Beyond team-directed reasoning: Participatory intentions contribute to a theory of collective agency

Hein Duijf

Abstract

Philosophical accounts of collective intentionality typically rely on members to form a personal intention of sorts, viewed as a *mental state*. This tendency is opposed by recent economic literature on team-directed reasoning (as studied by Bacharach, Gold, and Sugden), which focuses on the *reasoning process* leading up to the formation of the members' intentions. Our formal analysis bridges these paradigms and criticizes the team-directed reasoning account on two counts: first, team-directed reasoning is supposed to transcend traditional game and decision theory by adopting a certain collectivistic reasoning method. However, we show that team-directed reasoning yields the same action recommendations as a certain I-mode we-intention type. Accordingly, an important part of we-mode reasoning can be reduced to I-mode reasoning with certain preferences. Second, contrary to the claims of team-directed-reasoning theorists, we refute that team-directed reasoning surpasses pro-group intentions in selecting cooperatively rational solutions. That is, in some scenarios team-directed reasoning fails to guarantee successful cooperation whereas pro-group intentions succeed in doing so. We therefore propose to revise team-directed reasoning and introduce a third we-intention type, called participatory intentions. We prove that participatory intentions guarantee that a best group action is performed whenever either team-directed reasoning or pro-group intentions do.

Keywords: Team-directed Reasoning · Collective Agency · Participatory Intentions · We-intentions · Game Theory

1 Introduction

The philosophical debate on collective intentionality has recently been enriched by the economics literature on team-directed reasoning (see Gold and Sugden

2007, Hakli et al. 2010).[1] On the one hand, various philosophical accounts rely on members to form a *personal intention* of sorts.[2] On the other hand, it has been proposed that the economic literature on team-directed reasoning (Bacharach 1999; 2006, Sugden 1993; 2000; 2003) can be applied to gain a better understanding of collective intentionality (Gold and Sugden 2007, Hakli et al. 2010). Instead of relying on members' personal intentions, the latter approach appeals to the *reasoning process* leading up to the formation of the intention and argues that this focus is needed to grasp collective intentionality. In short, philosophers focus on the intentions of the members, while economists focus on their reasoning process. We discuss these accounts in reverse order.

To explain team-directed reasoning and to contrast it with traditional game and decision theory, let us consider the Hi-Lo game depicted in Figure 1. It seems that (*high*, *high*) is the only rational solution, but if the players in this game are guided by traditional game and decision theory, they have no reason for preferring one action over the other: when applying individualistic dominance reasoning, the row player sees that *high* is a best response to *high* and *low* is a best response to *low*. At best, individual dominance reasoning gives her conditional recommendations: to, for instance, choose *low* if she expects the other player to choose *low*. At worst, it gives her an indeterminate unconditional recommendation. Similar objections apply to expected utility theory. Instead, one may want to apply Nash equilibrium reasoning: it follows from (*high*, *high*) and (*low*, *low*) being the only pure Nash equilibria that Nash equilibrium reasoning also fails to give a determinate recommendation. As a response to this, 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 equilibrium (*high*, *high*) as the only rational solution.[3] In any case, such a 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.[4] This inadequate response to the Hi-Lo game by traditional game and decision theory stands to be corrected, which is what team-directed reasoning (as studied by Bacharach, Sugden, and Gold) has been designed for.

Bacharach (2006, Ch. 1) and Sugden (2000, Sections 2, 3, 7 and 8) argue that traditional game and decision theory needs to be augmented with a collectivis-

[1] Although Bacharach (2006), Gold and Sugden (2007) call their approach 'team reasoning', we follow Sugden (2000) and refer to it as 'team-directed reasoning' to emphasize that this is a mode of reasoning of an *individual* agent, not of a team.

[2] Philosophical accounts of collective intentionality build on participatory intentions (Kutz 2000), contributory intentions (discussed and rejected by Gilbert (2009)), intentions that we J (Bratman 2014), and we-intentions (Tuomela 2000; 2005, Searle 1990).

[3] Harsanyi and Selten (1988) argue for Pareto dominance as a principle of equilibrium selection.

[4] Sugden (2000, pp. 179–182) and Bacharach (2006, pp. 35–68) provide more elaborate treatments of these objections to traditional game and decision theory.

	high	*low*
high	(2,2)	(0,0)
low	(0,0)	(1,1)

Figure 1: The Hi-Lo game.

tic reasoning method to successfully address the Hi-Lo game. Team-directed reasoning appeals to the *reasoning process* by which an individual agent reasons about what to do. An individual agent engaged in team-directed reasoning "works out the best feasible combination of actions for all the members of her team, then does her part in it" (Bacharach 2006, p. 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*.[5] Similar reasoning prescribes *high* for the column player. Team-directed reasoning hence entails that *high* is the only rational option, and in turn selects (*high*, *high*) as the only rational outcome. Problem solved.

Philosophical theories of collective intentionality typically rely on the members to form personal intentions of sorts. We focus on Tuomela's philosophical theory of sociality (Tuomela 2000; 2005; 2006; 2007), which relies on the distinction between the I-mode and the we-mode, which encompasses reasoning, acting, and possibly more: "The we-mode involves functioning as a group member and not as a private person while the I-mode is concerned only with functioning as a private person" (Tuomela 2006, p. 49). Following Hakli et al. (2010, pp. 315–318), it is useful to divide the we-mode reasoning process up into three stages: the first results in the formation of "a group preference matrix", the second reaches "a joint intention to act", and in the third "the agents select their part-actions".[6] We focus solely on the third stage of the we-mode reasoning process.

Our study is thus cast against the background of collective intentional action and focuses on the members' personal intentions, which we call *we-intentions* because they relate to a collective intention. It is important to note that a we-

[5] Team reasoning is generally taken to presuppose a group preference (see Bacharach 1999, Sugden 2000). In the team-reasoning literature there has been some discussion on the relation between the group preference and the individual preferences. For example, Bacharach "allowed in principle that the group objective might be welfare decreasing for some members" (Gold 2012, p. 195) and according to Sugden (2000, p. 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." Nonetheless, for the current Hi-Lo game, it is uncontroversial that the group prefers (*high*, *high*) over (*low*, *low*).

[6] Hakli et al. (2010, p. 318) extensively discuss these three stages and argue that the result of the first stage is common knowledge, but the latter two "can be performed by the individual agents autonomously as generally supposed in non-cooperative game theory".

intention can be held in the I-mode or in the we-mode. For our purposes, the notions of *pro-group I-mode* and *we-mode* in decision making are essential.

> The pro-group I-mode is concerned with promoting the group's interests. (Hakli et al. 2010, p. 296)

That is, a pro-group I-mode agent transforms her preferences and adopts the group's objectives as if they were her own personal objectives. Then she decides what to do by way of individualistic reasoning. In such a case, we say that she forms a *pro-group intention*, meaning that she intends to further the group's objective.[7] Note that adopting a pro-group intention requires her to take the perspective of the group agent to determine the group's objective.[8] This is the first I-mode we-intention type we consider.

> We-mode reasoning and Bacharach's team reasoning yield the same action recommendations in game-theoretic choice situations. (Hakli et al. 2010, p. 301)

We show that the results of team-directed reasoning can be explained by I-mode we-intentions with a particular content, which we call *team-directed intentions*: an individual agent having such a team-directed intention intends to perform an individual action that is compatible with a best group action. An agent adopting a team-directed intention transforms her preferences and adopts as her personal objective performing an individual action that is compatible with a best group action. This requires her to put herself in the shoes of the group agent to determine the best group actions. Then she decides what to do by way of individualistic reasoning. We show that team-directed reasoning and team-directed intentions yield the same action recommendations.[9] This means that

[7] Compare Bacharach's (1999, p. 128 – notation adapted) notion of a "group benefactor": "Let us call a type of player who reasons individualistically but whose payoff function coincides with that of a team $\mathcal{G}$ a *benefactor* of $\mathcal{G}$." Hakli et al. (2010, p. 301) argue that "pro-group I-mode reasoning, in cases in which agents adopt the group preferences, and Bacharach's reasoning as a team benefactor yield the same action recommendations." So, a pro-group I-mode reasoner, a group benefactor, and a pro-group intention adopter yield the same action recommendations.

[8] Hakli et al. (2010, Section 3.2) describe the formation of group preferences by means of we-mode reasoning. It thus seems possible that a pro-group I-mode reasoner partially reasons in the we-mode, though not till the very end. To be more precise, a pro-group I-mode reasoner only follows the first stage of we-mode reasoning.

[9] When assuming the natural candidate for the group preference in the Hi-Lo game, the theory of team reasoning solves the Hi-Lo game letting the agents adopt a particular reasoning method, our theory of team intentions, instead, can be interpreted as solving the Hi-Lo game by letting the agents adopt a particular intention.

an important part of the distinguished we-mode reasoning can be reduced to I-mode reasoning with a team-directed intention.[10]

Our characterization of team-directed reasoning by way of team-directed intentions reveals that it is surprisingly weak. Indeed, a team-directed intention requires an individual agent to perform an individual action that is *only compatible* with a best group action. Reform is required. Therefore, we introduce a third I-mode we-intention type: *participatory intentions*, which require an individual agent to *promote* the realization of a best group action.[11] That is, she adopts as her objective that a best group action is performed.[12] As with team-directed intentions, the agent is required to take the group perspective to determine the best group actions. Then she decides what to do by way of individualistic reasoning.

In this paper we investigate the classes of games for which these three (I-mode) we-intentions – pro-group, team-directed, and participatory intentions – guarantee successful cooperation. The action recommendations resulting from these we-intention types are derived by means of individualistic reasoning: we adopt the intuition that an agent should avoid inadmissible, i.e. dominated, actions. Our results, depicted in Figure 2, are two-fold: (1) Pro-group intentions and team-directed intentions are on a par: in some scenarios team-directed intentions succeed in guaranteeing that a cooperatively rational solution is selected whereas pro-group intentions fail to do so, and vice versa. The class of games for which team-directed intentions guarantee successful cooperation is hence not a proper superset of the corresponding class for pro-group intentions, and vice versa. (2) Participatory intentions surpass both pro-group intentions and team-directed intentions: in any scenario, if either team-directed or pro-group intentions guarantee that a best group action is performed, then so do participatory intentions. Or, equivalently, the class of games for which participatory intentions guarantee successful cooperation is a superset of both the corresponding class for pro-group intentions and the corresponding class for team-directed intentions. Our theory of participatory intentions is therefore the prevalent account of cooperation.

[10] This opposes Hakli et al. (2010, Thesis (5) on p. 307): "We-mode reasoning is not reducible to pro-group I-mode reasoning, i.e. it is not definable by or functionally constructable from I-mode reasoning." These opposing results derive from the fact that Hakli et al. (2010) assume that I-mode reasoning is connected to equilibrium-based reasoning, whereas we assume that I-mode reasoning is connected to dominance reasoning.

[11] Our use of the term 'participatory intentions' is inspired by Kutz (2000), although we do not study the relation between our notions. Kutz writes: "On the one hand, collective activity is an ineliminable part of the content of agents' participatory intentions. [. . .] On the other hand, participatory intentions are simply a special class of ordinary intentions, differentiated by their group-oriented content" (pp. 85–86). He thus has in mind a kind of I-mode we-intention.

[12] Note the minor, yet crucial difference: pro-group intentions require the individual agent to adopt the *group's preferences* as her own, whereas a participatory intention can be viewed as requiring the individual agent to adopt as her personal objective that a *best group action is performed*.

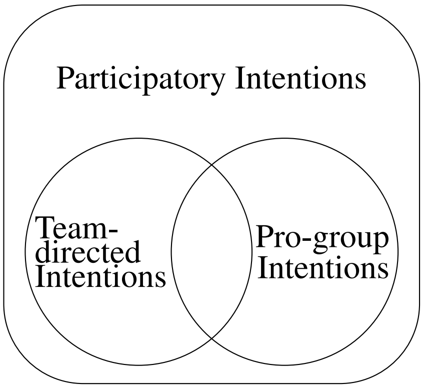

Figure 2: Selecting cooperatively rational outcomes: each of the three areas depicts the class of games for which the respective intention type guarantees successful cooperation.

To achieve all of this, we present a formal analysis. In Section 2 we introduce game-theoretic models and the admissibility requirement, which captures the intuition that one should avoid dominated actions. We then extend traditional game forms with intentions. Based on these models, we introduce a modal logic of agency and intentionality in Section 3. This language includes group-STIT[13] operators $[\mathcal{H}\ \texttt{stit}]\varphi$ expressing that 'group $\mathcal{H}$ sees to it that φ', operators $[\mathcal{H}\ \texttt{prom}]\varphi$ expressing that 'group $\mathcal{H}$ promotes φ', and intention operators $[\mathcal{H}\ \texttt{int}]\varphi$ expressing that 'group $\mathcal{H}$ intends to φ'. In Section 4 we use this language to accurately characterize team-directed intentions, and show that they yield the same action recommendations as team-directed reasoning. In Section 5 we show that, in the context of group $\mathcal{G}$'s collective intention to φ, i.e. $[\mathcal{G}\ \texttt{int}]\varphi$, this language allows us to accurately distinguish between the three previously mentioned we-intention types: pro-group intentions $[i\ \texttt{int}]\varphi$, team-directed intentions $[i\ \texttt{int}]\langle i\ \texttt{stit}\rangle[\mathcal{G}\ \texttt{prom}]\varphi$, and participatory intentions $[i\ \texttt{int}][\mathcal{G}\ \texttt{prom}]\varphi$. In Section 6 we raise our objection to team-directed intentions: we provide a game in which pro-group intentions succeed in guaranteeing successful cooperation whereas team-directed intentions fail to do so. More importantly, in Section 7 we establish that the class of games for which participatory intentions guarantee successful cooperation is a superset of both the corresponding class for pro-group intentions and the corresponding class for team-directed intentions. Our theory of participatory intentions therefore surpasses both team-directed reasoning and pro-group intentions in explaining and predicting cooperative behaviour. Finally, we conclude with a discussion in Section 8.

[13] The acronym STIT stands for 'seeing to it that'. The Chellas STIT operator that we use was introduced by Chellas (1992).

2 Games, intentions, and admissibility

We use strategic games to study and highlight the differences between various we-intention types and team-directed reasoning. A strategic game describes a strategic scenario involving a finite set N of individual agents. Each individual agent i is assigned a finite set of available actions A_i. The Cartesian product $\prod_{i \in N} A_i$ of all individual agents' sets of available actions gives the full set A of action profiles.

Definition 1 (Strategic game form)**.** *A strategic game form S is a tuple $\langle N, (A_i) \rangle$, where N is a finite set of individual agents and for each agent i in N it holds that A_i is a non-empty and finite set of actions available to agent i. The set of action profiles in S is $A = \prod_{i \in N} A_i$.*[14]

For each group $\mathcal{G} \subseteq N$ the set of available group actions, denoted by $A_{\mathcal{G}}$, is the Cartesian product $\prod_{i \in \mathcal{G}} A_i$ of the available individual group members' actions. We use $a_{\mathcal{G}}$ and $a'_{\mathcal{G}}$ as variables for group actions in the set $A_{\mathcal{G}}$, and omit subscripts for action profiles from A. Given a group action $a_{\mathcal{G}}$ and a subgroup $\mathcal{F} \subseteq \mathcal{G}$, we let $a_{\mathcal{F}}$ denote the subgroup action that is $\mathcal{F}$'s component subgroup action of the group action $a_{\mathcal{G}}$. Conclusively, we use $-\mathcal{G}$ to denote the relative complement $N - \mathcal{G}$.

To study various types of we-intentions, we need to extend these strategic game forms to include intentions. Though philosophers have studied various guises of intentions, we restrict our attention to future-directed intentions as studied in the planning theory of intentions advanced by Bratman (1987).[15] There are two different types of future-directed intentions: I can intend to perform a certain action, or I can intend to realize a certain state of affairs. We focus primarily on intentions to realize a certain state of affairs. We assume that an action profile fully determines the future state of the world.[16] An intention is then identified with a set of action profiles. Intuitively, an intention $J \subseteq A$ is an intention to realize the aspects that all outcomes of the action profiles in J have in common. We restrict to agents having just one single intention. So the intention of an – individual or collective – agent $\mathcal{H}$ is given by a set of action

[14] Note the absence of payoffs or utilities. In this paper we focus on the agents' intentions, which will be added in Definition 2. In game-theoretic terms such a strategic game form is a *game form*. A *game*, in contrast, also includes a vector of numerical payoffs, one for each individual agent.

[15] These future-directed intentions (such as my intention to submit this paper by the end of the month) have been distinguished from intentions in action (such as my typing with the intention to finish this introduction) and intentional acts (such as my typing these words intentionally) (see Anscombe (1963)).

[16] In this regard, our framework is very similar to that of van Hees and Roy (2008). Alternatively, if one wants to retain indeterminism, an agent called 'nature' can be added to model the indeterminacy. That is, once every agent has made her choice, the exact outcome is determined by nature's move.

profiles $\mathsf{Int}_{\mathcal{H}}$. This induces the reading that an agent $\mathcal{H}$ intends to φ if and only if her intention $\mathsf{Int}_{\mathcal{H}}$ is represented by φ. To model that rational intentions are feasible, we require that $\mathsf{Int}_{\mathcal{H}} \neq \emptyset$.

Definition 2 (Strategic game with intentions)**.** *A* strategic game with intentions *(SGI) S is a tuple $\langle N, (A_i), (\mathsf{Int}_{\mathcal{H}})\rangle$ where $\langle N, (A_i)\rangle$ is a strategic game form, and, for every group of agents $\mathcal{H}$, $\mathsf{Int}_{\mathcal{H}}$ is a non-empty set of profiles from A (so $\mathsf{Int}_{\mathcal{H}} \subseteq A \setminus \emptyset$). In picturing the intentions, we often use utilities $(u_{\mathcal{H}})$ where*

$$u_{\mathcal{H}}(a) = \begin{cases} 1, \textit{ if } a \in \mathsf{Int}_{\mathcal{H}} \\ 0, \textit{ otherwise.}^{17} \end{cases}$$

To interpret our logical language that will be introduced in Section 3, an SGI is extended to a strategic model with intentions *(SMI) by adding a propositional valuation $\pi : A \to 2^{\mathcal{P}}$ (where $\mathcal{P}$ is the set of propositional letters).*

Intentions provide a "*filter of admissibility* for options" (Bratman 1987, p. 33, emphasis in original). Although Bratman does not use this term in any decision-theoretic sense, we accept the intuition that an agent intending to φ is required to avoid inadmissible actions, i.e. avoid dominated actions.[18] Admissibility captures the idea that an agent takes all actions for the other agents into consideration; none is entirely ruled out.[19] Avoiding inadmissible actions is more restrictive than avoiding *strictly* dominated actions. Admissibility has a long tradition in decision theory (see the discussion in Kohlberg and Mertens (1986, Section 2.7)). Therefore we take 'providing a filter of admissibility' to mean that our intentions require us to choose an admissible action, i.e. one that is not dominated.[20]

[17] Although our conceptual analysis involves intentions, a decision theorist can view our formal analysis as being restricted to only binary utility functions.

[18] In their axiomatic approach to decision theory, Luce and Raiffa (1957, see Section 13.3, and p. 306) express the admissibility requirement in Axiom 5 and write: "Axioms 1 through 5, and to a lesser extent 6, seem quite innocuous and, so far as we are aware, all serious proposals for criteria satisfy them."

[19] Selten (1975) argues that even rational players, having made their choice, may with non-zero probability do something else by accident. In addition, Pearce (1984, Lemma 4) shows that an action is admissible if and only if it maximizes expected utility with respect to a probability function that assigns positive probability to every move of the opponent. This means that an expected utility maximizer should avoid inadmissible actions.

[20] Note that we do not employ iterated admissibility, i.e. iterated deletion of dominated actions (see the discussion in Kohlberg and Mertens (1986, Section 2.7) and the discussion of epistemic characterizations of iterated admissibility in Brandenburger et al. (2008, Section 2.6)). One of the main reasons for refraining from doing so is that iterated admissibility is subject to some paradoxes: for instance, it is well known that the order in which dominated strategies are eliminated can affect the outcome of the process. Furthermore, we believe the core results (Results 1 through 4) of this paper are sustained if iterated admissibility is taken to entail that inadmissible actions are avoided.

In decision theory it is common to derive this dominance ordering from exogenously given utilities. Here, however, we are interested in the dominance orderings resulting from the endogenously adopted intentions. We therefore submit that a dominance ordering is relative to a certain intention. A group action $a_{\mathcal{G}}$ is admissible with respect to an intention to J if and only if no other group action promotes the realization of J, regardless of what the group's non-members do, better than $a_{\mathcal{G}}$ does. So we require that an agent intending to φ should also promote φ, that is, should perform an action that is admissible with respect to φ.

Since we only consider intentions to realize a state of affairs, the adopted dominance ordering is relative to the realization of some state of affairs, instead of maximizing payoffs, as is usual in traditional game and decision theory.[21] Our dominance ordering is therefore explicitly relative to a certain state of affairs. Moreover, note that altering the state of affairs impacts the resulting dominance ordering. In decision-theoretic terms, I may want to optimize my own happiness, but I could instead aim at optimizing our collective, perhaps average or minimum, happiness; pursuing these different states of affairs may result in different dominance orderings. Still, the principle by which the dominance ordering results from a certain state of affairs is uniform.

The principle guiding the dominance orderings combines the sure-thing principle and reasoning by cases.[22] More specifically, a group action $a_{\mathcal{G}}$ weakly dominates $a'_{\mathcal{G}}$ with respect to a certain state of affairs if and only if $a_{\mathcal{G}}$ promotes realizing that state of affairs at least as well as $a'_{\mathcal{G}}$, regardless of what the group's non-members do:

Definition 3 (Admissibility). *Let $S = \langle N, (A_i)\rangle$ be a strategic game form. Let $J \subseteq A$ be a state of affairs. A group action $a_{\mathcal{G}}$* weakly dominates *$a'_{\mathcal{G}}$* with respect to *J, notation $a_{\mathcal{G}} \succeq_J a'_{\mathcal{G}}$, if and only if for all actions $a''_{-\mathcal{G}} \in A_{-\mathcal{G}}$ we have that $(a'_{\mathcal{G}}, a''_{-\mathcal{G}}) \in J$ implies $(a_{\mathcal{G}}, a''_{-\mathcal{G}}) \in J$.* Dominance*, notation $\succ_J$, is derived from weak dominance: $a_{\mathcal{G}} \succ_J a'_{\mathcal{G}}$ if and only if $a_{\mathcal{G}} \succeq_J a'_{\mathcal{G}}$ and $a_{\mathcal{G}} \not\preceq_J a'_{\mathcal{G}}$. A group action $a_{\mathcal{G}}$* is admissible with respect to *J if and only if it is not dominated with respect to J by any group action in $A_{\mathcal{G}}$.*

When we represent the intention by a binary utility function rather than a set of possible worlds, this definition translates to: $a_{\mathcal{G}} \succeq_J a'_{\mathcal{G}}$ if and only if for all

[21] However, note that, although we refrain from modelling beliefs and degrees of beliefs, these are intimately related as we interchangeably model the intentions by using a set of possible worlds and a utility function (see Definition 2).

[22] Savage (1954, p. 21) writes: "I know of no other extralogical principle governing decisions that finds such ready acceptance." Our personal inspiration is from Horty (1996; 2001), who provided a similar analysis in deontic logic, which is the formal study of obligations and permissions, by introducing "an ordering on actions available to the agent through a state-by-state comparison of their results", where "we will identify the states confronting the agent at any given moment with the possible patterns of actions that might be performed at that moment by all other agents" (Horty 2001, p. 67 and p. 66).

			j	
		t_1	t_2	t_3
	s_1	(1,1,1)	(1,0,0)	(0,1,1)
i	s_2	(1,1,1)	(0,0,0)	(0,0,0)
	s_3	(1,0,0)	(1,1,0)	(0,0,0)

Figure 3: Strategic Game with Intentions S_2, where the intentions are represented by a triple of utilities expressing i's, j's, and $\{i, j\}$'s intention.

$a''_{-\mathcal{G}} \in A_{-\mathcal{G}}$ we have $u_J(a'_{\mathcal{G}}, a''_{-\mathcal{G}}) \geq u_J(a_{\mathcal{G}}, a''_{-\mathcal{G}})$ (where $u_J(a) = 1$ if $a \in J$ and 0 otherwise). This is thus equivalent to the standard weak dominance ordering, as studied in traditional game and decision theory.

Note that this definition implies that for every group of agents and every state of affairs there is at least one admissible group action with respect to that state of affairs (because A is assumed to be finite).

To illustrate this dominance ordering and admissibility, consider game S_2 in Figure 3. First, note that this example illustrates that we do not presuppose any connection between personal and group intentions. Secondly, regarding the dominance ordering, since agent i's intention is not realized at (s_2, t_2), we can see that for agent i only actions s_1 and s_3 are admissible with respect to her intention. For agent j, action t_3 is dominated by t_1 with respect to her intention, while t_1 and t_2 are incomparable and, hence, both admissible. That is, $t_1 \succ_{\mathsf{Int}_j} t_3$, $t_1 \not\succeq_{\mathsf{Int}_j} t_2$, and $t_2 \not\succeq_{\mathsf{Int}_j} t_1$. For the group $\mathcal{G}$, consisting of agents i and j, the admissible group actions with respect to its collective intention $\mathsf{Int}_{\mathcal{G}}$ are (s_1, t_1), (s_2, t_1), and (s_1, t_3). Since we accept the intuition that an agent intending to φ should avoid inadmissible actions, this means that agent i can choose s_1 and agent j can choose t_2, with regard to their respective individual intention. It follows that the individual intentions allow (s_1, t_2), which is inadmissible with respect to the collective intention $\mathsf{Int}_{\mathcal{G}}$. The individual intentions hence do not guarantee that a best *group action* is performed, which is not a surprise.

3 Modal logic of agency and intentionality

To highlight the conceptual differences between the three we-intention types discussed in the introduction, in this section we introduce a modal-logical language. This logical language is an extension of group-STIT logics:

Definition 4 (Syntax). *Given a set of propositional letters $\mathcal{P}$ and a set of agents N, the syntax is given by*

$$\varphi ::= p \mid \varphi \wedge \varphi \mid \neg\varphi \mid \Box\varphi \mid [\mathcal{H}\ \texttt{stit}]\varphi \mid [\mathcal{H}\ \texttt{int}]\varphi \mid [\mathcal{H}\ \texttt{prom}]\varphi,$$

where p ranges over $\mathcal{P}$ and $\mathcal{H}$ ranges over subsets of N. Standard abbreviations are used for dual versions of the modal operators, for example $\Diamond\varphi \leftrightarrow \neg\Box\neg\varphi$ and $\langle\mathcal{H}\ \texttt{stit}\rangle\varphi \leftrightarrow \neg[\mathcal{H}\ \texttt{stit}]\neg\varphi$.

We briefly discuss intuitive readings of these operators before considering the respective formal semantics and providing a more detailed conceptual discussion. Formulas of this language are interpreted at action profiles. The central agency operator is the modality $[\mathcal{H}\ \texttt{stit}]\varphi$, which stands for 'group $\mathcal{H}$ sees to it that φ holds'. Modalities $[\mathcal{H}\ \texttt{int}]\varphi$ and $[\mathcal{H}\ \texttt{prom}]\varphi$ are interpreted as 'group $\mathcal{H}$ intends to φ' and 'group $\mathcal{H}$ promotes φ', respectively. The modality $\Box\varphi$ is a universal modality and is therefore interpreted as 'φ holds at any action profile'. The models and the dominance ordering, presented in the previous section, are used to provide formal semantics for this logical language by using truth conditions for the syntactic clauses:

Definition 5 (Semantics)**.** *Let $S = \langle N, (A_i), (\mathsf{Int}_\mathcal{H}), \pi\rangle$ be an SMI, and $\mathcal{H} \subseteq N$ be a group of agents. The truth conditions are given by a recursive definition:*

$$\begin{array}{rcl}
S, a \vDash p & \Leftrightarrow & p \in \pi(a) \\
S, a \vDash \varphi \wedge \psi & \Leftrightarrow & S, a \vDash \varphi \textit{ and } S, a \vDash \psi \\
S, a \vDash \Box\varphi & \Leftrightarrow & \textit{every } b \in A \textit{ satisfies } S, b \vDash \varphi \\
S, a \vDash [\mathcal{H}\ \texttt{stit}]\varphi & \Leftrightarrow & \textit{every } b \in A \textit{ with } b_\mathcal{H} = a_\mathcal{H} \textit{ satisfies } S, b \vDash \varphi \\
S, a \vDash [\mathcal{H}\ \texttt{prom}]\varphi & \Leftrightarrow & a_\mathcal{H} \textit{ is admissible with respect to } \{b \mid S, b \vDash \varphi\}^{23} \\
S, a \vDash [\mathcal{H}\ \texttt{int}]\varphi & \Leftrightarrow & \textit{we have } \mathsf{Int}_\mathcal{H} = \{b \mid S, b \vDash \varphi\}.^{24}
\end{array}$$

These semantics implement the idea from STIT theories that 'the agent sees to it that φ' means that the truth of φ is guaranteed by an action or choice of the agent. When Ann empties her glass of milk, the nature of her action on this view is to constrain the possible worlds to those where the glass of milk is emptied. Hence, an action a_i is identified with a subset of the possible worlds, namely those action profiles b satisfying $b_i = a_i$. This induces the reading that an agent sees to it that φ only if she performs an action a_i, thereby constraining the possible worlds to only φ-worlds.

It may be useful to add that the STIT modality $[\mathcal{H}\ \texttt{stit}]\varphi$ can be interpreted, relative to a profile a, as 'group $\mathcal{H}$ guarantees that φ holds regardless of what

[23] A technical remark: the semantics for the $[\mathcal{H}\ \texttt{prom}]$ operator could also have been given by neighbourhood semantics, where the neighbourhoods for a group $\mathcal{H}$ of a profile a would be given by the collection $\{M \subseteq A \mid a_\mathcal{H}$ is admissible with respect to $M\}$.

[24] Note the equality sign, which means that we employ neighbourhood semantics. If we had employed the standard possible worlds semantics, it would have read '$\mathsf{Int}_\mathcal{H}$ consists only of φ-worlds', meaning $\mathsf{Int}_\mathcal{H} \subseteq \{a \in A \mid a$ satisfies $\varphi\}$. Konolige and Pollack (1993) were the first to model intentions using neighbourhood semantics instead of the standard possible worlds semantics.

the others do'. Indeed, the truth condition of $S, a \vDash [\mathcal{H}\ \texttt{stit}]\varphi$ is equivalent to requiring that for every $b_{-\mathcal{H}} \in A_{-\mathcal{H}}$ we have $S, (a_{\mathcal{H}}, b_{-\mathcal{H}}) \vDash \varphi$. This reveals that the group has enough control to ensure φ.

The dual STIT modality $\langle\mathcal{H}\ \texttt{stit}\rangle\varphi$ expresses that 'group $\mathcal{H}$ allows for φ', that is, group $\mathcal{H}$'s action does not rule out φ. At a profile $a \in A$, this means that there is an action profile $b_{-\mathcal{H}} \in A_{-\mathcal{H}}$ for the others such that $S, (a_{\mathcal{H}}, b_{-\mathcal{H}}) \vDash \varphi$. Or, equivalently, action $a_{\mathcal{H}}$ is compatible with φ.

The modality $[\mathcal{H}\ \texttt{prom}]\varphi$ expresses that the group $\mathcal{H}$ performs a best group action with respect to realizing φ, where 'best' is equated with 'admissible'. The underlying intuition is that there may be multiple best actions, each of which is admissible. So, this operator expresses that group $\mathcal{H}$ avoids actions that are inadmissible with respect to φ. This operator hence refers to admissibility, as introduced in Definition 3.[25] Our intuition that an agent intending to φ is required to avoid actions that are inadmissible with respect to φ, thus means that an agent intending to φ is required to promote φ.

The semantics of the intention operator emphasize that we only consider intentions to realize a certain state of affairs, which is represented by a set of worlds. The truth condition for $[\mathcal{H}\ \texttt{int}]\varphi$ employs neighbourhood semantics and thus induces the reading that the group's intention is represented by φ. These neighbourhood semantics are usually employed to avoid problems concerning logical omniscience, that is, to avoid intentions being closed under logical implications.[26] Our motivation is different: we accepted the intuition that an agent intending to φ is required to avoid inadmissible actions. But if $\vDash \varphi \to \psi$, then an action that is admissible with respect to φ need not be admissible with respect to ψ.[27] So if the agent's intention is closed under logical implications, this will result in a practical dilemma: she cannot perform an action that is both admissible with regard to her initial intention and admissible with regard to all its logical consequences.

Since our aim is to contribute to a conceptual analysis of collective agency, and team-directed reasoning in particular, a complete logical investigation is well beyond our current ambition. Still, it is useful for our current purposes to examine some logical properties of the multi-modal logic we have introduced:

[25] A similar operator has been used by Broersen (2011) to model attempts. Whereas he uses maximizing expected utility, we adopted admissibility as the underlying decision principle. However, although there might be some connections to attempts, here we do not want to argue that our $[\mathcal{H}\ \texttt{prom}]$-operator adequately models attempts.

[26] For this reason, Konolige and Pollack (1993, p. 178) use neighbourhood semantics to model intentions.

[27] To see this, reconsider the SGI S_2 in Figure 3: We have already noticed that s_3 is admissible with respect to Int_i. However, s_3 fails to be admissible with respect to $\mathsf{Int}_i \cup \{(s_1, t_3)\}$, since s_1 dominates s_3 with respect to $\mathsf{Int}_i \cup \{(s_1, t_3)\}$.

Proposition 1. *Let φ be any formula in our language, and let $\mathcal{H}$ be a group of agents, possibly a singleton. Then*

1. $\vDash \Diamond[\mathcal{H}\ \texttt{prom}]\varphi$*, one is always able to promote* φ*,*
2. $\nvDash [\mathcal{H}\ \texttt{prom}]\varphi \to [\mathcal{H}\ \texttt{stit}]\varphi$*, promoting* φ *does not entail ensuring* φ*,*
3. $\vDash [\mathcal{H}\ \texttt{stit}]\varphi \to [\mathcal{H}\ \texttt{prom}]\varphi$*, guaranteeing* φ *entails promoting* φ*,*
4. $\vDash [\mathcal{H}\ \texttt{prom}]\varphi \wedge \Diamond\varphi \to \langle\mathcal{H}\ \texttt{stit}\rangle\varphi$*, promoting* φ *while* φ *is possible entails allowing* φ*,*
5. $\vDash [\mathcal{H}\ \texttt{prom}]\varphi \wedge \Diamond[\mathcal{H}\ \texttt{stit}]\varphi \to [\mathcal{H}\ \texttt{stit}]\varphi$*, promoting* φ *while being able to ensure* φ *entails guaranteeing* φ*,*
6. $\vDash \Diamond[\mathcal{H}\ \texttt{stit}]\varphi \to ([\mathcal{H}\ \texttt{prom}]\varphi \leftrightarrow [\mathcal{H}\ \texttt{stit}]\varphi)$*, if one is able to guarantee* φ*, then promoting* φ *is equivalent to ensuring* φ

Proof. Let $S = \langle N, (A_i), (\mathsf{Int}_{\mathcal{H}}), \pi\rangle$ be any SMI.
1. Follows from the fact that we only consider finite strategic games.
2. The previous item shows that if this were a validity, then $\vDash \Diamond[i\ \texttt{stit}]\varphi$ would follow. In other words, one is always able to ensure φ, no matter its logical form. This is, however, not always the case.
3. Suppose $S, a \vDash [\mathcal{H}\ \texttt{stit}]\varphi$. Then for any $c_{-\mathcal{H}} \in A_{-\mathcal{H}}$ we have $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$. Hence for any $b \in A$ and any $c_{-\mathcal{H}} \in A_{-\mathcal{H}}$ we have that $S, (b_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$ implies $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$. In other words, $a_{\mathcal{H}} \succeq_{\varphi} b_{\mathcal{H}}$, so $a_{\mathcal{H}}$ is admissible with respect to φ.
4. Suppose $S, a \vDash [\mathcal{H}\ \texttt{prom}]\varphi \wedge \Diamond\varphi$. Then there is a profile, say b, such that $S, b \vDash \varphi$. We argue by contradiction that $S, a \vDash \langle\mathcal{H}\ \texttt{stit}\rangle\varphi$: suppose $S, a \nvDash \langle\mathcal{H}\ \texttt{stit}\rangle\varphi$. Then for any $c_{-\mathcal{H}} \in A_{-\mathcal{H}}$ we have $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \nvDash \varphi$. Hence, for any $c_{-\mathcal{H}} \in A_{-\mathcal{H}}$ we have that $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$ implies $S, (b_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$, vacuously, i.e. $b_{\mathcal{H}} \succeq_{\varphi} a_{\mathcal{H}}$. Moreover, since $S, b \vDash \varphi$, we have $b_{\mathcal{H}} \not\preceq_{\varphi} a_{\mathcal{H}}$ and therefore $b_{\mathcal{H}} \succ_{\varphi} a_{\mathcal{H}}$. This shows that $a_{\mathcal{H}}$ is not admissible with respect to φ, which contradicts with our assumption that $S, a \vDash [\mathcal{H}\ \texttt{prom}]\varphi$.
5. We argue by contradiction. Suppose $S, a \vDash [\mathcal{H}\ \texttt{prom}]\varphi$ and $S, b \vDash [\mathcal{H}\ \texttt{stit}]\varphi$, yet $S, a \nvDash [\mathcal{H}\ \texttt{stit}]\varphi$. Recall from the proof of item 3 that $S, b \vDash [\mathcal{H}\ \texttt{stit}]\varphi$ entails that $b_{\mathcal{H}}$ weakly dominates every group action. In particular, $a_{\mathcal{H}} \preceq_{\varphi} b_{\mathcal{H}}$. Our assumption that $S, a \nvDash [\mathcal{H}\ \texttt{stit}]\varphi$ gives a $c_{-\mathcal{H}} \in A_{-\mathcal{H}}$ such that $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \neg\varphi$. In particular, $S, (b_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$ implies $S, (a_{\mathcal{H}}, c_{-\mathcal{H}}) \vDash \varphi$ does not hold. Hence $a_{\mathcal{H}} \prec_{\varphi} b_{\mathcal{H}}$, which contradicts $S, a \vDash [\mathcal{H}\ \texttt{prom}]\varphi$.
6. Follows immediately from the previous item and item 3

□

Item **1** establishes that one is always able to promote φ, irrespective of its logical form. The fact that this includes infeasible properties, in particular logical inconsistencies, may seem unsatisfactory; however, for infeasible φ it does not matter what one does, because one's choice of action does not change the fact that φ will not be realized. Items **2** and **3** show that guaranteeing φ is logically stronger than promoting φ. Item **4** expresses that if φ is feasible, promoting φ entails that one performs an action that is compatible with φ. Or, equivalently, if one performs an action that is incompatible with φ, then one is surely not promoting φ. Items **5** and **6** prove that, although promoting φ is logically weaker than guaranteeing φ, if one is able to ensure φ, then promoting φ is equivalent to guaranteeing φ. This shows that one is definitely devoted to realizing φ if one promotes φ.

We refrain from a logical analysis of the intention operator, because our primary interest is in how intentions constrain the choice of strategy of the respective agents. Most importantly, as mentioned before, we accept that an agent intending to φ is required to avoid actions that are inadmissible with respect to φ. That is, an agent intending to φ is required to promote φ. Individual intentions hence guarantee that individual agents choose actions that are admissible with regard to their respective intentions. To generalize, intentions guarantee a certain property ψ only if it is the case that whenever the agents choose actions that are admissible with respect to their intentions, then ψ holds. It is natural to interpret this as a counterfactual: if agent i intends to φ, and therefore performs an individual action that is admissible with respect to φ, then ψ will hold. That is, whenever the agent promotes φ then ψ will hold.

Definition 6. *Let S be a SMI, let $a \in A$ be a profile, let φ and ψ be formulas in our logical language, and let i be an individual agent. Then we say that, in S, $[i\ \mathtt{int}]\varphi$ guarantees ψ if and only if $S, a \vDash \Box([i\ \mathtt{prom}]\varphi \rightarrow \psi)$. (Note that we do not assume $S, a \vDash [i\ \mathtt{int}]\varphi$.)*

More generally, let $\mathcal{G}$ be a group of agents and let ψ and φ_i be formulas in our language, one for each $i \in \mathcal{G}$. Then we say that, in S, these intentions guarantee ψ if and only if $S, a \vDash \Box(\bigwedge_{i \in \mathcal{G}}[i\ \mathtt{prom}]\varphi_i \rightarrow \psi)$.

The latter can be semantically interpreted as follows: let P be any property, which need not be expressible in our logical language. Then we say that, in S, the intentions $\{J_i\}_{i \in \mathcal{G}}$ guarantee P if and only if for every $b \in A$ such that b_i is admissible with respect to J_i for every $i \in \mathcal{G}$, the property P holds at b.

Our focus, later in the paper, is going to be on whether intentions guarantee that a best group action is performed. Since the three we-intention types – pro-group, team-directed, and participatory intentions (see Section 5) – are introduced to further cooperation, they are compared with respect to whether they guarantee that a best group action is performed (Section 7). This includes, for example, a study of whether pro-group intentions guarantee that a best group action is performed only if team-directed intentions do. (The answer turns out to be negative, leading us to the conclusion that team-directed intentions and pro-group intentions are on an equal footing.)

4 A reduction of team-directed reasoning to intentions

As a motivation for the three I-mode we-intention types that are to be introduced in Section 5, we discuss here the reduction of an important part of we-mode reasoning to I-mode reasoning with a particular we-intention. That is, we show that team-directed reasoning and this I-mode we-intention yield the same action recommendations.

We henceforth presuppose a collective intention to φ and investigate what team-directed reasoning and the we-intention types amount to.[28] Just as an individual agent's intention requires her to perform an individual action that is admissible with respect to that individual intention, a collective intention requires the group to perform a group action that is admissible with respect to its collective intention. Or, equivalently, the collective intention provides a filter of admissibility for the available group actions in which the admissible group actions are best. The group should therefore perform a group action that promotes the realization of what is collectively intended.

To illustrate the benefit of team-directed reasoning in Strategic Games with Intentions, consider the game depicted in Figure 4, which is inspired by the Hi-Lo game. Note that, in the context of group $\mathcal{G}$'s collective intention to φ, the best group action is (*high*, *high*), because it is the only group action that ensures φ. An agent engaged in team-directed reasoning hence first identifies (*high*, *high*) as the unique best group action and then decides to perform her part in that combination, therefore recommending *high* to both agent i and agent j.[29] So, in this game, team-directed reasoning of individual agents i and j ensures that group $\mathcal{G}$ performs a best group action.

[28] In our models, we henceforth highlight that the collective intention is presupposed by boldfacing the utility function that represents it.

[29] Sugden (2003, p. 168) writes: "By 'team reasoning, narrowly defined' I mean a mode of reasoning, followed by one individual, which prescribes that he should perform his part of whichever profile is best for the team. This mode of reasoning may be embedded in a larger logic which specifies the conditions under which team reasoning, narrowly defined, should be used. In this paper, I will formulate such a logic, which I will call a logic of 'team reasoning'." Our analysis is thus limited to 'team reasoning, narrowly defined'; we do not study the larger logic.

		j				
		high	*low*		*high*	*low*
i	*high*	φ (1,1,1)	(1,0,0)		φ (1,1,1)	(1,0,0)
	low	(0,1,0)	φ (0,0,1)		(0,1,0)	(0,0,0)
		u_1			u_2	
				k		

Figure 4: The alternative Hi-Lo game S_3: a three-player game, where $\{i, j\} = \mathcal{G}$ collectively intends to φ, showing only the intentions of i, j, and $\mathcal{G}$, respectively.

To show that team-directed reasoning yields the same action recommendations as a certain I-mode we-intention, we introduce *team-directed intentions*, which are I-mode we-intentions with a certain type of content:

Definition 7 (Team-directed intentions)**.** *Suppose group $\mathcal{G}$ collectively intends to φ. Let $i \in \mathcal{G}$ be a member of the group $\mathcal{G}$. Individual agent i's* team-directed intention *is an intention to act compatibly with a best group action, that is,* $[i\ \texttt{int}]\langle i\ \texttt{stit}\rangle[\mathcal{G}\ \texttt{prom}]\varphi$*.*

An agent adopting a team-directed intention transforms her preferences and adopts as her personal objective performing an individual action that is compatible with a best group action. This may require her to put herself in the shoes of the group agent to determine the best group actions. Then she decides what to do by way of individualistic reasoning. That is, she decides what to do by determining the individual actions that are admissible with respect to her team-directed intention.

To illustrate these team-directed intentions, we reconsider the alternative Hi-Lo game S_3 in Figure 4. So far, we have not discussed the individual intentions in this SMI. Note that $[\mathcal{G}\ \texttt{stit}]\varphi$ only holds in (*high*, *high*, u_1) and (*high*, *high*, u_2), so $[\mathcal{G}\ \texttt{prom}]\varphi$ also only holds in those profiles (see Proposition 1 item 6). This entails that $\langle i\ \texttt{stit}\rangle[\mathcal{G}\ \texttt{prom}]\varphi$ is only satisfied in profiles where i's component is *high*, which is exactly those in which i's component in the utility triple equals 1. Hence, in the alternative Hi-Lo game S_3, agent i has the team-directed intention $[i\ \texttt{int}]\langle i\ \texttt{stit}\rangle[\mathcal{G}\ \texttt{prom}]\varphi$. It can be similarly shown that j has the team-directed intention.

An individual agent performs an action that is admissible with respect to her team-directed intention if and only if she performs an individual action that is her component of a best combination of actions that the members can perform. Although the team reasoning literature mostly presupposes that there is a *unique* best group action, we follow Bacharach (1999, p. 120 – adapted notation): "I shall say that agent i team reasons (for $\mathcal{G}$) when she first computes a best profile a^* (in terms of $u_{\mathcal{G}}$), next computes a_i^*, and lastly decides to do a_i^* because this is the component under her control of a best profile." It is unsurprising that actions resulting from team-directed reasoning coincide with actions admissible with respect to the team-directed intention:

Result 1. *Let $S = \langle N, (A_i), (\mathsf{Int}_{\mathcal{H}}), \pi\rangle$ be an SMI. Suppose group $\mathcal{G}$ collectively intends to φ. Let $a \in A$. Then team-directed reasoning admits the individual action a_i if and only if $S, a \vDash \langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi$, which is, in turn, equivalent to saying that the individual action a_i is admitted by the team-directed intention.*

Proof. First note that for any profile $a \in A$ we have that $S, a \vDash [i\ \mathtt{prom}]\langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi$ if and only if $S, a \vDash \langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi$: since, by Proposition 1, $\vDash \Diamond[\mathcal{G}\ \mathtt{prom}]\varphi$, we have $\vDash \Diamond\langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi$. Because $[i\ \mathtt{stit}]$ is an S5 operator, the 5 axiom and T axiom derive $\langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi \leftrightarrow [i\ \mathtt{stit}]\langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi$. Hence, these two validities show that $\vDash \Diamond[i\ \mathtt{stit}]\langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi$. Therefore, again by Proposition 1, $[i\ \mathtt{prom}]\langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi$ is equivalent to $[i\ \mathtt{stit}]\langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi$, which is equivalent to $\langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi$.

Secondly, recall that team-directed reasoning implies that individual agent i first identifies a best combination that the group members can perform, say $a'_{\mathcal{G}}$, and then decides to perform the individual action that is her part in that combination, which is a'_i. Hence, in the context of a collective intention to φ, for any $a \in A$, the following are equivalent: (1) team-directed reasoning admits a_i, (2) there is an $a'_{\mathcal{G}-i} \in A_{\mathcal{G}-i}$ such that $(a_i, a'_{\mathcal{G}-i})$ is a best group action that the group members can perform, (3) there is a $a'_{-i} \in A_{-i}$ such that $S, (a_i, a'_{-i}) \vDash [\mathcal{G}\ \mathtt{prom}]\varphi$, and finally (4) $S, a \vDash \langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi$. □

This result shows that the results of team-directed reasoning can be equally well explained by team-directed intentions. Or, equivalently, it shows that an important part of we-mode reasoning can be reduced to I-mode reasoning with a team-directed intention. So, the effects of the agency transformation in team-directed reasoning can be mirrored by the preference transformation in team-directed intentions.

This connection shows that we need not focus on the *mental processes* by which collective intentions are formed, because it suffices to study the *mental states* of the members.[30] Our result therefore complements the analysis by Gold and Sugden (2007), who argue that team reasoning leads to collective intentions:

[30] Compare Hakli et al. (2010, p. 299): "We study processes of we-reasoning concentrating on the difference between we-mode reasoning and pro-group I-mode reasoning. The difference is not in the aims of the agents because in both cases the agents aim at the benefit of the group. Rather, the difference is in the reasoning process: It is individualistic in the I-mode case."

> Team reasoning results in the formation of intentions. ... references to the group are noneliminable parts of the reasoning process that led to the formation of the intention. Thus, it is natural to regard the intentions that result from team reasoning as collective intentions. (Gold and Sugden 2007, p. 126)

If all of this is correct, then the connection between team-directed intentions and team-directed reasoning, established by our result, reveals that it is equally natural to suppose that team-directed *intentions* are essential for collective intentions. In line with the philosophical literature, this purports a relation between personal and collective intentionality. (Our analysis was the other way around: we started by supposing a collective intention, only then analysing which kind of intentions are supported by team-directed reasoning.)

Moreover, our reduction moves team-directed reasoning into the ballpark of preference transformations and dominance reasoning. Team-directed reasoning is supposed to fill a gap in traditional game and decision theory by adding some kind of collectivistic method of reasoning, but our result shows that this new reasoning method is not needed. After all, dominance reasoning coupled with the preference transformation in team-directed intentions yields the same action recommendations as team-directed reasoning.

5 Three I-mode we-intention types

Which individual attitudes are warranted in the context of a collective intention? In the following, our logical language is used to formalize and study three types of we-intentions. To adequately conceptualize the intricate differences between these we-intention types, they are meticulously represented by particular formulas in our logical language. As we have witnessed in the previous section, such a detailed representation can bring about new conceptual insights and thereby enhance our understanding of the subtleties in we-intentions. Our running example to illustrate these types of we-intentions is displayed in Figure 5.

Suppose group $\mathcal{G}$ $(= \{i, j\})$ collectively intends to φ. A member i of $\mathcal{G}$ can have various we-intentions that our formalism is able to distinguish. She could adopt the collective goal as her own, instead of furthering her personal goals, and pursue it to the best of her abilities, expressed by $[i\ \mathtt{int}]\varphi$ and coined a *pro-group intention*. This way, she ignores the contributions others can make and does her best to realize the collective intention regardless of what others do. Various game-theoretic enterprises try to explain cooperative behaviour by transforming the preferences of the group members. Although we do not study

j

		t_1	t_2	t_3	t_1	t_2	t_3
	s_1	φ (1,1,1)	φ (1,1,1)	(1,0,0)	φ (1,1,1)	φ (1,1,1)	(1,0,0)
i	s_2	φ (1,1,1)	(1,0,0)	(1,0,0)	φ (1,1,1)	(1,0,0)	(1,0,0)
	s_3	(0,0,0)	(0,0,0)	φ (0,0,1)	(0,0,0)	(0,0,0)	(0,0,0)
		u_1			u_2		

k

Figure 5: Strategic game model S_4: a three-player game, where $\{i, j\} = \mathcal{G}$ collectively intends to φ, showing only the intentions of i, j, and $\mathcal{G}$, respectively.

its relation with such preference transformations, the same intuition underlies this first we-intention type:

Definition 8 (Pro-group intentions)**.** *Suppose group $\mathcal{G}$ collectively intends to φ. Let $i \in \mathcal{G}$ be a member of the group $\mathcal{G}$. Agent i's* pro-group intention *is an intention to promote the group's objective, which is to realize φ, that is, $[i\ \texttt{int}]\varphi$.*

In game S_4 of Figure 5, we observe that agents i and j would have the pro-group intention if their intentions were represented by the third component of the triple of intentions. We consider which individual actions would be admissible if the agents were to adopt the pro-group intention: since $S_4, (s_2, t_2, u_1) \nvDash \varphi$ and $S_4, (s_1, t_2, u_1) \vDash \varphi$, we have $s_2 \not\succeq_\varphi s_1$ and $s_1 \succeq_\varphi s_2$. And, because $S_4, (s_3, t_2, u_1) \nvDash \varphi$, we also have $s_3 \not\succeq_\varphi s_1$. Likewise, by comparing (s_1, t_3, u_1) and (s_3, t_3, u_1), we derive that $s_1 \not\succeq_\varphi s_3$. Hence, if agent i adopted the pro-group intention, only s_1 and s_3 would be admissible with respect to her intention. By symmetry, if agent j adopted the pro-group intention, only t_1 and t_3 would be admissible with respect to her intention.

Discontent with preference transformations, Bacharach, Sugden, and Gold argue, on various occasions, that team-directed reasoning is more appropriate for explaining and predicting cooperative behaviour. Instead, they propose an *agency* transformation. When engaging in team-directed reasoning, a member i of $\mathcal{G}$ first identifies a best combination of individual actions that the group members can perform and then decides to perform the individual action that is her part of that combination. According to our understanding (see Result 1), this can be similarly explained by her adopting the (I-mode) we-intention to perform an individual action that is her part of a best combination of actions that the group members can perform, expressed by $[i\ \texttt{int}]\langle i\ \texttt{stit}\rangle[\mathcal{G}\ \texttt{prom}]\varphi$ and called a *team-directed intention*. An agent adopting a team-directed intention transforms her preferences and adopts as her personal objective performing an individual action that is compatible with a best group action. This may require her to put herself in the shoes of the group agent to determine the best group actions. Then she decides what to do by way of individualistic reasoning. That is, she decides what to do by determining the individual actions that are admissible with respect to her team-directed intention. This way, the reasoning process is left individualistic; only the content of her intention makes irreducible reference to the group. To make this section self-contained, we repeat the definition of team-directed intentions:

Definition 7 (Team-directed intentions)**.** *Suppose group $\mathcal{G}$ collectively intends to φ. Let $i \in \mathcal{G}$ be a member of the group $\mathcal{G}$. Individual agent i's* team-directed intention *is an intention to act compatibly with a best group action, that is,* $[i\ \texttt{int}]\langle i\ \texttt{stit}\rangle[\mathcal{G}\ \texttt{prom}]\varphi$.

Reconsider game S_4 of Figure 5. First, note that, for the group $\mathcal{G}$ consisting of agents i and j, (s_1, t_1), (s_2, t_1), and (s_1, t_2) are the only group actions that are admissible with respect to φ. Indeed, since these are the only group actions that ensure that φ is realized, these are the best group actions (see, for instance, Proposition 1 item **6**). Observe that, in S_4, only agent i has the team-directed intention. That is, agent i's intention in S_4 is represented by the individual actions that are compatible with an admissible group action. It then follows that only s_1 and s_2 are admissible with respect to agent i's team-directed intention. By symmetry, if agent j adopted the team-directed intention, only t_1 and t_2 would be admissible with respect to agent j's team-directed intention.

Now we direct our attention to participatory intentions. Observing that team-directed reasoning corresponds to I-mode we-intentions of the form $[i\ \texttt{int}]\langle i\ \texttt{stit}\rangle[\mathcal{G}\ \texttt{prom}]\varphi$ invites a natural objection. This objection originates from the oddity of using three consecutive modalities to express a team-directed intention. Team-directed intentions merely require the members to perform an individual action that is *only compatible* with a best group action. This is a very weak demand. Therefore, we introduce a third I-mode we-intention type: *participatory intentions*, which require an individual agent to *promote* the realization of a best group action. That is, she adopts as her objective that a best group action is performed and then decides what to do by way of individualistic reasoning.

Definition 9 (Participatory intentions)**.** *Suppose group $\mathcal{G}$ collectively intends to φ. Let $i \in \mathcal{G}$ be a member of the group $\mathcal{G}$. Agent i's* participatory intention *is an intention that the group promotes the group's objective, which is to realize φ, regardless of what others do, that is,* $[i\ \texttt{int}][\mathcal{G}\ \texttt{prom}]\varphi$.

Note the minor, yet crucial, difference: pro-group intentions require the individual agent to adopt the *group's objectives* as her own, whereas a participatory intention requires her to adopt as her personal objective that a *best group action* *is performed*. This reveals that participatory intentions can be viewed as a preference transformation. This preference transformation is not the result of the aggregation of preferences, but crucially relies on group notions. After all, an agent forming such a participatory intention is required to answer the question "What should *we* do?" before being able to answer the question "What should I do (with respect to my participatory intention)?". The appeal to an agency transformation in the team-directed-reasoning account is thus taken up by participatory intentions. However, in contrast to altering the reasoning process, a participatory intention alters the preferences.

Reconsider game S_4 of Figure 5. Again, note that (s_1, t_1), (s_2, t_1), and (s_1, t_2) are the only group actions that are admissible with respect to φ for the group $\mathcal{G}$ consisting of agents i and j. Observe that, in S_4, only agent j has the participatory intention. It follows that the individual action t_3 is incompatible with a best group action, and that t_1 dominates t_2. Therefore, only t_1 is admissible with respect to agent j's participatory intention. By symmetry, if agent i adopted the participatory intention, we derive that only s_1 is admissible with respect to agent i's participatory intention.

In this section, we have introduced and discussed three we-intention types, but why bother? Our discussions of game S_4 illustrate that these three we-intention types yield different action recommendations. If agent i adopted the pro-group intention, then individual actions s_1 and s_3 would be admissible. If instead she adopted the team-directed intention, then individual actions s_1 and s_2 would be admissible. Finally, only individual action s_1 would be admissible if she adopted the participatory intention. This shows that there is a significant difference between these three we-intention types.

6 An objection to team-directed reasoning

It may be thought that even though team-directed intentions and pro-group intentions yield different action recommendations, they nonetheless yield the same cooperative outcomes. In this section, we show that the alternative Hi-Lo game illustrates that this is false. Since team-directed reasoning succeeds in selecting cooperatively rational outcomes in this particular game whereas pro-group intentions fail to do so, it might then be thought that team-directed intentions surpass pro-group intentions with regard to guaranteeing successful cooperation in *all* scenarios. We show that this is also false. This means that team-directed reasoning is on an unsatisfactory par with pro-group intentions.

The alternative Hi-Lo game S_3 of Figure 4 presents a scenario in which team-directed intentions ensure successful cooperation. Pro-group intentions do not fare well in that example: since neither individual action dominates the other

		j	
		L	R
i	U	φ (1,1,1)	φ (1,1,1)
	D	φ (1,1,1)	(1,1,0)

Figure 6: Strategic model with intentions S_5, where the group $\{i, j\} = \mathcal{G}$ collectively intends to φ and both agents adopted the team-directed intention.

with respect to the group's objective, both are admissible. Pro-group intentions hence fail to dismiss all inferior group actions, for instance (*high*, *low*). We take this to reveal that there are games in which pro-group intentions fail to explain the obvious collective incentives, whereas team-directed reasoning succeeds.

There are, however, games revealing the opposite. In Game S_5 of Figure 6, team-directed intentions do not have much to offer: since group $\mathcal{G}$'s actions (U, L), (U, R), and (D, L) ensure φ, they are the best group actions in the context of the collective intention to φ. So, any individual action is compatible with a best group action. In particular, team-directed intentions admit D and R for agents i and j respectively, and hence do not dismiss (D, R). Pro-group intentions do not suffer from this flaw: they recommend U over D and L over R for agents i and j, respectively. So, whereas team-directed intentions risk performing an inferior group action, namely (D, R), pro-group intentions guarantee that a best group action is performed, namely (U, L). This illustrates that there is a game in which team-directed reasoning fails to select cooperatively rational outcomes whereas pro-group intentions succeed in doing so.

We agree with Bacharach (2006, p. 60, Section 8.3) that a theory should only be determinate when our intuitions are. We should hence not presuppose the determinacy of reason.[31] The team-directed reasoning literature, however, mostly presupposes that there is a *unique* best group action.[32] We do not wish to impose this restriction, and introduce a more general theory of cooperation.[33] When dropping the uniqueness assumption, the above discussion reveals that

[31] In contrast, Harsanyi and Selten (1988, p. 13) write: "Clearly a theory telling us no more than that the outcome can be any one of these equilibrium points will not give us much useful information. We need a theory selecting one equilibrium point as the solution of the game."

[32] As mentioned before, Bacharach (1999, p. 120 – adapted notation) is a notable exception: "I shall say that agent i team reasons (for $\mathcal{G}$) when she first computes a best profile a^* (in terms of $u_{\mathcal{G}}$), next computes a_i^*, and lastly decides to do a_i^* because this is the component under her control of a best profile."

[33] Tamminga and Duijf (2017) have a similar spirit and "study a strong sense of joint action in which members of a group, using team reasoning, design and then publicly adopt a group plan". Since such a group plan can be indeterminate, their focus is on "structural conditions that a group plan must meet in order to successfully coordinate the individual actions of the group members".

team-directed reasoning does not surpass pro-group intentions.

Our discussion seems to contradict the result of Bacharach (1999, Theorem 2), which is interpreted as showing that "team reasoning differs from, and is more powerful than, adopting the group's objective and then reasoning in the standard individualistic way" (p. 144).[34] Our results are, however, mutually consistent. The difference between our approach and Bacharach's is the employed individualistic reasoning method: we adopt the admissibility requirement, which states that individuals reason in such a way as to avoid dominated actions; Bacharach relies on equilibrium reasoning, which states that individuals determine an "individualistic best reply" (p. 127).

To summarize, we have discussed a game in which team-directed reasoning is indeterminate whereas pro-group intentions yield a determinate cooperative outcome. In fact, this establishes that team-directed intentions are on an equal footing with pro-group intentions.

7 Participatory intentions prevail

The three we-intention types – pro-group, team-directed, and participatory intentions (see Section 5) – have been introduced to predict and explain cooperative behaviour and incentives. Here we focus on their *effectiveness* with regard to guaranteeing that a best group action is performed. If these intention types have anything to say about cooperation they should certainly advance best group actions across a wide range of games.[35] We thus investigate the classes of games for which these we-intention types guarantee that a best group action is performed. The results of this section are collected in Figure 2 on page 6. As the title of this section already reveals, we prove that our theory of participatory intentions, in contrast with team-directed intentions or pro-group intentions, is the *prevalent* account of cooperation.

Let us start by comparing team-directed intentions and pro-group intentions. We will rephrase our discussion of the previous section. Recall that the alternative Hi-Lo game (see Figure 4 on page 16) presents a scenario in which team-directed intentions succeed in guaranteeing that a best group action is performed whereas pro-group intentions fail to do so. In contrast, the discussion of the game in Figure 6 has revealed that there are also scenarios in which

[34] Hakli et al. (2010, Thesis (3), p. 306) agree with Bacharach and write: "The we-mode tends to create more collective order than the pro-group I-mode: It can decrease the amount of equilibria but it cannot increase them."

[35] Bacharach (2006, p. 58 – adapted notation) writes: "There are three requirements for a good theory of why people play *high* in Hi-Lo: [. . .] (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."

pro-group intentions succeed in guaranteeing successful cooperation whereas team-directed intentions fail to do so. This means that the class of games for which team-directed intentions guarantee that a best group action is performed is not a proper superset of the corresponding class for pro-group intentions, and vice versa. So it is unclear which of these theories offers the best account of cooperation. This results in a stalemate when we study the range of problems for which they guarantee successful cooperation.

Result 2. *Team-directed intentions and pro-group intentions are on a par with regard to guaranteeing successful cooperation.*

Now we turn to participatory intentions and team-directed intentions. It seems uncontroversial to claim that whenever an individual agent decides to perform an individual action that is incompatible with any best group action, then she is certainly not promoting the realization of a best group action. Indeed, this follows logically: (a) In the first part of the proof of Result 1 we show that $\langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi$ entails $[i\ \mathtt{prom}]\langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi$. (b) From items **1** and **4** of Proposition 1 it then follows immediately that

$$\models [i\ \mathtt{prom}][\mathcal{G}\ \mathtt{prom}]\varphi \to [i\ \mathtt{prom}]\langle i\ \mathtt{stit}\rangle[\mathcal{G}\ \mathtt{prom}]\varphi.$$

Participatory intentions hence refine team-directed intentions' action recommendations.

Do participatory intentions guarantee that a best group action is performed if team-directed intentions do? Suppose that we are considering a scenario in which the following holds: when the agents adopt the team-directed intention and therefore perform an individual action that is admissible with respect to it, then ψ will hold. Since participatory intentions refine the action recommendations yielded by team-directed intentions, this means that if the agents adopted the participatory intention, then ψ will also hold. This immediately implies that whenever team-directed intentions guarantee that a best group action is performed, then so do participatory intentions. In particular, this shows that participatory intentions guarantee that a best group action is performed in scenarios in which there is a unique best group action, since team-directed intentions are effective in those scenarios.

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

Result 3 (below) can be viewed as generalizing and strengthening this justification for our theory of participatory intentions. On a positive note, our result establishes that whenever team-directed reasoning is successful in picking out cooperatively rational solutions, then so are participatory intentions. On a negative note, there are scenarios in which team-directed reasoning fails in this respect, whereas participatory intentions succeed.[36] The team-directed reasoning account of cooperation is hence surpassed by our theory of participatory intentions with regard to guaranteeing successful cooperation.

Result 3. *Participatory intentions surpass team-directed intentions with regard to guaranteeing successful cooperation.*

Let us now turn to participatory intentions and pro-group intentions. Our objection against team-directed reasoning and in favour of pro-group intentions originates from game S_5 in Figure 6. So, let us investigate how participatory intentions fare in that game. Recall that, for group $\mathcal{G}$, group actions (U, L), (U, R), and (D, L) are the best group actions in the context of the collective intention to φ. Since these coincide with the φ-worlds, the results of pro-group intentions and participatory intentions coincide in this game. That is, the action recommendations yielded by these we-intention types are identical in this game. In particular, this shows that participatory intentions champion the objection posed to team-directed reasoning.

Can we come up with a different objection against participatory intentions, in favour of pro-group intentions? That is, does a scenario exist in which participatory intentions fail to guarantee that a best group action is performed whereas pro-group intentions succeed in doing so? It is tempting to think that the action recommendations yielded by participatory intentions refine those of pro-group intentions. This is, however, *not* the case.[37] Still, the answer to the questions is negative: the following result shows that participatory intentions also surpass pro-group intentions when it comes to guaranteeing successful cooperation.[38]

Result 4. *Participatory intentions surpass pro-group intentions with regard to guaranteeing successful cooperation.*

[36] The discussion below shows that the game in Figure 6 is a case in point.

[37] To see this, imagine if we slightly change game S_4 in Figure 5: drop φ at (s_1, t_2, u_2) and thus remove (s_1, t_2, u_2) from $\mathsf{Int}_{\mathcal{G}}$. One can check that if agent i adopted the pro-group intention, only s_1 and s_3 would be admissible. In contrast, if agent i adopted the participatory intention, only s_1 and s_2 would be admissible. This example hence proves the point.

[38] Because the action recommendations of participatory intentions do not refine those of pro-group intentions, the proof that participatory intentions surpass team-directed intentions is more complicated than that of Result 3.

To prove this result, we rely on the following lemma:

Lemma 1. *Let S be an SMI. Let $a \in A$ satisfy $S, a \vDash \bigwedge_{j \in \mathcal{G}} [j\ \mathtt{prom}][\mathcal{G}\ \mathtt{prom}]\varphi$ and $S, a \not\vDash [i\ \mathtt{prom}]\varphi$. Then there is a profile $b \in A$ satisfying:*

1. $S, (b_i, a_{-i}) \vDash [i\ \mathtt{prom}]\varphi$
2. $b_i \preceq_{[\mathcal{G}\ \mathtt{prom}]\varphi} a_i$
3. $S, (b_i, a_{-i}) \vDash \bigwedge_{j \in \mathcal{G}} [j\ \mathtt{prom}][\mathcal{G}\ \mathtt{prom}]\varphi$

Proof. Assume all the mentioned assumptions. Since $S, a \not\vDash [i\ \mathtt{prom}]\varphi$, there is a $b \in A$ satisfying $b_i \succ_\varphi a_i$ and $S, b \vDash [i\ \mathtt{prom}]\varphi$. Hence, for any $c_{-i} \in A_{-i}$ we have that $S, (a_i, c_{-i}) \vDash \varphi$ implies $S, (b_i, c_{-i}) \vDash \varphi$. Then for any $c_{-i} \in A_{-i}$ we have $(b_i, c_{\mathcal{G}-i}) \succeq_\varphi (a_i, c_{\mathcal{G}-i})$. Hence for any $c_{-i} \in A_{-i}$ we have that $S, (a_i, c_{-i}) \vDash [\mathcal{G}\ \mathtt{prom}]\varphi$ implies $S, (b_i, c_{-i}) \vDash [\mathcal{G}\ \mathtt{prom}]\varphi$. In other words, $b_i \succeq_{[\mathcal{G}\ \mathtt{prom}]\varphi} a_i$. Since $S, a \vDash [i\ \mathtt{prom}][\mathcal{G}\ \mathtt{prom}]\varphi$, this entails $b_i \preceq_{[\mathcal{G}\ \mathtt{prom}]\varphi} a_i$, and $S, (b_i, a_{-i}) \vDash [i\ \mathtt{prom}][\mathcal{G}\ \mathtt{prom}]\varphi$. To prove the third item, it remains to prove that for every $j \in \mathcal{G} - i$ we have $S, (b_i, a_{-i}) \vDash [j\ \mathtt{prom}][\mathcal{G}\ \mathtt{prom}]\varphi$. This follows immediately from the fact that the j-th component in (b_i, a_{-i}) equals that in a and the assumption that $S, a \vDash \bigwedge_{j \in \mathcal{G}} [j\ \mathtt{prom}][\mathcal{G}\ \mathtt{prom}]\varphi$. □

Proof of Result 4. Let $S = \langle N, (A_i), (\mathsf{Int}_\mathcal{H}), \pi \rangle$ be an SMI. Suppose $\mathcal{G}$ collectively intends to φ, and suppose that pro-group intentions guarantee that a best group action is performed. Let $a \in A$ satisfy $S, a \vDash \bigwedge_{i \in \mathcal{G}} [i\ \mathtt{prom}][\mathcal{G}\ \mathtt{prom}]\varphi$. Using the previous lemma, we can show, by induction on the size of $\mathcal{F} := \{j \in \mathcal{G} \mid S, a \not\vDash [j\ \mathtt{prom}]\varphi\}$, that there is a $b \in A$ satisfying (1) $S, (b_\mathcal{F}, a_{-\mathcal{F}}) \vDash \bigwedge_{j \in \mathcal{G}} [j\ \mathtt{prom}]\varphi$ and (2) $b_j \preceq_{[\mathcal{G}\ \mathtt{prom}]\varphi} a_j$ for every $j \in \mathcal{G}$. In light of the assumption that pro-group intentions guarantee that a best group action is performed, (1) implies $S, (b_\mathcal{F}, a_{-\mathcal{F}}) \vDash [\mathcal{G}\ \mathtt{prom}]\varphi$. Using (2), this can be shown, by induction on the size of $\mathcal{F}$, to imply $S, a \vDash [\mathcal{G}\ \mathtt{prom}]\varphi$. This shows that participatory intentions guarantee that a best group action is performed. □

This result emphasizes that our theory of participatory intentions surpasses team-directed reasoning *and* pro-group intentions in selecting cooperatively rational solutions. Indeed, in *any* scenario in which pro-group intentions guarantee successful cooperation, participatory intentions do so too. So, to cooperate successfully, it is generally better if all members take up the participatory intention.

Considered together, the results of this section purport that our theory of participatory intentions, in contrast with team-directed reasoning or pro-group intentions, is the prevalent account of cooperation. Since team-directed reasoning does not surpass pro-group intentions in some scenarios, this provides ample justification for our theory of participatory intentions beyond the team-directed-reasoning account. After all, whereas team-directed reasoning surpasses pro-group intentions with regard to guaranteeing successful cooperation only in *some* scenarios, participatory intentions surpass both in *all* scenarios. Our theory of participatory intentions therefore best explains and predicts cooperation.

8 Discussion

The team-directed-reasoning account of cooperation (as studied by Bacharach, Sugden, and Gold) has been criticized on two counts: first, team-directed reasoning is supposed to transcend traditional game and decision theory by adopting a certain collectivistic reasoning method. However, we have shown that an important part of we-mode reasoning reduces to I-mode reasoning with a particular preference expressed by the team-directed intention. That is, the action recommendations yielded by team-directed reasoning coincide with those yielded by team-directed intentions. This moves team-directed reasoning into the domain of preference transformations coupled with dominance reasoning. Traditional game and decision theory can thus explain team-directed reasoning as a particular preference transformation.

Secondly, we have shown that the team-directed-reasoning account does not surpass pro-group intentions in selecting cooperatively rational solutions. Similar to traditional preference transformation theories, pro-group intentions require an individual agent to adopt the group's objective as her own. These pro-group intentions are on a par with team-directed intentions. That is, in some scenarios that lack a unique best group action, pro-group intentions succeed in selecting a cooperatively rational solution, whereas team-directed intentions fail to do so. The class of games for which team-directed intentions guarantee successful cooperation is hence not a proper superset of the corresponding class for pro-group intentions, and vice versa. So it is unclear which of these theories offers the best account of cooperation. This results in a stalemate when we study the range of problems for which they guarantee successful cooperation. Reform is required.

This deadlock is resolved by our theory of participatory intentions: a participatory intention requires a member to promote the realization of a best group action, regardless of what others do. Participatory intentions' action recommendations refine those of team-directed intentions, because the latter require a member to act only compatibly with a best group action. This entails that in *any* scenario in which team-directed intentions guarantee successful cooperation, participatory intentions do so too. In contrast, participatory intentions do not refine the action recommendations yielded by pro-group intentions. Nonetheless, in *any* scenario in which pro-group intentions guarantee that a best group action is performed, participatory intentions do so too. Our theory of participatory intentions thus overcomes the deadlock and is the prevalent account of cooperation.

Alternatively, since we focus solely on the third stage of we-mode reasoning,[39] one could view our study as attempting to understand the logical form of the we-intentions resulting from team-directed reasoning. The logical machinery helps to address this question more precisely than what has been done before. We revise the theory by showing that we-mode reasoning at the third stage may be improved by taking these resulting we-intentions to be participatory intentions rather than team-directed intentions.

Acknowledgements

Special thanks go to Jan Broersen, John-Jules Meyer and Allard Tamminga for helpful comments and fruitful discussions. I would like to thank the audiences at Collective Intentionality X in The Hague and at the Logic and Interactive Rationality seminar in Amsterdam for their helpful suggestions. I am grateful for the extensive and extremely useful reports of the referees of this journal. Finally, I gratefully acknowledge financial support from the European Research Council, ERC-2013-CoG Project Number 616512, and thank the Centre for Philosophy of Natural and Social Science (CPNSS) at the London School of Economics for the facilities and support during my visit.

References

Anscombe, G. E. M. (1963), *Intention*, Harvard University Press, Cambridge.

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*, Princeton University Press, Princeton.

Brandenburger, A., Friedenberg, A. and Keisler, H. J. (2008), 'Admissibility in games', *Econometrica* **76**(2), 307–352.

Bratman, M. E. (1987), *Intention, Plans, and Practical Reason*, Harvard University Press, Cambridge.

[39] These three stages have been discussed in Section 1 (see Hakli et al. 2010, Section 4).

Bratman, M. E. (2014), *Shared Agency: A Planning Theory of Acting Together*, Oxford University Press, Oxford.

Broersen, J. (2011), Modeling attempt and action failure in probabilistic stit logic, *in* T. Walsh, ed., 'Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence', AAAI Press, pp. 792–797.

Chellas, B. F. (1992), 'Time and modality in the logic of agency', *Studia Logica* **51**(3/4), 485–517.

Gilbert, M. (2009), 'Shared intention and personal intentions', *Philosophical Studies* **144**(1), 167–187.

Gold, N. (2012), Team reasoning, framing and cooperation, *in* S. Okasha and K. Binmore, eds, 'Evolution and Rationality: Decisions, Co-operation and Strategic Behaviour', Cambridge University Press, New York, pp. 185–212.

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

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

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

Horty, J. F. (1996), 'Agency and obligation', *Synthese* **108**(2), 269–307.

Horty, J. F. (2001), *Agency and Deontic Logic*, Oxford University Press, New York.

Kohlberg, E. and Mertens, J.-F. (1986), 'On the strategic stability of equilibria', *Econometrica* **54**(5), 1003–1037.

Konolige, K. and Pollack, M. E. (1993), A representationalist theory of intention, *in* R. Bajcsy, ed., 'Proceedings of the Thirteenth International Joint Conference on Artifical Intelligence', Vol. 1, pp. 390–395.

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

Luce, R. D. and Raiffa, H. (1957), *Games and Decisions*, John Wiley & Sons, New York.

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

Savage, L. (1954), *The Foundations of Statistics*, John Wiley & Sons, New York.

Searle, J. (1990), Collective intentions and actions, *in* P. R. Cohen, J. Morgan and M. E. Pollack, eds, 'Intentions in Communication', MIT Press, Cambridge, pp. 401–415.

Selten, R. (1975), 'Reexamination of the perfectness concept for equilibrium points in extensive games', *International Journal of Game Theory* **4**(1), 25–55.

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. (2003), 'The logic of team reasoning', *Philosophical Explorations* **6**(3), 165–181.

Tamminga, A. and Duijf, H. (2017), 'Collective obligations, group plans and individual actions', *Economics & Philosophy* **33**(2), 187–214.

Tuomela, R. (2000), 'Collective and joint intention', *Mind & Society* **1**(2), 39–69.

Tuomela, R. (2005), 'We-intentions revisited', *Philosophical Studies* **125**(3), 327–369.

Tuomela, R. (2006), 'Joint intention, we-mode and I-mode', *Midwest Studies in Philosophy* **30**(1), 35–58.

Tuomela, R. (2007), *The Philosophy of Sociality*, Oxford University Press, New York.

van Hees, M. and Roy, O. (2008), Intentions and plans in decision and game theory, *in* B. Verbeek, ed., 'Reasons and Intentions', Ashgate Publishing Limited, pp. 207–226.

HEIN DUIJF
Utrecht University
Janskerkhof 13, 3512 BL Utrecht
The Netherlands
h.w.a.duijf@uu.nl
`http://www.uu.nl/staff/hwaduijf/`