes *high* for the column player. Team reasoning therefore entails that *high* is the only rational option, and selects (*high*, *high*) as the only rational outcome. It hence solves the Hi-Lo game.

To help clarify the scope and limits of my study, let me briefly discuss two questions that are relevant for operationalizing team reasoning. First, what are team preferences (in contrast to individual preferences)? Team reasoning is generally taken to presuppose team preferences that are used to determine the best combination of individual actions that the team can perform (see Bacharach 1999; Sugden 2000). There are only a handful of proposals in the literature that specify what this team preference generally involves. Sugden (2010, 2011, 2015) seems to rely on a notion of mutual advantage or benefit.[24] Bacharach (2006: 59) on the other hand, hypothesizes that a team reasoner "ranks all act-profiles, using a Paretian criterion".[25] I follow the trend in team

[23] One could say that traditional rational choice theory yields a false positive in the case of the Hi-Lo game: the theory does not rule out all intuitively bad choices. In contrast, another complaint towards rational choice theory says that it also yields false negatives in that it rules out cooperation in the prisoners' dilemma: the theory rules out intuitively plausible choices. Although team reasoning might address this second problem, it is beyond the current paper to argue that it successfully does so.

[24] Karpus and Radzvilas (2018) develop a detailed formal account of mutual advantage.

[25] There has been some discussion on the relation between the team preference and the individual preferences. For example, Gold (2012: 195) writes that Bacharach "allowed in principle that the group objective might be welfare decreasing for some members" and according to Sugden (2000: 176) "the preferences of a team are not necessarily reducible to, or capable of being constructed out of, the preferences that govern the choices that the members of the team make as individuals".

reasoning and will suppose that the team preference is given. The benefit of doing so is that the framework could incorporate any theory of team preferences.

It is helpful to add that team reasoning, interpreted strictly, does not operate on the Hi-Lo game, as represented in Figure 2, since the game only shows the individual preferences. Because team reasoning relies on team preferences, we need to make some assumption regarding these team preferences. It is nonetheless uncontroversial that the group prefers (*high*, *high*) over (*low*, *low*). Team reasoning thus yields a convincing argument for choosing *high* in the Hi-Lo game.[26]

Second, how can the theory of team reasoning be extended to apply to a greater variety of problems – not just the Hi-Lo game?[27] In this essay, I will focus on interdependent decision problems where the team preferences determine a unique best group action.[28] The application of team reasoning is unambiguous in these cases and the Hi-Lo game indicates that traditional rational choice theory can fall short even in these idealized cases. The main reasons for excluding cases where there are multiple best group actions are that team reasoning theorists have only sparingly addressed these cases and it is hard to formulate general desiderata that we would like our decision theory to satisfy.[29]

The discussion of the Hi-Lo game and these two questions for team reasoning should help the reader grasp the idea of team reasoning. Now, let me set forth a simple model of team reasoning. A *team game* extends a material game by adding participation states and team preferences.[30] The participation states of the agents are specified by a function P that assigns to each agent i the team $P(i) \subseteq N$ in which she is active. For each group $\mathcal{G}$ in the range of P, the team preference function $v_{\mathcal{G}}$ assigns to each action profile a in A a value $v_{\mathcal{G}}(a) \in \mathbb{R}$. This illustrates that a given team game presupposes that individuals team reason from the perspective of the team they are active in and rely on the associated team preference.

Definition 5 (Team Game). A *team game* S is a tuple $\langle N, (A_i), X, o, (m_i), P, (v_{\mathcal{G}})\rangle$, where $\langle N, (A_i), X, o, (m_i)\rangle$ is a material game, P assigns to each agent i in N the team

[26] In recent years, Sugden (2015: 156) seems to embrace a theory of team reasoning that relies on notions of *mutual advantage* and *conventions*. For example, he writes: “If individuals are to cooperate effectively, they need to be ready to play their parts in mutually beneficial practices that seem to them to be – and perhaps really are – less than ideal.” His new theory of team reasoning is, for instance, consistent with (*low*, *low*)-play in the Hi-Lo game. Nonetheless, I will stick to the original idea that team reasoning is designed to explain why *high* is uniquely rational.

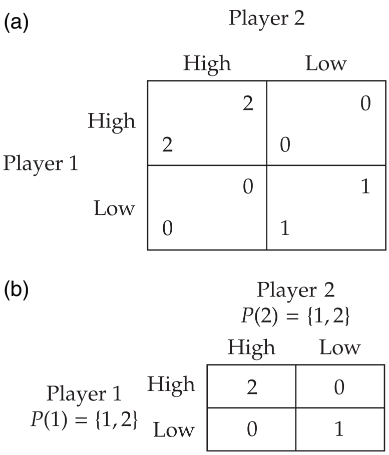
[27] Compare Bacharach (2006: 58) – amended notation: “There are three requirements for a good theory of why people play *high* in Hi-Lo: (i) that it imply observed behaviour, that is, the almost universal choice of *high* in normal circumstances; (ii) that it do so intelligibly to us, which (to the extent that *high* intuitively and stably seems to us the only rational thing to do) involves displaying *high* as uniquely rational – that is, giving principles of rationality which are themselves persuasive, and showing they dictate doing *high*; and (iii) that it be part of a unified theory of a wide range of problems, not just Hi-Lo – for example, all problems of cooperation.”

[28] Bacharach (1999) and Duijf (2018a; b) discuss team reasoning in cases where there are multiple best group actions.

[29] Team reasoning struggles with scenarios in which there are multiple best group actions available. In (Duijf 2018*b*: 445–447), I illustrate this defect by presenting a simple ambiguous Hi-Lo game and argue that this undermines the application of team reasoning to typical collective action problems.

[30] A team game could be viewed as a simplification of what Bacharach (1999) calls unreliable team interactions.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press



(a)

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

(b)

		Player 2 $P(2) = \{1,2\}$	
		High	Low
Player 1 $P(1) = \{1,2\}$	High	2	0
	Low	0	1

Figure 3. An illustration of a team game that yields *high* in the material Hi-Lo game. (a) The material Hi-Lo game. (b) Team game of the material Hi-Lo game, where the numbers represent the team preference of the group $\{1, 2\}$.

$P(i) \subseteq N$ that she is active in, and $v_{\mathcal{G}}$ is the team preference function that assigns to each action profile a in A a value $v_{\mathcal{G}}(a) \in \mathbb{R}$, one for each group $\mathcal{G}$ in the range of P.[31]

As indicated before, team reasoning theorists assume that players do not reason in line with traditional rational choice theory – at least in cases where the agent is active in a non-singleton team. Let us specify how team reasoning works in a given team game S. An agent i first determines the group action for the group $P(i)(= \mathcal{G})$ that best promotes the team preference $v_{\mathcal{G}}$ and then selects the individual action that is her part in that group action. Under the assumption that there is a unique best group action, team reasoning yields a unique individual action for each individual.

To illustrate, let us see how team games might explain the choice of *high* in the Hi-Lo game. First, one could postulate the participation function that assigns to each of the two individuals the team consisting of both. Second, one can postulate that the team preference function is identical to the personal material payoff functions. The resulting team game is depicted in Figure 3. This team game supports the conclusion that each individual will choose *high*. Let us see why. An individual will first determine the group action that best promotes the team preferences. In this particular case the group action (*high*, *high*) does so. Then, she selects the individual action that is her part of this group action, which yields *high*.

5. Team reasoning as a payoff transformation theory

What is the relation between team reasoning and payoff transformation theories? It should be noted that team games generalize motivational games. That is, for each motivational game S we can simply define a corresponding team game by setting

[31] One could think of the team preference function as, what Sugden (2000: 197) calls, "the team-directed preference": "For an individual who engages in team-directed reasoning, her team-directed preferences constitute a ranking of outcomes which she uses, by way of that reasoning, to determine which strategy she chooses." From this perspective, my definition of a team game requires that two agents' team-directed preferences are identical when they are active in the same team.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

$P(i) = \{i\}$ for every agent i in N and where the team preference function is identical to the personal motivation function, that is, $v_i = u_i$. This entails that every motivational game can be viewed as a team game. Therefore, every payoff transformation theory can be incorporated as an agency transformation theory. Note, however, that the corresponding team game only involves participation states that indicate that each individual agent is active in the singleton 'team' consisting of only herself. It is thus a bit of a conceptual stretch to call such a team game an *agency* transformation. After all, the vital revision that team reasoning proposes is that people need not be active in their personal singleton 'team'.

Nevertheless, a payoff transformation theory may be implemented in the theory of team reasoning to yield predicted or observed behaviour in experimental settings. Consequently, experimental evidence for a particular payoff transformation theory is compatible with team reasoning. This should not come as a surprise, since team reasoning allows for both a payoff transformation and an agency transformation.

5.1. The impossibility result

What about the other direction, that is, can team reasoning be viewed as a payoff transformation? It is important to note that payoff transformation theories use rationality principles from orthodox rational choice theory to derive the predicted or observed behaviour in experimental settings. That is, these theories *only* change the utilities; the action recommendations then follow from standard rationality principles.

The theory of team reasoning, in contrast to payoff transformation theories, revises the rationality principles of traditional rational choice theory by adopting a new reasoning method. It is often claimed that team reasoning differs from certain payoff transformation theories:

> [T]eam reasoning differs from, and is more powerful than, adopting the group's objective and then reasoning in the standard individualistic way. (Bacharach 1999: 144)[32]

> Team reasoning is inherently non-individualistic and cannot be derived from transformational models of social value orientation. (Colman 2003: 151)

Two games play an important role in the theory of team reasoning: the Hi-Lo game (see Figure 2) and the famous prisoner's dilemma (see Figure 5a). In particular, team reasoning theorists claim that payoff transformation theories cannot explain the choice of *high* in the Hi-Lo game. In general, they advance the claim that the action recommendations of team reasoning cannot be explained by payoff transformations – at least, not in a satisfactory way. Call this *the incompatibility claim*. For example, Natalie Gold and Robert Sugden write:

[32] On the basis of Bacharach (1999: Theorem 2), Hakli *et al.* (2010: 307, thesis 5) conclude: "We-mode reasoning is not reducible to pro-group I-mode reasoning, i.e. it is not definable by or functionally reconstructable from I-mode reasoning."

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

> By using the concept of agency transformation, [team reasoning] is able to explain the choice of *high* in Hi-Lo. Existing theories of payoff transformation cannot do this. It is hard to see how any such theory could credibly make (*high*, *high*) the unique solution of Hi-Lo . . . .
>
> For Bacharach, the "strongest argument of all" in support of [the team-reasoning account] of cooperation [in the prisoner's dilemma] is that the same theory predicts the choice of *high* in Hi-Lo games. (Bacharach 2006: 173–174)[33]

In contrast to this consensus among team reasoning theorists, I will prove that every team game can be associated with a motivational game in such a way that they yield the same action recommendations. In other words, my results demonstrate that we do not need to transform both the *payoffs* and the *unit of agency*, rather, it suffices to transform only the payoffs.[34] I call this particular payoff transformation the theory of *participatory motivations* (see §5.2 for more details).

My discussion and results demonstrate that participatory motivations are able to explain both the choice of *high* in the Hi-Lo game and the choice to *cooperate* in the prisoner's dilemma. Before getting there, it will be helpful to ask whether team reasoning can be captured by the payoff transformations proposed by Fehr and Schmidt (1999), theories of social value orientations, and Rabin (1993), as discussed in §3. The interrogation of these payoff transformation theories will yield three general impossibility results.

Given the key role of the Hi-Lo game, it will be central to my investigation of payoff transformation theories. As a preliminary remark, the argument in §4 demonstrated that traditional rational choice theory fails to select *high* as the unique solution to the Hi-Lo game. It is therefore vital to remark that the same argument applies to any payoff transformation that transforms the material game associated with the Hi-Lo game into a motivational game in which the action profiles are assigned similar values by the personal motivation functions. More precisely, to rule out *low* for player 1, payoff transformation theories need to yield a motivation function u_1 that satisfies $u_1(\textit{high}, \textit{low}) > u_1(\textit{low}, \textit{low})$. Moreover, to explain the choice of *high* for player 1, the motivation function u_1 needs to satisfy $u_1(\textit{high}, \textit{high}) > u_1(\textit{low}, \textit{high})$.

Let us start with investigating the model proposed by Fehr and Schmidt (1999). Consider the material Hi-Lo game. Recall that team reasoning yields the action

[33] Gold and Sugden (2007: 117) concur: "One motivation for theories of team reasoning is that there are games that are puzzles for orthodox decision theory, in the sense that there exists some strategy that is at least arguably rational and that a substantial number of people play in real life, but whose rationality decision theory cannot explain and whose play it cannot predict." Compare Colman (2003: 183): "Payoff transformations are potentially useful for psychological game theory, notably in Rabin's (1993) "fairness equilibria" . . . but they cannot solve the payoff-dominance problem [i.e. the Hi-Lo game], although it would be pleasant indeed if such a simple solution were at hand."

[34] In contrast, in the conclusion of Bacharach (2006: 173), the editors, Gold and Sugden, write: "One might wonder whether Bacharach needs to transform *both* payoffs *and* agency. If payoffs have been transformed so that they represent the welfare of the two players as a group, doesn't conventional game theory provide an explanation of why each individual chooses *cooperate* [in the prisoner's dilemma]? Not necessarily . . . . Conventional game theory does not show that rational players of [the prisoner's dilemma] will choose *cooperate*. To show that, we need a transformation of the unit of agency."

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

recommendation to play *high* – at least, given the team game depicted in Figure 3. Is it possible to construct a motivational game, using payoff transformations proposed by Fehr and Schmidt (1999), that yields the same prescription? In particular, as noted before, this induces the requirement that $u_1(high, low) > u_1(low, low)$. Hence, this requirement entails that u_1 can decrease even if the personal material payoff increases while the relative material payoff remains unchanged. This is a strange proposition. It is, indeed, easy to verify that this is incompatible with the theory proposed by Fehr and Schmidt (1999). In fact, the resulting personal motivation function is the same as the material payoff function in the Hi-Lo game *regardless of the values of the parameters α and β*. In particular, for any values of the parameters α and β it will be the case that the resulting motivational game does not rule out (*low*, *low*). This is the inspiration for the first impossibility result.

To generalize this observation, consider a particular motivational game S'. Let us use A^{eq} to denote the set of action profiles that yield an equal distribution of material payoffs, that is, A^{eq} is the set $\{a \in A |$ for every $i, j \in N$ it holds that $m_i(a) = m_j(a)\}$. Given an agent i, we propose to call her personal motivation function *equity-Pareto* if it satisfies the following criterion:

(EP) for any action profiles $a, b \in A^{eq}$ it holds that $u_i(a) > u_i(b)$ if and only if for every agent j it holds that $m_j(a) > m_j(b)$.

In other words, agent i prefers the outcome associated with profile a more than the one associated with b if action profile a yields a higher material payoff for all agents and both a and b yield an equal distribution of material payoffs. That is, the personal motivations respect the Pareto principle on the set of action profiles that yield equal distributions. Or, equivalently, they can only violate the Pareto principle on action profiles that yield unequal distributions. It is easy to verify that the model proposed by Fehr and Schmidt (1999) validates this equity-Pareto principle. Moreover, it seems that any plausible payoff transformation theory that relies only on considerations of equality will conform to this principle.

Theorem 1 (First Impossibility Result). *Let S be the material game associated with the Hi-Lo game. Let S' be a motivational game that is based on S. Suppose the personal motivation functions are equity-Pareto. Then S' does not rule out* (low, low).

Proof. Observe that $A^{eq} = A = \{(high, high), (high, low), (low, high), (low, low)\}$. Note that $u_1(low, low) > u_1(high, low)$, since these action profiles each yield equal distributions of material payoff and (*low*, *low*) Pareto dominates (*high*, *low*). Hence, (*low*, *low*) is a Nash equilibrium in the motivational game S'.

The following is an immediate consequence of this impossibility result. Consider a particular payoff transformation theory. If this payoff transformation theory conforms to the equity-Pareto principle, then the action recommendations yielded by this payoff transformation theory are different from those yielded by team reasoning. This finding resonates with the incompatibility claim.

Let us proceed to theories of social value orientation. It can be shown that whenever a given SVO motivation function is compatible with choosing *high*,

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

then it is also compatible with *low*. In other words, there exists no SVO motivation function that yields the same action recommendations as team reasoning in the Hi-Lo game.[35]

Theorem 2 (Second Impossibility Result). *Let S be the material game associated with the Hi-Lo game. Let S′ be a motivational game that is based on S. Suppose the personal motivation functions are SVO motivation functions. Then, S′ cannot be compatible with* (high, high) *while ruling out* (low, low).

Proof. Observe that $A^{eq} = A = \{(high, high), (high, low), (low, high), (low, low)\}$. Therefore, for any profile a in A it holds that $u_1(a)$ can be given by $(w_1 + w_2)\cdot m_1(a)$, where $w_1, w_2 \in \mathbb{R}$. Assume that S' is compatible with $(high, high)$. Then, we must have $u_1(high, high) \geq u_1(low, high)$ and, therefore, $(w_1 + w_2) \geq 0$. However, this entails that $u_1(low, low) \geq u_1(high, low)$. Hence, (low, low) is a Nash equilibrium in the motivational game S'.

Let us now interrogate the model of Rabin (1993).[36] To investigate its application to the material Hi-Lo game, we need to investigate whether at (*low*, *low*) player 2 is willing to sacrifice her material payoff in order to help or punish the other. To do so, we need to first determine whether player 1 is being kind to player 2 at (*low*, *low*). As noted in §3, we need to consider the set of action profiles consisting of (*high*, *low*) and (*low*, *low*), and the associated material payoffs for player 2, which are 0 and 1, respectively. Given this set of action profiles, player 2's material payoff is highest at (*low*, *low*). This entails that player 1 is not being unkind to player 2 at the action profile (*low*, *low*) and it might even be that player 1 is being kind to player 2 (depending on the exact details of the kindness functions). Hence, player 2 might be willing to sacrifice her own material well-being in order to help player 1. However, given player 1's action, player 2 cannot improve player 1's material payoff. That is, $m_1(low, low) > m_1(low, high)$. Hence, player 2 is not motivated to deviate from (*low*, *low*). Similar reasoning shows that player 1 is also not so motivated. In other words, Rabin's proposed motivational game does not rule out (*low*, *low*).

To generalize this observation, take an arbitrary material game *S* and consider a particular motivational game *S*′. Recall that, for any action profile *a*, Rabin's kindness function that assesses whether agent *i* was kind to agent *j* only relies on the set

[35] I am indebted to Andrew M. Colman for making me aware of this general impossibility result and its importance. This impossibility result extends an earlier discussion by Colman (2003: 151–152) of linear payoff transformations: "Although [linear payoff transformation] seems plausible and may help to explain some social phenomena, it cannot explain the payoff-dominance phenomenon or incorporate team reasoning. Team reasoning is inherently non-individualistic and cannot be derived from transformational models of social value orientation." It may be interesting to note that Van Lange and Gallucci (2003: 178), in their commentary, respond by claiming that "a transformation analysis may very well account for the fact that people tend to be fairly good at coordinating in the Hi-Lo Matching game and related situations". However, Colman (2003: 182) replies by arguing that "they misunderstand the problem". In any case, I hope that this impossibility result helps to clarify the exact limitations of the theory of social value orientations.

[36] Team reasoning theorists have considered the question of whether Rabin's theory solves the Hi-Lo game. For instance, Bacharach (2006: 174) writes: "Reciprocity in Rabin's sense does not affect the equilibrium status of (*low*, *low*)." My third impossibility result generalizes this claim (see Theorem 3 below).

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

of action profiles where agent j conforms to a, that is, $A^{conf}(a_j)$. In particular, for the profile (*low*, *low*) this means that the kindness functions of players 1 and 2 only rely on the set $\{$(*high*, *low*), (*low*, *high*), (*low*, *low*)$\}$, that is, they do not rely on (*high*, *high*). Moreover, this means that any agent's kindness function at most relies on the set of action profiles $A^{conf}(a) := \{b \in A|$ there is an agent i such that $a_i = b_i\}$. The set $A^{conf}(a)$ can be thought of as those action profiles that may result from partial conformity to a – as opposed to full-scale deviation.

Consider any action profile a such that each player's material payoff at a is higher than her material payoff associated with any other action profile in $A^{conf}(a)$. Then, none of the agents is being unkind to others while some might even be kind to others (depending on the specific details of the kindness functions). In light of Rabin's first fact, the agents might be motivated to sacrifice their own material payoff in order to help others. However, under this condition, no agent can improve any other agent's material payoff. Moreover, under this assumption, no agent can improve her own material payoff. Hence, the three facts that Rabin wishes to accommodate entail that, under these circumstances, no player is willing to deviate from a.

Given an agent i, we propose to call her personal motivation function *partial-conformity-Pareto* if it satisfies the following criterion:

(PCP) for any action profiles $a, b \in A$ it holds that $u_i(a) \geq u_i(b)$ if

(i) $b \in A^{conf}(a)$, and
(ii) for every agent j it holds that $m_j(a) = \max_{c \in A^{conf}(a)} m_j(c)$.

In other words, no agent assigns a higher utility to action profile b than to a if action profile b partially conforms to a and a Pareto dominates all of the action profiles that partially conform to it. It should be clear that Rabin's models validate this principle. Moreover, it seems that any plausible payoff transformation theory that relies only on considerations of kindness that involve partial conformity will incorporate this principle.

Theorem 3 (Third Impossibility Result). *Let S be the material game associated with the Hi-Lo game. Let S' be a motivational game that is based on S. Suppose the personal motivation functions are partial-conformity-Pareto. Then S' does not rule out* (low, low).

Proof. Note that A^{conf}((*low*, *low*))= $\{$(*high*, *low*), (*low*, *high*), (*low*, *low*)$\}$ and note that for each player j it holds that m_j(*low*, *low*)= $\max_{c \in A^{conf}((\mathrm{low},\mathrm{low}))} m_j(c)$. Hence, u_1(*low*, *low*) $\geq u_1$(*high*, *low*), since conditions (PCP)(i) and (PCP)(ii) hold for (*low*, *low*) and (*high*, *low*). Hence, (*low*, *low*) is a Nash equilibrium in the motivational game S'.

Consider any payoff transformation theory that endorses the partial-conformity-Pareto principle. Then the action recommendations yielded by this payoff transformation theory are different from those yielded by team reasoning. This, again, concurs with the incompatibility claim.

What have we learnt from these impossibility results? Instead of drawing the conclusion that the individualistic assumption in rational choice theory needs to be abandoned in order to explain that (*high*, *high*) is the unique solution of the

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

Hi-Lo game, I take this to mean that the agents need to care about more than equality and reciprocity. The general principles underlying these impossibility results can help to establish that other payoff transformation theories are also incompatible with team reasoning. Moreover, these principles indicate that a rather unorthodox payoff transformation theory is required for explaining why *low* is ruled out. After all, any payoff transformation that conforms to (EP) or (PCP) will be compatible with *low* and, as a consequence, will be incompatible with team reasoning.

5.2. The possibility result

I demonstrate that team reasoning can be viewed as a payoff transformation that incorporates the intuition that people care about the group actions they participate in.[37] In particular, my proposal reflects the following two stylized ideas:

(*) People might be willing to sacrifice their own material well-being to do their part in a collective act, where their part is defined as the individual action they ought to perform if the group is to be successful in realizing a shared goal.

(**) Motivation $(*)$ has greater effect on behaviour as the material cost of sacrificing becomes smaller.

Team reasoning solves the Hi-Lo game by relying on a team game. In this section, the aim is to demonstrate that there are ways to define a motivational game on the basis of a given team game. It may be hard to determine which individual acts would qualify as doing your part in a collective act. For instance, in cases where agent i is active in a team that does not include all the agents she interacts with, it may be hard to determine which group action best promotes the team's utility. Moreover, there may be cases where there are multiple ways to achieve a certain goal and the group is indifferent between each of these. However, when agent i is active in the team consisting of all the agents, then the best group actions are exactly those group actions that yield the highest team utility. In the remainder of this section, we therefore restrict our attention to team games that uniquely identify the best group actions.

Although there are multiple ways to incorporate the idea that agents care about the group actions they participate in, for our current purposes it suffices to show that there is a motivation function that yields the same action recommendations as team reasoning does. For simplicity's sake, we assume that each individual agent i is active in the team consisting of all the agents – that is, $P(i) = N$. Let us introduce some notation to rigorously characterize the idea that agents care about the group actions they participate in. Recall that $A^{conf}(a_i)$ is the set of action profiles that are compatible with agent i's choice a_i. Let $V_i^{conf}(a_i) := \{v_{\mathcal{G}}(b) | P(i) = \mathcal{G} \text{ and } b \in A^{conf}(a_i)\}$ denote the

[37] My personal inspiration comes from Kutz's (2000: 67) theory of participatory intentions; he writes: "I argue that many cases of joint action are best explained by the intentions with which individual agents act. These intentions are what I call *participatory*: Individuals act with the intention of contributing to a collective outcome."

team utilities that are associated with the team that agent i is active in and with the action profiles in $A^{conf}(a_i)$. Consider the following motivation function:

$$u_i^{\max}(a) = \gamma_i \cdot m_i(a) + \delta_i \cdot \max V_i^{conf}(a_i)$$

where $\gamma_i, \delta_i \geq 0$, γ_i represents the utility from the agent's material payoff, and δ_i represents the utility from the group actions one participates in. (The subscript i for the parameters γ and δ is suppressed in case the omission does not give rise to ambiguity.) Let us call these *participatory motivation functions*.[38]

The benefit of including both the parameters γ and δ is that we can empirically test the trade-off between the material payoff and the 'participatory' utility. After all, it seems highly implausible that people will always sacrifice their own material payoff in order to promote participatory utility. Whereas team reasoning theorists typically invoke a strict dichotomy between individualistic reasoning and team reasoning, the proposed participatory motivations allow for a more gradual picture where participatory utility can outweigh material payoffs – without neglecting material payoffs altogether.[39]

To illustrate the participatory motivation function for the Hi-Lo game, consider Figure 4 which depicts the participatory motivational game that is associated with the team game of the Hi-Lo game (Figure 3). It should be clear that (*low*, *low*) fails to be a Nash equilibrium in the participatory motivational game if and only if $\delta > \gamma$. That is, participatory motivations rule out *low* if and only if the agents care more about the group actions they participate in than about their material payoff. Hence, under these assumptions, standard rationality principles can be applied to the participatory motivational game of the Hi-Lo game to yield *high*. More precisely, the application of the standard rationality principles in this participatory motivational game rules out *low*, as desired.

Theorem 4 (Possibility Result for Hi-Lo). *Let S be the material game associated with the Hi-Lo game. Then there exists a motivational game S′, that is associated with S, such that* (high, high) *is the only Nash equilibrium in S′.*

Proof. Consider the motivational game depicted in Figure 4 with $\delta > \gamma$. In the resulting motivational game, only (*high*, *high*) is a Nash equilibrium.

Before generalizing this possibility result, let us elaborate on one way to interpret the values assigned by the participatory motivations in the Hi-Lo game. Consider P1 in the participatory motivational game with $\gamma = 0$ and $\delta = 1$ (see Figure 4). What makes it the case that her utility is 1 at (*low*, *high*) while it is 2 at (*high*, *low*)? Her utilities could be taken to describe what it would feel like for her to end up in one of the cells of the game matrix. Although there is no distinction in the actual outcome, I suggest that the distinction be thought of as follows. At (*high*, *low*) P1 thinks of herself as having counterfactually

[38] Note that there are some obvious alternatives. We could postulate that agents care about the total goodness of the group actions they might participate in. This idea could be incorporated by the following motivation function: $u_i^{\Sigma}(a) = \gamma \cdot m_i(a) + \delta \cdot \Sigma V_i^{conf}(a_i)$. The study of these motivation functions and other alternatives has to be left for future research.

[39] Variants of team reasoning that invoke such a strict dichotomy include Bacharach's (2006) "circumspect team reasoning", Smerilli's (2012) "vacillation between frames".

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

(a)

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

(b)

		Player 2 $P(2) = \{1, 2\}$	
		High	Low
Player 1 $P(1) = \{1, 2\}$	High	2	0
	Low	0	1

(c)

		Player 2	
		High	Low
Player 1	High	$2\gamma + 2\delta$ $2\gamma + 2\delta$	δ 2δ
	Low	2δ δ	$\gamma + \delta$ $\gamma + \delta$

Figure 4. An illustration of the participatory motivational game associated with the Hi-Lo game. (a) Material game of the Hi-Lo game. (b) The team game associated with the Hi-Lo game. (c) The participatory motivational game of the Hi-Lo game.

participated in (*high*, *high*), while at (*low*, *high*) she thinks of herself as having counterfactually participated in (*low*, *low*). In other words, the difference between these outcomes is that P1 either sees herself as doing her part in the collective act (*high*, *high*) or as doing her part in the collective act (*low*, *low*). Since (*high*, *high*) yields a higher team utility than (*low*, *low*), her participatory utility is higher at (*high*, *low*) than at (*low*, *high*).[40]

It is important to realize that participatory motivations can explain cooperation in the prisoner's dilemma. To illustrate this fact, consider the game-theoretical model of the prisoner's dilemma in Figure 5a. To apply the theory of participatory motivations, let me *add* the assumption that the team utility is given by the sum of the individual material payoff.[41] That is, $v_N(a) = \Sigma_{i \in N} m_i(a)$. The participatory motivational game associated with prisoner's dilemma is depicted in Figure 5.

[40] Although these considerations about counterfactuals may seem unorthodox, the theory of fairness by Rabin (1993) can be viewed as relying on similar counterfactuals. Suppose P1's personal motivation is given by Rabin's theory. Then, P1's utility at (*high*, *low*) might be lower than at (*low*, *high*) because P1's feeling of being treated badly by P2 is stronger at (*high*, *low*) than at (*low*, *high*). Let us explain why. First, P2 is being unkind at (*high*, *low*), because P2 could have counterfactually yielded a higher payoff for P1 if P2 had chosen *high* instead of *low*. Second, it is plausible that P2 is being more unkind at (*high*, *low*) than at (*low*, *high*), since she could have counterfactually yielded payoff 2 for P1 at the former profile compared with payoff 1 for P1 at the latter profile.

[41] This is a working assumption, and it has some conceptual problems. For example, it may prescribe self-sacrifice, which goes against the idea of team reasoning for mutual benefit; and it requires interpersonal comparisons of utility. However, at this stage, the modest purpose is to illustrate the participatory motivations in some decision contexts.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

(a)

		Player 2	
		C	D
Player 1	C	3 3	4 0
	D	0 4	1 1

(b)

		Player 2 $P(2) = \{1, 2\}$	
		C	D
Player 1 $P(1) = \{1, 2\}$	C	6	4
	D	4	2

(c)

		Player 2	
		C	D
Player 1	C	$3\gamma + 6\delta$ $3\gamma + 6\delta$	$4\gamma + 4\delta$ 6δ
	D	6δ $4\gamma + 4\delta$	$\gamma + 4\delta$ $\gamma + 4\delta$

Figure 5. An illustration of the participatory motivational game associated with the prisoner's dilemma. (a) Material game of the prisoner's dilemma. (b) The team game associated with the prisoner's dilemma. (c) The participatory motivational game of the prisoner's dilemma.

In the participatory motivational game associated with the prisoner's dilemma, it is easy to see that (C, C) is the only Nash equilibrium if and only if the following inequalities obtain:

$$3\gamma + 6\delta > 4\gamma + 4\delta$$

$$6\delta > \gamma + 4\delta.$$

Hence, the agent cooperates if and only if $2\delta > \gamma$. That is, under these idealized assumptions, an agent cooperates if and only if she cares less than twice as much about her material payoff than about the group actions she participates in. The model thus demonstrates that an agent will cooperate depending on the (agent-specific) trade-off between material self-interest and participatory utility, i.e. depending on the exact values of γ_i and δ_i. It is an advantage that participatory motivations can be used to explain cooperation in the prisoner's dilemma.

Let us generalize the possibility result to team games that include a unique best group action. For this class of team games, team reasoning can be subsumed under participatory motivations. In particular, in these scenarios, we can prove that team reasoning yields the same action recommendations as participatory motivation functions with $\gamma = 0$ and $\delta > 0$.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

Theorem 5 (Possibility Result). *Let S be a material game. Let S_T be a team game based on S, and let i in N be an agent. Suppose that $P(i) = N$ and suppose that v_N yields a unique best group action among those in A. Let $a^* \in A$ be the unique best group action, and let S_P be any participatory motivational game with $\gamma = 0$ and $\delta > 0$. Then, a^* is the only Nash equilibrium in S_P.*

Proof. Take any action profiles $a, b \in A$ with $a_i = a_i^*$ and $b_i \neq a_i^*$. Hence $a^* \in A^{conf}(a_i)$ and $a^* \notin A^{conf}(b_i)$. Since a^* is the *unique* best group action, it holds that $\max V_i^{conf}(a_i) > \max V_i^{conf}(b_i)$. Since $\gamma = 0$, we get $u_i^{\max}(a) = \delta \cdot \max V_i^{conf}(a_i) > \delta \cdot \max V_i^{conf}(b_i) = u_i^{\max}(b)$.

It follows that a_i^* strictly dominates any other available individual action in the participatory motivation game. Therefore, a^* is the only Nash equilibrium in S_P.

In other words, the associated participatory motivational game S_P prescribes each player to perform the individual action that is her component in the best group action that is available to the team she is active in. Or, equivalently, the participatory motivations rule out any individual action that is not her component in the best group action.

It immediately follows that the action recommendations yielded by team reasoning in S_T are the same as those yielded by standard rational choice theory in S_P. Hence, the action recommendations of team reasoning can be equally well explained by a payoff transformation. Or, equivalently, it shows that an important part of team reasoning can be reduced to traditional individualistic reasoning with a particular motivation. So, the effects of the agency transformation in team reasoning can be simulated by a payoff transformation.

To clarify, it is an open question whether the general possibility result (Theorem 5) demonstrates that participatory motivations entirely exclude any form of collective reasoning. After all, the payoff transformations rely on the team game and, in particular, on the team utilities. *If* these team utilities can only be obtained via team reasoning, *then* the corresponding participatory motivations would require team reasoning. Given the lack of consensus on what team utilities are, it is plausible that whether these participatory motivations exclude team reasoning depends on the details of the particular account of team utilities.

One of the main justifications for team reasoning is that it advances cooperative behaviour in Hi-Lo games. Whether the cooperative incentives in the material Hi-Lo game actually lead to team reasoning is not at issue. Team reasoning sets out to address what makes *high* the *only* rational option. As such, the theory of participatory motivations is on an equal footing. After all, the action recommendations resulting from team reasoning coincide with those resulting from participatory motivations in the Hi-Lo game. This justification for the team-reasoning account of cooperation therefore transfers to the theory of participatory motivations.

The previous theorem can be viewed as generalizing this justification for the theory of participatory motivations beyond the Hi-Lo game. The result establishes that whenever there is a unique best group action, then players with certain participatory motivations successfully pick out this best group action. Hence, the team-reasoning account of cooperation is on a par with the theory of participatory motivations with regard to guaranteeing successful cooperation in these scenarios.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

The main possibility result entails that there is a payoff transformation theory that yields the same action recommendations as team reasoning. Therefore, if we compare theories on the basis of the action recommendations that they produce, then payoff transformations are at least as powerful as team reasoning. In other words, the action recommendations yielded by team reasoning can be derived from payoff transformation models. After all, participatory motivations can fully account for the behaviour predicted by team reasoning. It thus seems impossible for team reasoning theorists to uphold the incompatibility claim. We discuss some options to rectify the compatibility claim in the final section.

6. Discussion

The results of team reasoning cannot be explained by the particular payoff transformations of Fehr and Schmidt (1999), nor by those involving social value orientations, nor by those of Rabin (1993). More generally, I have shown that team reasoning cannot be explained by a wide class of payoff transformation theories (Theorems 1, 2 and 3). In contrast to this negative result, I have shown that there is a payoff transformation theory, which relies on participatory motivations, that does yield the same action recommendations as team reasoning (Theorems 4 and 5). This provides evidence against the incompatibility claim, which states that team reasoning cannot be derived from payoff transformation theories – at least, not in a credible way.

Before stepping back and drawing more general conclusions, let me briefly mention several properties of these participatory motivations. First, participatory motivations are process-oriented rather than outcome-based. That is, they incorporate the idea that people care not only about features of the outcome but also about how that outcome came about. In other words, it is conceivable that someone's utility differs between two identical outcomes if they were produced in a different manner. Second, participatory motivations may evaluate a given player's actions differently depending on the *concurrent* actions of the others rather than depending on the past actions of others or on the beliefs regarding the actions of others. In other words, it involves instantaneous reciprocity rather than sequential reciprocity or belief-based reciprocity.[42] Lastly, participatory motivations provide a simple model that allows for trade-offs between self-interest and team reasoning concerns (the latter is represented by the participatory utility).

With regard to the incompatibility claim, my results show that to validate this claim, team reasoning theorists need to go beyond the behavioural realm and, for example, include mental constructs in their theoretical and empirical explanandum. After all, the theory of participatory motivations is able to explain the *behavioural* predictions of team reasoning. It may be helpful to point out that the realm of mental constructs could include beliefs, expectations, preferences, motivations, sentiments, and perhaps even personal identity. To

[42] Although our discussion of the models proposed by Rabin (1993) abstracted away from the beliefs and expectations (see §3), Rabin's models and predictions were supposed to be expectation-based. That is, if I believe that you are unkind, then I am motivated to act unkindly. I take this to be a type of belief-based reciprocity.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

theoretically validate the incompatibility claim, one could try to argue that payoff transformation theories are unable to explain both what makes (*high*, *high*) rational and explain certain sets of beliefs and sentiments.

Let me supplement this observation with an example. In his recent book, Sugden (2018: Ch. 9) discusses the 'paradox of trust' (also see Isoni and Sugden (2019)). To illustrate this paradox, consider the normal-form representation of a trust game in Figure 6.[43] The investor's option of sending can be interpreted as investing one unit of material payoff in a way that will generate five units. The trustee then decides on how to distribute the costs and benefits of this investment. It is therefore natural to interpret the action profile (send, return) as a practice of trust. The paradox is that "in a theory in which individuals are motivated by reciprocity, two individuals cannot have common knowledge that they will both participate in a practice of trust" (Sugden 2018: 222). More precisely, such common knowledge would undermine the idea that the investor's choice of send is kind. Setting aside the question of how this paradox affects existing payoff transformation theories, it is important to note that the paradox involves a behavioural and a mental component: the paradox indicates the impossibility of a practice of trust when agents are motivated by reciprocity *and* commonly know that each participates. Phrased differently, these beliefs and motivations are *jointly* incompatible with the practice of trust.

Alternatively, team reasoning theorists might claim that its *mode of reasoning* cannot be reduced to that of individualistic reasoning.[44] Although I cannot say anything conclusive on this suggestion, a few remarks are in order. First, in the absence of an elaborate theory of reduction, it is hard to ascertain whether the claim is interesting nor whether it is true. Second, to empirically verify the claim, it seems like we would need methods to reliably determine the mode of reasoning that people employ. Lastly, in any case, the fact that the theory of participatory motivations yields the same action recommendations as team reasoning should be taken very seriously. After all, the most compelling argument for the team-reasoning account of cooperation concerns its action recommendations: it predicts *high* in the Hi-Lo game and it explains cooperation in the prisoner's dilemma.

Another theoretical move could be to argue that the notion of participatory motivations does not cohere with other mental constructs in an agent's psychological economy. For example, the participatory motivation suggests that an individual agent's evaluation of an outcome that was brought about by the best group action is distinct from her evaluation of that outcome if it were brought about in a different manner. This property is not standardly included in the outcomes of a game. For instance, in his seminal contribution to decision theory, Savage (1954) models actions as functions from states to outcomes. So standard preferences on outcomes cannot encompass these participatory motivations. However, as we alluded to in §3, within the literature on other-regarding preferences there is a wide class of payoff transformations that are process-oriented. We therefore need not revise the rationality principles of traditional rational choice theory; rather, we need to revise the assumption that motivations are outcome-based.

[43] Although the trust game is standardly represented in extensive-form, as opposed to our normal-form representation, this does not matter for our current purposes.

[44] I thank the anonymous reviewers of this journal for inspiring me to think about this possibility.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

		Trustee	
		Keep	Return
Investor	Hold	0 0	0 0
	Send	5 -1	2 2

Figure 6. Trust game.

Lastly, I would like to emphasize that I am sympathetic to the idea of team reasoning and to the ideas of fairness and reciprocity. It may turn out that a fusion between payoff transformation theories and team reasoning is needed to find the key to cooperation. My aim should thus not be understood as an attempt to reduce one to the other. Rather, by relying on this theoretical unification, we can understand the differences more rigorously and make systematic progress towards modelling human decision-making and cooperation.

Acknowledgements. The author would like to thank the audience at the University of Tampere and would especially like to thank Jan Broersen, Nathalie Gold, Raul Hakli, Jurgis Karpus, Werner Raub, Robert Sugden and Allard Tamminga for helpful comments and stimulating discussions on this paper and related topics. The author gratefully acknowledges funding from the European Research Council (the ERC-2013-CoG project REINS, no. 616512, and the ERC-2017-CoG project SEA, no. 771074).

References

Anderson E. 2001. Unstrapping the straitjacket of 'preference': a comment on Amartya Sen's contributions to philosophy and economics. *Economics & Philosophy* **17**(1), 21–38.

Bacharach M. 1999. Interactive team reasoning: a contribution to the theory of co-operation. *Research in Economics* **53**(2), 117–147.

Bacharach M. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory*, ed. N. Gold and R. Sugden. Princeton, NJ: Princeton University Press.

Bardsley N. and A. Ule 2017. Focal points revisited: team reasoning, the principle of insufficient reason and cognitive hierarchy theory. *Journal of Economic Behavior & Organization* **133**, 74–86.

Bardsley N., J. Mehta, C. Starmer and R. Sugden 2010. Explaining focal points: Cognitive hierarchy theory versus team reasoning. *Economic Journal* **120**(543), 40–79.

Berg J., J. Dickhaut and K. McCabe 1995. Trust, reciprocity, and social history. *Games and Economic Behavior* **10**(1), 122–142.

Bernheim B.D. 1984. Rationalizable strategic behavior. *Econometrica* **52**(4), 1007–1028.

Bicchieri C. 2006. *The Grammar of Society*. Cambridge: Cambridge University Press.

Bolton G.E. and A. Ockenfels 2000. ERC: A theory of equity, reciprocity, and competition. *American Economic Review* **90**(1), 166–193.

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

Charness G. and M. Rabin 2002. Understanding social preferences with simple tests. *Quarterly Journal of Economics* **117**(3), 817–869.

Colman A.M. 2003. Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences* **26**(2), 139–153.

Colman A.M. and N. Gold 2018. Team reasoning: Solving the puzzle of coordination. *Psychonomic Bulletin & Review* **25**(5), 1770–1783.

Colman A.M., B.D. Pulford and J. Rose 2008. Collective rationality in interactive decisions: Evidence for team reasoning. *Acta Psychologica* **128**(2), 387–397.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

Deutsch M. 1949. A theory of co-operation and competition. *Human Relations* **2**(2), 129–152.

Dufwenberg M. and G. Kirchsteiger 2004. A theory of sequential reciprocity. *Games and Economic Behavior* **47**(2), 268–298.

Duijf H. 2018a. Beyond team-directed reasoning: Participatory intentions contribute to a theory of collective agency. *Logique et Analyse* **61**(243), 269–298.

Duijf H. 2018b. Responsibility voids and cooperation. *Philosophy of the Social Sciences* **48**(4), 434–460.

Falk A. and U. Fischbacher 2006. A theory of reciprocity. *Games and Economic Behavior* **54**(2), 293–315.

Fehr E., G. Kirchsteiger and A. Riedl 1993. Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics* **108**(2), 437–459.

Fehr E. and K.M. Schmidt 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* **114**(3), 817–868.

Fehr E. and K.M. Schmidt 2006. The economics of fairness, reciprocity and altruism – experimental evidence and new theories. In *The Handbook of the Economics of Giving, Altruism and Reciprocity*, ed. S.-C. Kolm and J.M. Ythier, **1**, 615–691. Amsterdam: Elsevier.

Føllesdal D. and R. Hilpinen 1971. An introduction. In *Deontic Logic: Introductory and Systematic Readings*, ed. R. Hilpinen, 1–35. Dordrecht: D. Reidel Publishing Company.

Forsythe R., J.L. Horowitz, N.E. Savin and M. Sefton 1994. Fairness in simple bargaining experiments. *Games and Economic Behavior* **6**(3), 347–369.

Geanakoplos J., D. Pearce and E. Stacchetti 1989. Psychological games and sequential rationality. *Games and Economic Behavior* **1**(1), 60–79.

Gilbert M. 1989a. *On Social Facts*. London: Routledge.

Gilbert M. 1989b. Rationality and salience. *Philosophical Studies* **57**(1), 61–77.

Gold N. 2012. Team reasoning, framing and cooperation. In *Evolution and Rationality: Decisions, Co-Operation and Strategic Behaviour*, ed. S. Okasha and K. Binmore, 185–212. Cambridge: Cambridge University Press.

Gold N. and R. Sugden 2007. Collective intentions and team agency. *Journal of Philosophy* **104**(3), 109–137.

Güth W., R. Schmittberger and B. Schwarze 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization* **3**(4), 367–388.

Hakli R., K. Miller and R. Tuomela 2010. Two kinds of we-reasoning. *Economics & Philosophy* **26**(3), 291–320.

Harsanyi J.C. 1995. A new theory of equilibrium selection for games with complete information. *Games and Economic Behavior* **8**(1), 91–122.

Harsanyi J.C. and R. Selten 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Heath J. 2015. Methodological individualism. In *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), ed. E. N. Zalta. https://plato.stanford.edu/archives/spr2015/entries/methodological-individualism/.

Hilpinen R. 1971. *Deontic Logic: Introductory and Systematic Readings*. Dordrecht: D. Reidel Publishing Company.

Hodgson D.H. 1967. *Consequences of Utilitarianism*. Oxford: Clarendon Press.

Hollis M. and R. Sugden 1993. Rationality in action. *Mind* **102**(105), 1–35.

Hurley S.L. 1989. *Natural Reasons*. New York, NY: Oxford University Press.

Isoni A. and R. Sugden 2019. Reciprocity and the paradox of trust in psychological game theory. *Journal of Economic Behavior & Organization* **167**, 219–227.

Karpus J. and M. Radzvilas 2018. Team reasoning and a measure of mutual advantage in games. *Economics & Philosophy* **34**(1), 1–30.

Kutz C. 2000. *Complicity: Ethics and Law for a Collective Age*. Cambridge: Cambridge University Press.

McClintock C.G. 1972. Social motivation – a set of propositions. *Behavioral Science* **17**(5), 438–454.

Messick D.M. and C.G. McClintock 1968. Motivational bases of choice in experimental games. *Journal of Experimental Social Psychology* **4**(1), 1–25.

Murphy R.O. and K.A. Ackermann 2014. Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review* **18**(1), 13–41.

Nash J.F. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences USA* **36**(1), 48–49.

Nash J.F. 1951. Non-cooperative games. *Annals of Mathematics* **54**, 286–295.

Okasha S. 2016. On the interpretation of decision theory. *Economics & Philosophy* **32**(3), 409–433.

https://doi.org/10.1017/S0266267120000413 Published online by Cambridge University Press

Osborne M.J. and A. Rubinstein 1994. *A Course in Game Theory*. Cambridge, MA: MIT Press.

Pearce D.G. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica* **52**(4), 1029–1050.

Perea A. 2012. *Epistemic Game Theory*. Cambridge: Cambridge University Press.

Pulford B.D., A.M. Colman, C.L. Lawrence and E.M. Krockow 2017. Reasons for cooperating in repeated interactions: social value orientations, fuzzy traces, reciprocity, and activity bias. *Decision* **4**(2), 102–122.

Rabin M. 1993. Incorporating fairness into game theory and economics. *The American Economic Review* **83**(5), 1281–1302.

Regan D. 1980. *Utilitarianism and Co-operation*. New York, NY: Oxford University Press.

Risse M. 2000. What is rational about Nash equilibria? *Synthese* **124**(3), 361–384.

Savage L.J. 1954. *The Foundations of Statistics*. New York, NY: John Wiley & Sons.

Schelling T.C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Segal U. and J. Sobel 2007. Tit for tat: Foundations of preferences for reciprocity in strategic settings. *Journal of Economic Theory* **136**(1), 197–216.

Smerilli A. 2012. We-thinking and vacillation between frames: Filling a gap in Bacharach's theory. *Theory and Decision* **73**(4), 539–560.

Sugden R. 1991. Rational choice: a survey of contributions from economics and philosophy. *The Economic Journal* **101**(407), 751–785.

Sugden R. 1993. Thinking as a team: towards an explanation of nonselfish behavior. *Social Philosophy and Policy* **10**(1), 69–89.

Sugden R. 2000. Team preferences. *Economics & Philosophy* **16**(2), 175–204.

Sugden R. 2010. Opportunity as mutual advantage. *Economics & Philosophy* **26**(1), 47–68.

Sugden R. 2011. Mutual advantage, conventions and team reasoning. *International Review of Economics* **58**(1), 9–20.

Sugden R. 2015. Team reasoning and intentional cooperation for mutual benefit. *Journal of Social Ontology* **1**(1), 143–166.

Sugden R. 2018. *The Community of Advantage: A Behavioural Economist's Defence of the Market*. Oxford: Oxford University Press.

Tuomela R. 2013. *Social Ontology: Collective Intentionality and Group Agents*. New York, NY: Oxford University Press.

Van Lange P.A.M. 1999. The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology* **77**(2), 337–349.

Van Lange P.A.M. and D. Balliet 2015. Interdependence theory. In *APA Handbook of Personality and Social Psychology: Vol. 3. Interpersonal Relations*, ed. M. Mikulincer, P.R. Shaver, J.A. Simpson and J.F. Dovidio, 65–92. Washington, DC: American Psychological Association.

Van Lange P.A.M. and M. Gallucci 2003. Bridging psychology and game theory yields interdependence theory. *Behavioral and Brain Sciences* **26**(2), 177–178.

von Neumann J. and O. Morgenstern 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Hein Duijf is a postdoctoral researcher at the VU Amsterdam. His current research takes inspiration from behavioural sciences (including social psychology, economics and social sciences) and agent-based modelling to study the epistemic risks and potential benefits of group deliberation. He has a PhD in Philosophy and Artificial Intelligence from Utrecht University. This second line of scholarship embraces moral responsibility, collective agency, machine ethics and deontic reasoning.

Cite this article: Duijf H (2021). Cooperation, fairness and team reasoning. *Economics and Philosophy* **37**, 413–440. https://doi.org/10.1017/S0266267120000413