Coherent Social Action

Michael Wooldridge[1]

Abstract. Formal analyses of social action for Distributed A.I. (DAI) have focussed, almost exclusively, on scenarios in which participating agents have a joint intention to act. While such scenarios are significant, there are many examples of artificial and natural social systems in which joint intention not only does not occur, but is not even a practical possibility. This paper proposes that a deep theory of social action should account for the whole spectrum of social action types within the same framework. It is argued that coherence is an attribute which unifies different types of social action, and is therefore a possible starting point for a deep theory. A discussion and subsequent formalisation of coherence is then presented. This model of coherence is used as the foundation upon which to build a new formalisation of team action. The framework in which these formalisations are presented is a new quantified multi-modal logic.

1 INTRODUCTION

Over the past decade, many accounts of social action have appeared in the literature of DAI, philosophy, and the social sciences (see [6] for a recent review). Most of these accounts have focussed on the phenomenon of *team action*, in which a close-knit, highly organised collective cooperates to achieve some common goal. Although accounts differ on details, a common theme is that this type of social action is characterised by a mental state, shared among participating agents, which is variously referred to as mutual-, we-, or joint-intention.

While team action is undoubtedly important, there are many other important types of social action that team action theories cannot account for. There are several reasons for this. First, 'mutual' mental states, a feature common to most accounts of team action, are not realisable in practice [4]; even approximations to them are likely to entail heavy communication overheads. Secondly, there are many examples of social actions in which the actors do not have the complex beliefs and goals implied by theories of collective intention. For example, consider driving in ordinary automobile traffic: this is not *team* action, (in the sense of [7]), but it is, nevertheless, a social activity of interest to DAI. There are even extreme examples of artificial and natural social actions in which the actors apparently have no cognitive state. For example, Steels demonstrated that a network of cognitively simple agents could achieve near-optimal performance in a non-trivial cooperative task [12]. Steels' agents do not have any explicitly held beliefs or goals, and are thus literally *indescribable* by most accounts of collective intentions.[2] The same is true of many natural examples of swarm intelligence and emergent functionality. So team actions, characterised by some form of collective intention, represent just one point in a broad spectrum of different types of social action. What distinguishes these different action types is the cognitive state of the actors, and in particular the amount and complexity of the beliefs and goals that agents have relating to their role in the social action, and how this role relates to that of others. At one extreme, (team action), agents have many mutual beliefs, goals, and expectations about their role and that of others; at the other extreme, there are social systems in which agents do not even appear to *have* beliefs or goals.

This paper argues the case for a *deep theory* of social action, which accounts for the whole spectrum of social action types within the same general framework. Since current models of team action cannot function as deep theories, we are forced to look elsewhere. It is proposed that *coherence* is a possible starting point for a deep theory, as the notion of coherence runs through the whole social action spectrum. For example, when we view Steels' Mars explorer system, we do not see a group of agents acting in a random manner. Instead, we see that the actions are related, and that the global behaviour of the system emerges from the individual behaviours. In short, we see *coherent* action. Similarly, a group of complex reasoning agents cooperating on some task will use their communication, reasoning and representation abilities to ensure that they are acting coherently; if the collective action ever starts becoming incoherent, then the group will take steps to correct this. The difference is simply that in the latter case, agents have complex internal models of the task at hand and the agents they are working with, whereas in the former case, the agents have no such models.

The remainder of this paper represents some first steps towards a deep theory of social action, based on the notion of coherence. The following section presents a new logic for representing the beliefs, goals, and actions of agents and groups of agents. Following a discussion, this logic is used in §3 to formalise the notion of coherence. This formalisation is then used as the basis upon which to develop a new model of team action. Some conclusions are presented in §4. Finally, it is worth commenting on the aims of this work. It *is* intended to contribute to the theoretical foundations of DAI, by providing a model which can be used to help researchers understand what is involved in social action. But the work is *not* intended to be social science research; while such research is often of use in DAI, it is not the case that DAI theory must be social science theory.

2 A FORMAL FRAMEWORK

This section develops a many-sorted quantified multi-modal logic for representing the beliefs, goals, and actions of agents and groups of agents. This logic both draws upon and extends the formalisms of [2, 8, 13, 14]. Note that although the logic *is* completely defined, space restrictions mean the presentation must be somewhat terse, and a discussion of the logic's properties is unfortunately not possible here.

[1] Department of Computing, Manchester Metropolitan University, Chester Street, Manchester M1 5GD, United Kingdom

[2] We ignore knowledge implicit in the state of Steels' agents, *à la* situated automata theory [9]; it is at least clear that Steels' agents do not have *explicitly represented* beliefs and goals.

ECAI 94. *11th European Conference on Artificial Intelligence* Edited by A. Cohn
Published in 1994 by John Wiley & Sons, Ltd.

Informally, the operators of the language have the following meanings. The operator **true** is a logical constant for truth. (Bel i φ) and (Goal i φ) mean that agent i has a belief, or goal of φ respectively. The = operator is first-order equality. The $\in$ operator allows us to relate agents to groups of agents; it has the expected set-theoretic interpretation, so ($i \in g$) means the agent denoted by i is a member of the group denoted by g. The $\sqsubseteq$ operator defines a partial ordering on action terms: ($\alpha \sqsubseteq \alpha'$) means that the actions referred to in α are a subset of those referred to in α'. The (Agts α g) operator means that the group denoted by g are precisely the agents required to perform the actions referred to in α. The A operator is a *path quantifier*: Aφ means that φ is a *path formula* that is satisfied in all the futures that could arise from the current state.[3] The operators $\neg$ (not) and $\vee$ (or) have classical semantics, as does the universal quantifier $\forall$; the remaining classical connectives and existential quantifier are assumed to be introduced as abbreviations. (Happens α) is a path formula that means that the action α happens next; $\alpha; \alpha'$ means action α immediately followed by α'; $\alpha|\alpha'$ means either α or α' happen next; $\varphi?$ is a test action, which occurs if φ is satisfied in the current state; $\alpha*$ means the action α iterated.

Readers that are unfamiliar with quantified modal logics, or that do not wish to read the formal definition of the logic, might wish to skip the remainder of this section, and move directly to §3.

Definition 1 *The language contains the following symbols: the propositional connectives $\neg$ (not) and $\vee$ (or), and universal quantifier $\forall$; the operator symbols* Bel, Goal, Happens, Agts, $\in$, =, $\sqsubseteq$, *and* A*; the action constructor symbols* ;, |, ?, *and* $*$*; a countable set Pred of predicate symbols --- each symbol $P \in$ Pred is associated with a natural number called its arity, given by arity(P); a countable set Const of constant symbols, the union of the mutually disjoint sets $Const_{Ag}$ (agent constants), $Const_{Ac}$ (action sequence constants), $Const_{Gr}$ (group constants), and $Const_U$ (other constants); a countable set Var of variable symbols, the union of the mutually disjoint sets Var_{Ag}, Var_{Ac}, Var_{Gr} and Var_U; the punctuation symbols*), (, *'·' and comma ','.*

Definition 2 *A term is either a constant or a variable; the set of terms is Term. The sort of a term is either Ag, Ac, Gr or U; if s is a sort then $Term_s = Const_s \cup Var_s$; thus $\tau_s \in Term_s$.*

The syntax of (well-formed) formulae is constructed from these symbols according to the rules presented informally above. Note that we demand that a predicate P is applied to $arity(P)$ terms.

It is assumed that the world may be in any of a set S of *states*. A state *transition* is caused by the occurrence of a *primitive action* (or *event*): the set of all primitive actions is D_{Ac}. From any state, there is at least one — and perhaps many — possible actions, and hence resultant states. The binary relation R on S is used to represent all possible courses of world history: $(s, s') \in R$ iff the state s could be transformed into state s' by the occurrence of a primitive action that is possible in s. Clearly, R will *branch* infinitely into the future from every state. A labelling function *Act* maps each arc in R to the action associated with the transition. The world is populated by a non-empty set D_{Ag} of *agents*. A *group* over D_{Ag} is simply a non-empty subset of D_{Ag}; the set of all such groups is D_{Gr}. Agents and groups may be related to one-another via a simple (typed) set theory. Agents have beliefs and goals, and are (idealised) reasoners. The beliefs of an agent are given by a *belief accessibility relation* on S in the usual way; similarly for goals. Every primitive action α is associated with an agent, given by $Agt(\alpha)$. Finally, the world contains other individuals (e.g., chairs, pints of beer) given by the set D_U. A complete definition of the language semantics will now be given. First, *paths* (a.k.a. fullpaths) will be defined: a path represents a possible course of events through a branching time structure.

Definition 3 *If S is a non-empty set and R is a total binary relation on S then a path over S, R is an infinite sequence $(s_u : u \in \mathbb{N})$ such that $\forall u \in \mathbb{N}$, $s_u \in S$ and $(s_u, s_{u+1}) \in R$. The set of all paths over S, R is given by paths(S, R). The head of a path $p = (s_0, \ldots)$ is its first element s_0, and is given by hd(p).*

Next, we present the technical apparatus for dealing with the denotation of terms.

Definition 4 *The domain of quantification, D, is $D_{Ag} \cup (D^*_{Ac}) \cup D_{Gr} \cup D_U$, (where S^* denotes the set of non-empty sequences over S). If $n \in \mathbb{N}$, then the set of n-tuples over D is D^n. An interpretation for constants, I, is a sort-preserving bijection $I : Const \rightarrow D$. A variable assignment, V, is a sort-preserving bijection $V : Var \rightarrow D$.*

The function $[\![\ ]\!]_{I,V}$ gives the denotation of a term relative to I, V.

Definition 5 *If $\tau \in Term$, then $[\![\tau]\!]_{I,V}$ is $I(\tau)$ if $\tau \in Const$, and $V(\tau)$ otherwise. Reference to I, V will usually be suppressed.*

Definition 6 *A model, M, is a structure:*

$$\langle S, R, D_{Ag}, D_{Ac}, D_{Gr}, D_U, Act, Agt, B, G, I, \Phi \rangle$$

where: S is a non-empty set of states; $R \subseteq S \times S$ is a total binary relation on S; D_{Ag} is a non-empty set of agents; D_{Ac} is a non-empty set of actions; D_{Gr} is the set of groups over D_{Ag}; D_U is a non-empty set of other individuals; $Act : R \rightarrow D_{Ac}$ associates a primitive action with each arc in R; $Agt : D_{Ac} \rightarrow D_{Ag}$ gives the agent of each primitive action; $B : D_{Ag} \rightarrow powerset(S \times S)$ associates a transitive, euclidean, serial belief accessibility relation with every agent in D_{Ag}; $G : D_{Ag} \rightarrow powerset(S \times S)$ associates a serial goal accessibility relation with every agent in D_{Ag}, such that $\forall i \in D_{Ag}$, $G(i) \subseteq B(i)$; $I : Const \rightarrow D$ is an interpretation for constants; and finally $\Phi : Pred \times S \rightarrow \bigcup_{n \in \mathbb{N}} D^n$ gives the extension of each predicate symbol in each state, such that $\forall P \in Pred$, $\forall n \in \mathbb{N}$, $\forall s \in S$, if $arity(P) = n$ then $\Phi(P, s) \subseteq D^n$ (i.e., Φ preserves arity).

The semantics of the language are defined via the satisfaction relation, '$\models$', which holds between *interpretation structures* and formulae. For state formulae, an interpretation structure is a triple $\langle M, V, s \rangle$, where M is a model, V is a variable assignment and s is a state. For path formulae, an interpretation structure is a triple $\langle M, V, p \rangle$, where p is a path. The rules defining the satisfaction relation are given in Figure 1. The rules make use of some syntactic abbreviations. First, we write $occurs(\alpha, u, v, p)$ if action α occurs between 'times' $u, v \in \mathbb{N}$ on the (possibly finite) path p:

[3] There is a distinction made in the language between *path* and *state* formulae: state formulae are evaluated with respect to the 'current state' of the world, whereas path formulae are evaluated with respect to a course of events. The well-formed formulae of the language are identified with the set of state formulae [3, 8].

$$
\begin{array}{rcl}
occurs(\alpha, u, v, (s_0, \ldots)) & \text{iff} & [\![\alpha]\!] = (\alpha_1, \ldots, \alpha_n), n \leq v - u, \\
 & & \quad \text{and } \forall w \in \{1, \ldots, n\}, \\
 & & \quad Act(s_{u+w-1}, s_{u+w}) = \alpha_w \\
 & & \quad \text{(where } \alpha \in Term_{Ac}) \\
occurs(\alpha; \alpha', u, v, p) & \text{iff} & \exists w \in \{u, \ldots, v\} \text{ s.t.} \\
 & & \quad occurs(\alpha, u, w, p) \text{ and} \\
 & & \quad occurs(\alpha', w, v, p) \\
occurs(\alpha | \alpha', u, v, p) & \text{iff} & occurs(\alpha, u, v, p) \text{ or} \\
 & & \quad occurs(\alpha', u, v, p) \\
occurs(\varphi?, u, v, p) & \text{iff} & \langle M, V, hd(p) \rangle \models \varphi \\
occurs(\alpha*, u, v, p) & \text{iff} & \exists w_1, \ldots, w_x \in I\!N \text{ s.t. } (w_1 = 0) \\
 & & \quad \text{and } (w_1 < \cdots < w_x) \\
 & & \quad \text{and } \forall y \in \{1, \ldots, x\}, \\
 & & \quad occurs(\alpha, w_y, w_{y+1}, p)
\end{array}
$$

Two functions are required, that return all the primitive actions referred to in an action sequence, and the agents required for an action term, respectively.

$$
\begin{array}{rcl}
actns((\alpha_1, \ldots, \alpha_n)) & \stackrel{\text{def}}{=} & \{\alpha_1, \ldots, \alpha_n\} \\
agents(\alpha) & \stackrel{\text{def}}{=} & \{i \mid \exists \alpha' \in actns([\![\alpha]\!]) \text{ s.t. } Agt(\alpha') = i\} \\
 & & \quad \text{(where } \alpha \in Term_{Ac})
\end{array}
$$

Some Derived Operators. A number of derived operators will now be introduced. First, the usual connectives of linear temporal logic: $\varphi \mathcal{U} \psi$ means φ is satisfied *until* ψ becomes satisfied; $\Diamond \varphi$ means φ is *eventually* satisfied; $\Box \varphi$ means φ is *always* satisfied. These connectives are used to build path formulae. The path quantifier E is the dual of A; thus $\mathsf{E}\varphi$ means φ is a path formulae satisfied on *at least one* possible future.

$$
\begin{array}{rcl@{\qquad}rcl}
\varphi \mathcal{U} \psi & \stackrel{\text{def}}{=} & (\text{Happens } (\neg\psi?; \varphi?)*; \psi?) & \Box\varphi & \stackrel{\text{def}}{=} & \neg\Diamond\neg\varphi \\
\Diamond\varphi & \stackrel{\text{def}}{=} & \textbf{true}\,\mathcal{U}\,\varphi & \mathsf{E}\varphi & \stackrel{\text{def}}{=} & \neg\mathsf{A}\neg\varphi
\end{array}
$$

The next operator allows us to relate agents and groups of agents: $(\mathsf{Singleton}\ g\ i)$ means g is a singleton group with i as the only member.

$$(\mathsf{Singleton}\ g\ i) \stackrel{\text{def}}{=} \forall j \cdot (j \in g) \Rightarrow (j = i)$$

The Prim operator defines the conditions under which an action term denotes a primitive action; $\sqsubset$ has the obvious meaning; and $(\mathsf{Agt}\ \alpha\ i)$ means i is the only agent of α.

$$
\begin{array}{rcl}
(\mathsf{Prim}\ \alpha) & \stackrel{\text{def}}{=} & \forall \alpha' \cdot (\alpha' \sqsubseteq \alpha) \Rightarrow (\alpha' = \alpha) \\
(\alpha \sqsubset \alpha') & \stackrel{\text{def}}{=} & (\alpha \sqsubseteq \alpha') \wedge \neg(\alpha = \alpha') \\
(\mathsf{Agt}\ \alpha\ i) & \stackrel{\text{def}}{=} & \forall g \cdot (\mathsf{Agts}\ \alpha\ g) \Rightarrow (\mathsf{Singleton}\ g\ i)
\end{array}
$$

Finally, the mutual belief of φ in a group of agents g is $(\mathsf{M\text{-}Bel}\ g\ \varphi)$; the mutual goal of φ in g is $(\mathsf{M\text{-}Goal}\ g\ \varphi)$. Mutual mental states are defined as *fixed points*.

$$
\begin{array}{rcl}
(\mathsf{M\text{-}Bel}\ g\ \varphi) & \stackrel{\text{def}}{=} & \forall i \cdot (i \in g) \Rightarrow (\mathsf{Bel}\ i\ \varphi \wedge (\mathsf{M\text{-}Bel}\ g\ \varphi)) \\
(\mathsf{M\text{-}Goal}\ g\ \varphi) & \stackrel{\text{def}}{=} & \forall i \cdot (i \in g) \Rightarrow (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Goal}\ i\ \mathsf{A}\Diamond\varphi))
\end{array}
$$

3 COHERENCE IN SOCIAL ACTION

This section begins by discussing and subsequently formalising *coherent social action*, using the language developed in §2. This formalisation is then used as a foundation upon which to build a new model of team action.

Coherent Social Action. Coherence in DAI is a measure of 'how well … [a] system performs along some dimension of evaluation' [1, p19]. It follows that coherence may be evaluated in many different ways: in terms of solution quality, efficiency, or conceptual clarity, for example. In this paper, however, the specific interpretation given to the term will be: *the effectiveness with which the primitive actions in a complex action conspire to bring about a particular goal*. Note that this definition is neutral on the subject of cognitive state: it does not require that agents have complex internal models. It does not even require that agents have any conception of the goal they are acting coherently to achieve. Coherence merely requires that we, as external observers of a system, judge that the agents appear to be acting effectively to achieve the goal. This definition is thus applicable both to systems of cognitively simple agents (e.g., Steels' Mars explorer system [12]), and to systems with agents capable of complex representation and reasoning tasks (as in [7]). Coherence, as we have defined it, is therefore an *external* concept, relying only on the actions that agents perform in the world, as opposed to an internal one, defined in terms of mental state [11, 6].

So, when may an arbitrary conglomeration of primitive actions be said to be coherent with respect to a goal? We propose that the actions must satisfy at least the following two conditions: (i) the actions should actually *achieve* the goal; and (ii) the actions should, in a sense to be explained shortly, be *minimal* with respect to the goal. To understand the first condition, consider that if the goal is not true after the actions are performed, then the actions could hardly be said to effectively conspire to achieve the goal. Moreover, it is not sufficient simply to require that the goal is a *possible* consequence of the actions. The goal must be a *necessary* consequence, in that every time the actions are performed, the goal state is subsequently satisfied. This notion of an action α achieving a goal φ is formalised as follows.

$$(\mathsf{Achieves}\ \alpha\ \varphi) \stackrel{\text{def}}{=} \mathsf{A}((\mathsf{Happens}\ \alpha) \Rightarrow (\mathsf{Happens}\ \alpha; \varphi?))$$

The reader may like to compare this definition with the dynamic logic $[\alpha]\varphi$ (see, e.g., [5]).

The purpose of the second condition is to ensure that each of the primitive actions *contributes* something to the achievement of the goal. This requirement may be best illustrated through a simple example. Suppose I have a goal of owning a pint of beer; then the 'primitive' action of ordering a beer is a good candidate for my next action, since it has the goal as a necessary consequence. Now add to this primitive the action of ordering a packet of peanuts. The resulting conglomerate action still has owning a pint of beer as a necessary consequence, and yet it includes an action that is clearly redundant with respect to the goal; ordering peanuts contributes nothing to owning a pint of beer. For this reason, we would say that the conglomerate action was *incoherent* with respect to the goal. Formally, we shall call the second property *minimality*, and say that an action sequence α is minimal with respect to a goal φ if there is no sub-action of α that has φ as a necessary consequence. If α is minimal with respect to φ, we write $(\mathsf{Min}\ \alpha\ \varphi)$.

$$(\mathsf{Min}\ \alpha\ \varphi) \stackrel{\text{def}}{=} \neg(\exists \alpha' \cdot (\alpha' \sqsubset \alpha) \wedge (\mathsf{Achieves}\ \alpha'\ \varphi))$$

In mathematical terms, a minimal action is a *least fixed point*. Coherent social action can now be informally defined.

> **Coherent social action:** Group g perform social action α that is coherent with respect to φ iff: (i) the agents required to perform α are just those in g; (ii) the agents actually do α; (iii) α achieves φ; and (iv) α is minimal with respect to φ.

State Formulae Semantics

$\langle M, V, s\rangle \models$ **true**

$\langle M, V, s\rangle \models (P\ \tau_1, \ldots, \tau_n)$ iff $\langle [\![\tau_1]\!], \ldots, [\![\tau_n]\!]\rangle \in \Phi(P, s)$

$\langle M, V, s\rangle \models (\mathsf{Bel}\ i\ \varphi)$ iff $\forall s' \in S$, if $(s, s') \in B([\![i]\!])$ then $\langle M, V, s'\rangle \models \varphi$

$\langle M, V, s\rangle \models (\mathsf{Goal}\ i\ \varphi)$ iff $\forall s' \in S$, if $(s, s') \in G([\![i]\!])$ then $\langle M, V, s'\rangle \models \varphi$

$\langle M, V, s\rangle \models (\mathsf{Agts}\ \alpha\ g)$ iff $agents(\alpha) = [\![g]\!]$

$\langle M, V, s\rangle \models (\tau_1 = \tau_2)$ iff $[\![\tau_1]\!] = [\![\tau_2]\!]$

$\langle M, V, s\rangle \models (i \in g)$ iff $[\![i]\!] \in [\![g]\!]$

$\langle M, V, s\rangle \models (\alpha \sqsubseteq \alpha')$ iff $actns([\![\alpha]\!]) \subseteq actns([\![\alpha']\!])$

$\langle M, V, s\rangle \models \mathsf{A}\varphi$ iff $\forall p \in paths(S, R)$, if $hd(p) = s$ then $\langle M, V, p\rangle \models \varphi$

$\langle M, V, s\rangle \models \neg\varphi$ iff $\langle M, V, s\rangle \not\models \varphi$

$\langle M, V, s\rangle \models \varphi \vee \psi$ iff $\langle M, V, s\rangle \models \varphi$ or $\langle M, V, s\rangle \models \psi$

$\langle M, V, s\rangle \models \forall x \cdot \varphi$ iff $\langle M, V \dagger \{x \mapsto d\}, s\rangle \models \varphi$
for all $d \in D$ s.t. x and d are of the same sort

Path Formulae Semantics

$\langle M, V, p\rangle \models (\mathsf{Happens}\ \alpha)$ iff $\exists u \in \mathbb{N}$ s.t. $occurs(\alpha, 0, u, p)$

$\langle M, V, p\rangle \models \varphi$ iff $\langle M, V, hd(p)\rangle \models \varphi$ (where φ is a state formula)

$\langle M, V, p\rangle \models \neg\varphi$ iff $\langle M, V, p\rangle \not\models \varphi$

$\langle M, V, p\rangle \models \varphi \vee \psi$ iff $\langle M, V, p\rangle \models \varphi$ or $\langle M, V, p\rangle \models \psi$

$\langle M, V, p\rangle \models \forall x \cdot \varphi$ iff $\langle M, V \dagger \{x \mapsto d\}, p\rangle \models \varphi$
for all $d \in D$ s.t. x and d are of the same sort

Figure 1. Semantics

Formally, the conditions of satisfaction for the performance of a coherent social action α by group g with respect to φ are:

$$(\mathsf{CSA}\ g\ \alpha\ \varphi) \stackrel{\text{def}}{=} (\mathsf{Agts}\ \alpha\ g) \wedge (\mathsf{Happens}\ \alpha) \wedge (\mathsf{Achieves}\ \alpha\ \varphi) \wedge (\mathsf{Min}\ \alpha\ \varphi).$$

Let us now look at the implications of this definition. First, consider some possible objections to it. The most obvious objection may be illustrated through the following scenario, due to Searle [10]. A group of people in a park suddenly run to a tree. If the people are dancers, and the choreography calls upon them to converge on the tree, then this action could be viewed as being cooperative. But if it has just started raining, and the people are trying to avoid getting wet by running for shelter, so that their actions are motivated by individual desires, then Searle argues that this is not cooperation. Both cases would be recognised as coherent social action by the definition we have just presented. But a distinction can *only* be made by appealing to an internal perspective, and any system which did not lend itself to such an analysis would presumably not be regarded as cooperative. Thus, systems such as Steels' Mars explorer would not be regarded as cooperative [12]. This seems unreasonable: although a designer might find it useful to build systems that have internalised beliefs and goals, it is possible to build efficient DAI systems that do not have such internal models, as Steels' work demonstrates. Internalised beliefs and goals are not *necessary* for efficient social action.

Secondly, it could be argued that the requirements for coherence are too strong. Consider again Steels' Mars explorer system. The agents in this system walk in random directions as part of their programming: it is therefore inevitable that they will perform redundant actions, and as a result they will be incoherent by this definition. But from the point of view of a designer, the definition recognises those actions that do not include redundancy. In general, it is precisely such behaviour, (that does not include redundancy), that we desire of DAI systems [1]. So the definition captures an important and useful, if idealised, design concept. Note that although minimality is an ideal, which comparatively few systems might achieve in practice, it is nevertheless a useful concept for analysis — consider the widespread use of other idealised concepts (such as mutual belief) in the analysis of multi-agent systems.

To the best of the author's knowledge, no other attempts to formalise coherence have appeared in the literature. However, other authors have considered similar notions. Probably the closest is the work of Singh, who used the necessary consequence property in his definition of group intentions [11]; the distinction between internal and external views is also due to Singh. Werner used a somewhat similar idea to capture the notion of a group planning to do an action [13]; in an earlier paper, we also used a related idea to try to capture a notion of joint goals [14]. However, the intention in all these accounts is quite different to that here.

It was suggested above that coherence should form the centrepiece of a deep theory of cooperative action; the next step is therefore to investigate how the concept fits into a general framework for describing social actions.

Team Action. This type of social action is characterised by a strong pattern of mental states loosely corresponding to the joint-intentions of [7]. The participants in a TeamAction must have knowledge of the group that is acting, with mutual beliefs in the group about aspects of the action. Team action also requires that participants *care* about the status of the team activity, and in particular must ensure that the group is kept informed about their beliefs concerning its likely outcome. Informally, team action can be defined as follows.

Team action: Group g perform team action α with respect to φ iff g have a joint intention of performing the coherent social action α with respect to φ.

Team action thus requires joint intention, which in turn requires some subsidiary concepts. (These definitions are adapted from [7].)

First, we write $(\mathsf{W\text{-}Goal}\ g\ \varphi)$ iff every agent in group g either (i) believes that φ is not true, but that φ is possible, and has a goal that φ is eventually true, or (ii) believes that φ is true, and has a goal that this becomes mutually believed in g, or (iii) believes that φ is impossible, and has a goal that this becomes mutually believed in g.

$$\begin{aligned}(\mathsf{W\text{-}Goal}\ g\ \varphi) \stackrel{\text{def}}{=}\ & \forall i \cdot (i \in g) \Rightarrow \\ & (((\mathsf{Bel}\ i\ \neg\varphi) \wedge (\mathsf{Bel}\ i\ \mathsf{E}\Diamond\varphi) \wedge (\mathsf{Goal}\ i\ \mathsf{A}\Diamond\varphi)) \vee \\ & ((\mathsf{Bel}\ i\ \varphi) \wedge (\mathsf{Goal}\ i\ \mathsf{A}\Diamond(\mathsf{M\text{-}Bel}\ g\ \varphi))) \vee \\ & ((\mathsf{Bel}\ i\ \mathsf{A}\Box\neg\varphi) \wedge \\ & (\mathsf{Goal}\ i\ \mathsf{A}\Diamond(\mathsf{M\text{-}Bel}\ g\ \mathsf{A}\Box\neg\varphi))))\end{aligned}$$

$(\mathsf{W\text{-}M\text{-}Goal}\ g\ \varphi)$ means it is mutually believed in g that g have a weak goal (W-Goal) of φ.

$$(\mathsf{W\text{-}M\text{-}Goal}\ g\ \varphi) \stackrel{\text{def}}{=} (\mathsf{M\text{-}Bel}\ g\ (\mathsf{W\text{-}Goal}\ g\ \varphi))$$

We write $(\mathsf{J\text{-}P\text{-}Goal}\ g\ \varphi)$ iff (i) it is mutually believed in g that φ is not true, and (ii) g have a mutual goal of φ, and (iii) until it is mutually believed in g that either φ is true or φ is impossible, g have a weak mutual goal (W-M-Goal) of φ.

$$\begin{aligned}(\mathsf{J\text{-}P\text{-}Goal}\ g\ \varphi) \stackrel{\text{def}}{=}\ & (\mathsf{M\text{-}Bel}\ g\ \neg\varphi) \wedge (\mathsf{M\text{-}Goal}\ g\ \mathsf{A}\Diamond\varphi) \wedge \\ & ((\mathsf{W\text{-}M\text{-}Goal}\ g\ \varphi)\ \mathcal{U} \\ & ((\mathsf{M\text{-}Bel}\ g\ \varphi) \vee (\mathsf{M\text{-}Bel}\ g\ \mathsf{A}\Box\neg\varphi)))\end{aligned}$$

Finally, we write $(\mathsf{J\text{-}Intend}\ g\ \alpha)$ iff (i) it is mutually believed in g that g are the agents of α, and (ii) g have a joint persistent goal (J-P-Goal) that eventually, g mutually believe they are about to do α, and then α happens.

$$\begin{aligned}(\mathsf{J\text{-}Intend}\ g\ \alpha) \stackrel{\text{def}}{=}\ & \\ & (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Agts}\ \alpha\ g)) \wedge \\ & (\mathsf{J\text{-}P\text{-}Goal}\ g\ \mathsf{A}\Diamond(\mathsf{Happens}\ (\mathsf{M\text{-}Bel}\ g\ \mathsf{A}(\mathsf{Happens}\ \alpha))?;\alpha))\end{aligned}$$

The conditions of satisfaction for the performance of a team action α by group g with respect to φ, written $(\mathsf{TeamAction}\ g\ \alpha\ \varphi)$, are then:

$$(\mathsf{TeamAction}\ g\ \alpha\ \varphi) \stackrel{\text{def}}{=} (\mathsf{J\text{-}Intend}\ g\ \mathsf{A}(\mathsf{Happens}\ (\mathsf{CSA}\ g\ \alpha\ \varphi)?))$$

This definition ensures that the group not only intend to do the action, they intend to do it *coherently*. Thus, if any agent in g no longer believes that (i) the action will be performed, or (ii) that the action no longer achieves the goal, or (iii) that the action is no longer minimal, in that it includes some redundant actions, then that agent must have a goal of bringing this to the attention of the group. Thus TeamAction is much stronger than simply jointly intending to do the action.

4 REMARKS

This paper began by proposing that team action theories, typically based on some model of collective intention, fail to account for many important and interesting examples of cooperative action that occur in real social systems. It was argued that a deep theory of social action should account for the whole range of social action types in the same general framework; coherence was proposed as a principle that unifies different types of social action. A discussion and subsequent formalisation of coherence was then presented. This formalisation was used as the starting point from which to develop a new formalisation of team action. The framework in which the formalisations were presented was a new quantified multi-modal logic, which was rigorously defined in §2.

Future work will focus on two key areas. First, the logic presented in §2 will be extended, for example to allow quantification over complex actions. Secondly, the model of coherence will be refined, in order to make it more fine grained. Our ultimate goal is to formalise a model of coherence that can be directly used in the analysis of implemented DAI systems.

ACKNOWLEDGEMENTS

Thanks to Afsaneh Haddadi, Nick Jennings, Daniel Mack, and the anonymous refereesfor helpful comments.

REFERENCES

[1] A. H. Bond and L. Gasser, editors. *Readings in Distributed Artificial Intelligence*. Morgan Kaufmann Publishers, Inc., 1988.

[2] P. Cohen and H. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[3] E. A. Emerson and J. Y. Halpern. 'Sometimes' and 'not never' revisited: on branching time versus linear time temporal logic. *Journal of the ACM*, 33(1), 1986.

[4] J. Y. Halpern. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3), 1990.

[5] D. Harel. Dynamic logic. In D. Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic Volume II — Extensions of Classical Logic*, pages 497–604. D. Reidel Publishing Company, 1984. (Synthese library Volume 164).

[6] N. Jennings. Commitments and conventions: The foundation of coordination in multi-agent systems. *Knowledge Engineering Review*, 8(3), 1993.

[7] H. Levesque, P. Cohen, and J. H. T. Nunes. On acting together. In *Proceedings of the National Conference on Artificial Intelligence (AAAI '90)*, Boston, MA, 1990.

[8] A. S. Rao and M. P. Georgeff. Social plans: Preliminary report. In *Decentralized AI 3 — Proceedings of the Third European Workshop on Modelling Autonomous Agents and Multi-Agent Worlds (MAAMAW '91)*. Elsevier/North Holland, 1992.

[9] S. Rosenschein and L. Kaelbling. The synthesis of digital machines with provable epistemic properties. In J. Y. Halpern, editor, *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning About Knowledge*. Morgan Kaufmann Publishers, Inc., 1986.

[10] J. R. Searle. Collective intentions and actions. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*. Bradford Books/MIT Press, 1990.

[11] M. P. Singh. Group intentions. In *Proceedings of the 10th International Workshop on Distributed A.I.*, 1990.

[12] L. Steels. Cooperation between distributed agents through self organization. In Y. Demazeau and J. P. Muller, editors, *Decentralized AI — Proceedings of the First European Workshop on Modelling Autonomous Agents in Multi-Agent Worlds (MAAMAW '89)*. Elsevier/North Holland, 1990.

[13] E. Werner. What can agents do together: A semantics of co-operative ability. In *Proceedings of the Ninth European Conference on Artificial Intelligence (ECAI '90)*. Pitman, 1990.

[14] M. Wooldridge and M. Fisher. A first-order branching time logic of multi-agent systems. In *Proceedings of the Tenth European Conference on Artificial Intelligence (ECAI '92)*, pages 234–238. John Wiley & Sons, August 1992.