This paper was published in: *International Journal of Cooperative Information Systems* Vol. 6, No. 1 (1997) 3-26.

THE STRUCTURE AND VERIFICATION OF PLAN-BASED JOINT INTENTIONS

WOLFGANG BALZER
University of Munich, LMU, MCMP, Germany

RAIMO TUOMELA
Academy Professor, Academy of Finland
Department of Philosophy, Helsinki, Finland

Tuomela's philosophical account of joint intentions is formalized in a special setting in which fully specified plans are available for the execution of the intended joint action. Using additional modal logical assumptions the definition is simplified and used to investigate how the presence of a joint intention can be efficiently checked.

Keywords: joint intention, formal model, plan, structure.

1. Introduction

Joint intentions have recently become an increasingly important topic in various fields. Intentional joint action, especially cooperation, among autonomous agents is not even conceptually possible without joint intention. Philosophers have previously analyzed joint intentions and related notions (Bratman 1993), (Searle 1990), (Tuomela 1984), (Tuomela and Miller 1988), (Tuomela 1995). Even if some technical work on single-agent actions and intentions has been done earlier, the formal modelling of joint intentions has started only recently (Levesque, Cohen and Nunes 1990), (Rao, Georgeff and Sonenberg, 1992), (Wooldridge and Jennings, 1994). While the philosophical accounts have made progress in bringing out the essential features of joint intention in a natural language setting, the formal accounts, while of course more precise, have mainly been restricted to simple, strongly idealized aspects of joint intention.

We will begin with Tuomela's analysis of joint intention (Tuomela 1995).[1] This analysis has the advantage of emphasizing the notion of an actor's intention

[1] There are weaker notions of collective intention, but we shall not discuss them in this paper.

to perform his part of a joint action as a key notion in the concept of joint intention – a feature crucial for formal analysis but largely absent in other approaches.

Our first main goal is to formalize Tuomela's basic notion of a plan-based joint intention – which is the most central notion for Computer applications – for the special case of completely specified plans. We aim at a simplification of Tuomela's[2] two-tier account which characterizes joint intentions essentially in terms of the individuals' we-intentions. This is achieved by adding some weak, natural assumptions on intentions and beliefs on the basis of which we can break down we-intentions to their constituent clauses. This is not meant to indicate, however, that joint inten tions may be conceptually reduced to individual intentions and beliefs. It Secms that conceptual reduction is not possible, primarily due to the notion of a part of a joint action. Though there is a conceptual connection between the concept of the intention to perform one's part of a joint action and the concept of the joint intention to perform the joint action, ontically nothing over and above individuals and their interrelations is needed.

Our second main goal is to use the precise characterization of plan-based joint in tentions in Order to investigate how the presence of joint intentions can be checked. In Sec. 5 we give the first such analysis for the case of joint intentions based on completely specified plans. Checking the presence of joint intentions is an essential ingredient of Cooperation and teamwork and thus is of utmost importance, both theoretically as well as for practical application, say, in communities of robots. Our results in Sec. 5 may be used as a specification for the development of programs which check the presence of joint intentions on the basis of individual beliefs, inten tions, causality, and meaning relations.

The present account differs from each of the three formal approaches mentioned above. First, while our account focusses on the notion of a given plan none of the other formal approaches uses plans as a central ingredient?[3] Second, reference to a plan provides formal access to the notion of 'a person's part' of a joint action which is missing in the other approaches, but central in the present one. Third, the mentioned authors all work with a possible world semantics including elements of a logic of time. At the present stage, however, we do not want to settle on a particular modal logic. Rather, we want to be free to consider different such logics and study their bearing on how joint intentions are checked and are built up. As the present analysis does not go to the full depth of theorem proving, and also for reasons of simplicity we do not explicitly introduce a possible worlds frame.[4] With an appropriate choice of a logic our definition of models with joint intention (D6) could be, we think, incorporated in a full possible worlds account. This would not by itself, however, solve the problems arising when time is made explicit. In view of Sec. 5 these problems are far from trivial, and cannot be addressed here. In particular, we have nothing to

[2] There are weaker notions of collective intention, but we shall not discuss them in this paper.

[3] Tuomela's and Bratman's informal analyses use the notions of plans and subplans, though.

[4] The axioms (A1), (A2), (A3), (A5) in Sec. 4 thus are adopted in a somewhat 'eclectic' way without modal logical backing.

say here about the persistence of (joint) intentions. Finally, in (Levesque, Cohen and Nunes 1990) and (Wooldridge and Jennings 1994) individual intentions are analyzed according to Cohen and Levesque's by now well known formula 'intention is choice with commitment', and stress is laid on commitment such that individual intentions are mainly represented by individual commitments. Our account, taking individual intentions.[5]

We start in Sec. 2 by restating and formalizing the definitions from (Tuomela 1995). In Sec. 3 we describe a formal framework tailored for the analysis of joint intentions. In Sec. 4 we look more closely into the relations of intentions, beliefs and mutual beliefs which leads to some simplification of the definitions of Sec. 2, and present the final formal definition of plan based-joint intention. In Sec. 5, we use this definition for a first analysis of how to check the presence of joint intentions.

2. The General Notion of Joint Intention

We base our analysis on previous work presented in (Tuomela 1995) but here we will concentrate on 'actual' we-intentions directed at a joint action, as contrasted with other forms, like standing we-intentions or we-intentions to bring about a future state.[6] Also, we will suppress a number of differentiations present in Tuomela's original account.

Consider a set $I = \{t_1, ..., i_n\}$ of individuals each of which is endowed with the capability of believing and intending. The content of an individual's belief or intention is expressed by means of a proposition. We use the expressions 'individual i believes that a'[7] and 'individual i intends to do a' in which p and a respectively denote a suitable proposition expressing the content of i's belief and the action intended. The content of the joint action which the individuals jointly intend to do also can be described by a proposition. In this paper we will

[5] The issue of 'persistency' of intentions and joint intentions thus remains implicit. Whether and how Cohen and Levesque's explicit account of that issue can be added to the present, plan-based approach, is an open question at the moment. Other accounts of joint intention (especially philosophical ones) are discussed in (Tuomela 1995). As the approach of this paper is formal rather than conceptual and as the extant philosophical accounts are non-formal it is not important here to discuss them. The central elements that distinguish our approach conceptually from accounts of e.g. (Bratman 1993) and (Searle 1990) accounts are that – in contrast to them – we are here analyzing plan-based (or, as one can arguably equivalently say, agreement-based) joint intentions and use the notion of intending to do one's part of a joint action (or plan) as a key concept. Our notion of a plan-based joint intention is a strong notion of joint intention. There are viable weaker notions, but they will not be discussed in the present paper. The only other analysis making use of the notion of a joint or shared plan of action that we are aware of is the formal account by (Rao, Georgeff and Sonenberg 1992). They do not either make serious use of the notion of one's intention to perform one's part of a joint action. As their approach using possible worlds semantics is formally very different, we cannot here make comparisons with their account. We cannot either here comment in detail any other formal accounts within AI – all of such accounts we are aware of also use possible worlds semantics.

[6] For a full analysis including standing we-intentions, see (Tuomela 1995), especially Chap. 3.

[7] 'Belief' may also refer to a disposition to acquire a belief, cf. (Audi 1982).

assume that the joint action is already decomposed such that for each individual there is a unique, agreed-upon 'part' of the joint action which that individual has to perform as an individual action.[8] We therefore can speak about 'the' part which each individual has to perform, 'his' or 'her' part of the joint action. We understand this decorn- position such that it belongs to a person's part to inform the other actors when he has done his part, when he comes to believe that it is impossible to do it, or when he finds that some presupposition of the whole enterprise is no longer satisfied.

In (Tuomela 1995), the definition of 'individuals $i_1, ..., i_n$' have the joint intention to perform a joint action w' is stated in two steps. First, individual we-intentions are characterized by (WI) below, and second, these are used to dehne joint intentions in (JI) below, $(J1\text{-iii})$ being treated as an underlying condition.

(WI) A member i of I we-intends to do w iff, based on an agreement to perform w jointly made by the members of I;
(i) i intends to do his part of w;
(ii) i has a belief to the effect that the joint action opportunities for an intentional performance of w will obtain;
(iii) i believes that there is mutual belief among the members of I to the effect that the joint action opportunities for an intentional performance of w will obtain.

(JI) Individuals $i_1, ..., i_n$ have the *joint intention to perform the joint action* w iff
(i) each $i_\sigma, \sigma \in \{1, ..., n\}$, we-intends to do w;
(ii) there is mutual belief in I to the effect that (i).
(iii) for each $\sigma \in \{1, ..., n\}$ $(WI\text{-i})$ holds for i_σ in part because of $(WI\text{-ii})$ and $(WI\text{-iii})$.

It is required that the beliefs $(WI\text{-ii})$ and $(WI\text{-iii})$ not be idle: the agent cannot intend to perform his part somehow 'accidentally' without its being based on $(WI\text{-ii})$ and $(WI\text{-iii})$. This is expressed in $(JI\text{-iii})$.

When there is a joint intention in I, then at the individual level each individual has the individual intention to perform her part. This intention implies the person's commitment to do her part, and, as the 'part' is conceived by the individual as being part of a joint action, also implies that the individual is committed to the joint action (although he of course is not committed to perform it alone by his action). Joint intentions involve joint commitments toward a joint action. A joint action is something that does not come about before all the participants have done their parts (in some cases a part need not involve any actual performance of anything). After they have done their parts the joint action comes about provided that 'the world cooperates' (viz. there are no external obstacles). A joint action has the character that if it is performed (or comes about or the 'action predicate is satisfied') then it is performed or 'satisfied' for

[8] In general, this decomposition is one of the most difficult features of the notion of joint action.

all the participants.[9] Accordingly, the participants are jointly committed to the joint action until it has been performed by them or some escape clause applies (for instance, they come to mutually believe that the joint action is impossible). If that commitment is carried out or is fulfilled for one of the participants it is fulfilled for all (collectively taken) and for everyone. Thus one is committed not only to perform one's preassigned part but also to 'follow up' the situation until the joint action has been performed (or the mentioned kind of escape clause applies), and the joint commitment entails the prima facie commitment to helping others to perform their parts of the joint action if needed.

Suppose Mary and John have agreed to clean their backyard jointly on Monday night. As a consequence they have the joint intention to do it. Each of them has the intention to do his (or her) part of the joint cleaning. They may have agreed on their parts in advance or left it for the occasion. Each of them is committed to the joint cleaning, indeed to the yard's becoming clean and to their cleaning it jointly. We can speak of a joint commitment here: Mary and John are both committed to the joint project and – hence – to performing their parts of it and to helping each other if needed. Their intentions to perform their parts of the joint project of course must be based on their mutual, communication-based understanding of the Situation in question: This involves that they share the joint plan (and joint intention) to clean the backyard on Monday night. Thus their intentions to perform their parts rely on their beliefs that the joint action opportunities will obtain and they must believe that these beliefs are shared, viz. that there is a mutual belief among them that the joint action opportunities concerning backyard cleaning obtain. Clauses (ii) and (iii) of (WI) must accordingly obtain in addition to the obvious requirement , and (i) must in part be grounded on (ii) and (iii) in Order not to be idle or (i) accidental.

In this paper we will deal with the special case of plan-based joint intentions. For this special case we will now formalize the above definitions. In (Tuomela 1995), it is argued that plan-based joint intention is at least extensionally equivalent to agreement- based joint intention. We therefore need not attempt to spell out the meaning of the phrase 'based on an agreement to perform w jointly made by the members of I' used in (WI).

Plans are related to the notions used in the above definitions in two ways. First, for each plan p there is a specification of its parts which can be performed by single individuals. This specification we capture by a partial function hispart expressing that in plan p action a is individual i's part. We write $hispart(i,p) = a$ ('$hispart$ of i in p is a' in the jargon of computer applicants). If i has no role to play in p then hispart is not defined for the pair $\langle i,p\rangle$. With respect to a, possibly large, set I of individuals and a given plan p, $hispart$ may be used to determine those individuals which are involved in the plan p. Individual i is involved in plan p iff $hispart(i,p)$ is defined. We assume that a fully specific plan contains fully specific 'stopping conditions' or revocability conditions. Whenever an individual action implies the success or the failure of the joint action the plan specifies a corresponding reaction (for instance, an Information) on the

[9] Cf. (Tuomela 1996).

individual's side as part of that individual's *hispart*. Moreover, it cannot be overemphasized that *hispart* is not a notion referring only to one individual. Even purely syntactically, *hispart* refers to plan p which is an entity involving many individuals. Consequently, as already noted, an individual intention to do $hispart(i, p')$ may well be related in its content to the joint plan.

Second, in connection with the joint action opportunities we will have to express that an individual i believes a plan p to be feasible: 'i believes that p is feasible'. This yields a simple way to deal with the phrase 'the joint action opportunities obtain' used in the above definitions when the joint action is specified by a plan. In this case, the phrase means 'the joint action opportunities for plan p obtain', which means nothing eise than 'p is feasible'. Thus, 'i believes that the joint action opportunities for p obtain' becomes 'i believes that p is feasible'. Our assuming a fully specific plan here amounts to the idealization that all conditions required for the joint action are satisfied. We thus can avoid relativization to 'the right conditions', which is crucial in general. In this section we take the notion of a joint plan as primitive; its definition is postponed to Sec. 3 below.[10]

In order to express clause (WI-vi) above we use a notion of precondition as a binary relation among propositions writing $precon(\Phi, \Phi')$ to express that the event represented by proposition Φ is a precondition of the event represented by proposition Φ'. This relation we take to hold in the sense of counterfactual implication: $precon(\Phi, \Phi')$ means that if Φ' were the case then Φ also would be the case.[11]

The above definitions refer to mutual beliefs, so we have to dehne the expression $mubel(I, \Phi)$: 'among the members of I there is mutual belief that Φ'. We do this in the standard, iterative way but for reasons of simplicity we cut off iteration after two steps. Though we believe there are good reasons for including at least a third iteration even in practical applications, we suppress this because it does not pose any new formal problems as compared to the two step iteration, and because three steps in Sec. 3 below will lead to some quite lengthy formulas.[12]

D1 For a set I of individuals and a proposition Φ we say that *in I there is mutual belief that* Φ, or $mubel(I.\Phi)$, iff
(1) for all $k \in I$: $bel(k, \Phi)$
(2) for all $kl, \in I$: $bel(k, bel(l, \Phi))$.

We thus arrive at the following list of primitive expressions: $bel(i, \Phi), int(i, a)$, $precon(\Phi, \Phi'), mubel(I, \Phi), jointplan(p), hispart(i, p) = a$, and $bel(i, feas(p))$ with i, Φ, Φ', a, p, I standing, respectively, for: individuals, propositions, actions, plans and sets of individuals, in terms of which (WI) and (Jl) can be formalized as follows.

[10] Compare (Sandu and Tuomela 1996) for an account of joint action and group action.
[11] Compare (Lewis 1973) for an analysis of counterfactuals.
[12] As this paper concentrates on the application aspects the 'fixed point' definition of mutual believes as used, e.g., in (Colombetti 1993) and (Rao, Georgeff and Sonenberg 1992) will be avoided.

(WI^*) Let I be a set of individuals, $k \in I$, and let p be a joint plan.
k *we-intends* p wrt. I, or simply: $we\text{-}intends(k,p,I)$ iff
(1) $int(k, hispart(k,p))$
(2) $bel(k, feas(p))$
(3) $bel(k, mubel(I, feas(p)))$.

(JI^*) Let $I = \{i_1, ..., i_n\}$ be a finite set of individuals and p be a joint plan.
$\{i_1, ..., i_n\}$ *have the plan-based joint intention to carry out* p iff
(1) for all k: $k \in I \leftrightarrow \langle k,p\rangle \in Dom(hispart)$
(2) for all $k \in I$: $we\text{-}intends(k,p,I)$
(3) $mubel(I, \forall k \in I(we\text{-}intends(k,p,I)))$
(4) for all $k \in I(precon(bel(k, mubel(l, feas(p))) \wedge bel(k, feas(p)) \wedge int(k, hispart(k,p))))$.

3. A Formal Frame

We will now describe a 'minimal' formal framework[13] in which (JI^*) can be restated in a fully explicit way. In doing so we will deviate from the widely used possi- ble world constructions and develop a setting which is 'non-standard' insofar as it contains a set of syntactical entities S – a set of sentences – as a component of the semantical structures. On the one hand, this greatly simplifies the formalism. On the other hand we think it is more natural – and in social science ultimately unavoidable – to include relations expressing a person's belief in a sentence (and other similar relations) directly in the structures and models.[14]

Sentences are used to represent both actions and the contents of intentions. We keep a formal distinction, however, between those sentences representing the actions which are of interest here, and 'ordinary' sentences as defined in D3 below. Theoretically, this distinction prevents our formal system from becoming highly circular; without it the System would allow for quantification over formal sentences. For practical applications, the distinction provides a well specified environment for 'plugging in' systems of action categories.[15] Such systems can be 'superimposed' on the structure $\langle A, \leq, \neg\rangle$ used in D4-4 below which represents the proper actions as performed by the individuals pursuing a plan. Thus both the sets A and S introduced below contain sentences, but only those occurring in S are formally specified. That elements of A are sentences is left as a matter for the informal Interpretation of A.

As a device used both in syntactical and semantical construction we introduce the notion of n-construction schemes for $n \in \mathbb{N}$.

D2 (a) (1) For all $\sigma \leq n$: $[\sigma]$ is an n-construction scheme.
(2) If f, g are n-construction schemes then so are $(f \parallel g)$ and $(f;g)$.
(b) Let J, A be non-empty sets. p is a *formal plan*[16] (in J and A) iff there

[13] Compare Refs. 11, 14 and 20 for alternative settings.
[14] Compare Ref. 12 as an alternative.
[15] Compare, for instance, Refs. 2 and 6 for such Systems.
[16] This is admittedly a very narrow notion of a plan which is used here only for reasons of simplicity.

exist $f, i_1, ..., i_n, a_1, ..., a_n$ such that
(1) f is a n-construction scheme
(2) $i_1, ..., i_n \in J$ and $a_1, ..., a_n \in A$
(3) $p = \langle f, i_1, a_1, ..., i_n, a_n \rangle$.

Here $\|$ and ; are taken from dynamic logics,[17] and in their later application to actions a, b are read as follows. $a \| b$: 'a and b are performed in parallel', $a;b$ reads: 'first a, then b'.

D3 The formal language L we need consists of the following:

- a set Var_1 of variables of sort 1 (ranging over individuals); syntactical variables: ξ, ξ_i
- a set Var_2 of variables of sort 2 (ranging over actions), syntactical variables: $\alpha, \alpha', \alpha_i$
- the usual logical and auxiliary symbols, including identity and pointed brackets
- the symbols **bel**, **int**, **can**, **pres**, **com**, **precon**, **feas**.

From these items we define

- *plan expressions*: A plan expression is any expression of the form $\langle f, \xi_1, \alpha_1, ..., \xi_n, \alpha_n \rangle$ where f is a n-construction schema
- *formulas*:
  (1) if u, v are variables of the same sort then $u = v$ is a formula
  (2) if ξ is a variable of sort 1 and α, α' are variables of sort 2 then $\mathbf{can}(\xi, \alpha)$, $\mathbf{com}(\alpha, \alpha')$, $\mathbf{pres}(\alpha, \alpha')$ and $\mathbf{int}(\xi, \alpha)$ are formulas
  (3) for every plan expression **e**, $\mathbf{feas}(\mathbf{e})$ is a formula
  (4) if Φ, Φ' are formulas then $\neg\Phi$, $\Phi \wedge \Phi'$, $\Phi \vee \Phi'$, $\Phi \rightarrow \Phi'$, and $\Phi \leftrightarrow \Phi'$ are formulas
  (5) if Φ is a formula and u is a variable of sort 1 or sort 2 then $\exists u \Phi(u)$ and $\forall u \Phi(u)$ are formulas
  (6) if Φ, Φ' are formulas and ξ is a variable of sort 1 then $\mathbf{bel}(\xi, \Phi)$ and $\mathbf{precon}(\Phi, \Phi')$ are formulas.

By S we denote the set of all closed formulas (sentences) of language L.

A plan expression $\langle f, \xi_1, \alpha_1, ..., \xi_n, \alpha_n \rangle$ is a schema that will be applied to sequences $\langle i_1, a_1, ..., i_n, a_n \rangle$ in which in $\langle i_\sigma, a_\sigma \rangle$ expresses that individual can perform the action $a_\sigma, \sigma \leq n$. Starting from such a sequence the plan expression yields a way for successively producing more complex actions in terms of $\|$ and ; according to the definition of the n-construction scheme f (Sec D5-a below). The sequence of all these actions plus the way they are put together (as captured by f), and the explicit list of the actors involved may be taken as a plan of the most primitive kind.

[17] Sec, for instance (Harel 1984).

The meaning of **bel**, **int**, **precon** was already explained in Sec. 2. **can** is used to express that an individual actually can perform an action. This term is interpreted as a strong success notion. When a person intends to do something and can do it then she will successfully do it, viz. not only try to do it but succeed in doing it. In a robot, can is simply stored by means of atomic sentences in the knowledge base, pres is a notion of factual presupposition among actions, which will be used in the definition of feasible plans (D5-b below). $\mathbf{pres}(\alpha_1, \alpha_2)$ is read 'α_1 presupposes α_2'. We will require that a sequence $(a; b)$ of actions is feasible in a plan only if b 'relies on' a, or presupposes a in a certain, minimal, factual[18] sense. Without this condition there would be no reason to put a and b in sequential order in the plan. They would be independent, and might as well be performed in parallel. A slightly weaker Interpretation of 'a presupposes b' is in counterfactual terms: 'if a were the case then b would be the case'. In a similar way the notion of compatibility of actions, represented by com, will serve to formulate a condition of factual compatibility for the parallel execution of two actions a, b in a plan (Sec D5-b below). $\mathbf{com}(\alpha, \alpha')$ is read 'α and α' are factually compatible' in the sense that they both can be performed without hindering each other or, more precisely, that their conjunction is possible. Finally, $\mathbf{feas}(\mathbf{e})$ expresses that the plan represented by plan expression $\mathbf{e}$ is feasible. Feasibility will be explicitly defined in D5-b below.

The structures in which this language may be interpreted are defined as follows.

D4 x is a *structure allowing for plan-based joint intentions* iff there exist $J, A, S, B, \leq, \neg, \|, ;, bel, int, can, pres, com, precon$ such that $x = \langle J, A, S, B, <, \|, ;, bel, int, can, pres, com, precon\rangle$ and

(1) J is a non-empty, finite set (of persons)
(2) A is a non-empty set (of actions)
(3) $\leq \subseteq A \times A$ ('implication on conceptual grounds'), and $\neg : A \to A$ is a partial function ('negation')
(4) $\langle A, \leq, \neg\rangle$ is an atomic *poset* with negation, and infima[19]
(5) S is the set of all sentences of L as defined in D3
(6) B is the set of all formal plans in J and A (Sec D2-b)
(7) $\|$ and $;$ are functions from $A \times A \to A$
(8) $int \subseteq J \times A$, $can \subseteq J \times A$, $pres \subseteq A \times A$, $com \subseteq A \times A$
(9) $bel \subseteq J \times S$, $precon \subseteq S \times S$
(10) for all $a, b \in A$: if $com(a, b)$ then none of the following hold if $\neg a$ respectively $\neg b$ are defined: $(a \leq \neg b)$, $(b \leq \neg a)$, $(\neg a \leq b)$, $(\neg b \leq a)$
(11) for all $a, b \in A$: if $pres(a, b)$ then not $a \leq b$.

$\leq$ is interpreted as 'implication on conceptual grounds' or 'implication in mean-

[18] 'Empirical' or, even better, 'factual' as contrasted to 'metaphysical' and 'conceptual'.

[19] That is, $\leq$ is transitive, reflexive, and anti-symmetric, for every subset $B \subseteq A$, the infimum AB of B with respect to $\leq$ exists, and $\neg$ is such that, for all $a \in A$: if $\neg a$ is defined then so is $\neg(\neg)$ and $(\neg(\neg) = a$, and for all $a, b \in A$: if $a \leq b$ and $\neg a, \neg b$ are defined then $\neg b \leq \neg a$. $0 = \vee A$ is then uniquely determined, and for each $a \in A$, $a \neq 0$ there exists an atom b in A such that $b \leq a$. b is an *atom* iff $b \neq 0$ and $\forall c \in A(0 \leq c \leq b \to c = 0$ or $c = b)$.

ing'. Action a 'implies' another action b, viz. $a \leq b$, in this sense if, whenever a is performed then, by the way language is used, also b is performed. Thus, for instance, walking implies moving, and kissing implies touching. The negation $\neg a$ of a is the action described by the negation of the sentence describing action a. As not every negated action sentence represents an action, $\neg$ cannot be defined for all $a \in A$.

The same symbols $\|$ and ; thus occur in the n-construction schemes and also are used as operators on actions, but no confusion is likely to arise from this. All the primitives of language L, except **feas**, have their obvious counterparts in a structure x, the counterpart of **feas** will now be defined (in D5-b below). As auxiliary notions we introduce the complex action $f(a_1, ..., a_n)$ constructed from given actions $a_1, ..., a_n$ by means of a n-construction scheme (D5-a). Finally, the notion of *hispart* is defined in D5-c.

D5 Let $x = \langle J, A, S, B, \leq, \neg, \|, ;, bel, int, can, pres, com, precon \rangle$ be a structure allowing for plan-based joint intentions.

(a) If f is a n-construction scheme and $a_1, ..., a_n \in A$ then $f(a_1, ..., a_n)$ is defined as follows.
(1) if $f = [\sigma]$ then $f(a_1, ..., a_n) = c$
(2) if $f = (g \parallel h)$ then $f(a_1, ..., a_n) = (g(a_1, ..., a_n) \parallel h(a_1, ..., a_n))$
(3) if $f = (g;h)$ then $f(a_1, ..., a_n) = (g(a_1, ..., a_n); h(a_1 ..., a_n))$.

(b) $feas \subseteq B$ is defined as follows. Let $p = \langle f, i_1, a_1, ..., i_n, a_n \rangle \in B$.
(1) if $f = [\sigma]$ then $feas(p)$ iff $can(i_\sigma, a_\sigma)$
(2) if $f = (g \parallel h)$ then $feas(p)$ iff $feas(g, i_1, a_1, ..., i_n, a_n) \wedge$ $feas(h, i_1, a_1, ..., i_n, a_n)) \wedge com(g(a_1, ..., a_n), h(a_i, ..., a_n))$
(3) if $f = (g;h)$ then $feas(p)$ iff $feas(g, i_1, a_1, ..., i_n, a_n) \wedge$ $feas(h(a, i_1, a_1, ..., i_n, a_n)) \wedge pres(h(a_1, ..., a_n), g(a_1, ..., a_n))$

(c) $hispart : J \times B \rightarrow A$ is defined as follows. For all i, b and $p = \langle f, i_1, a_1, ..., i_n, a_n \rangle$:
$hispart(i, p) = b$ iff $b = \wedge w$ where[20] $w = \{a_\sigma / \exists r \leq n : \langle i, a_\sigma \rangle = \langle i_r, a_r \rangle\}$.

A plan is feasible if all its 'atomic' actions aa can be performed by 'their' corresponding actor a_σ (D5-b-1), if in each step of parallel execution both parallel actions are compatible (D5-b-2), and if in each sequential step the 'later' action presupposes the 'earlier' one (D5-b-3). Obviously, if, for all $\sigma \leq n$, f contains a constituent $[\sigma]$ then $feas(\langle f, i_1, a_1, ..., i_n, a_n \rangle)$ implies $can(i_\sigma, a_\sigma)$, for all $\sigma \leq n$.

An *interpretation* $I = \langle \psi_1, \psi_2 \rangle$ of L in a structure x allowing for plan based joint intentions consists of two surjective mappings $\psi : Var_i \rightarrow J$ and $\psi' : Var_2 :\rightarrow A$. This induces a mapping I_1 assigning to each plan expression $e = \langle f, \xi_1, \alpha_1, ..., \xi_n, \alpha_n \rangle$ a corresponding formal plan

$$I_1(e) = \langle f, \psi_1(\xi_1), \psi_2(\alpha_1), ..., \psi_1(\xi_n), \psi_2(\alpha_n) \rangle.$$

Finally, *validity* of formulas ϕ in x under I ($x \models_I \Phi$) is defined as follows.

[20] $\wedge w$ denotes the infimum of w with respect to $\leq$, Sec D4-4 above.

(1) if u, v are variables of sort j, $j = 1, 2$, then $x \models_I \Phi$ iff a and v are of the same sort and, depending on the sort, their ψ_j-values are identical (j = 1,2)
(2) $x \models_I \mathbf{can}(\xi, \alpha)$ iff $can(\psi_i(\xi), \psi_2(\alpha))$, $x \models_I \mathbf{com}(\alpha, \alpha')$ iff $com(\psi_2(\alpha), \psi_2(\alpha'))$, $x \models_I \mathbf{pres}(\alpha, \alpha')$ iff $pres(\psi_2(\alpha), \psi_2(\alpha'))$, $x \models_I int(\xi, \alpha))$ iff $int(\psi_1(\xi), \psi_2(\alpha))$
(3) $x \models_I \mathbf{feas}(\mathbf{e})$ iff $feas(I_1(e))$
(4) $x \models_I \mathbf{precon}(\Phi, \Phi')$ iff $precon(\Phi, \Phi')$
(5) the usual conditions for $\wedge, \vee, \neg, \rightarrow, \leftrightarrow, \exists u \Psi(u)$ and $\forall u \Psi(u)$
(6) $x \models_I \mathbf{bel}(\xi, \Phi)$ iff $bel(\psi_1(\xi), \Phi)$.

Note that as a consequence of our particular way of dealing with sentences in clauses (4) and (6) the formulas Φ, Φ' occur on both sides. The 'interpretation' of a formula is that very formula. After these formal preparations we now come back to joint intentions.

4. Models of Plan-Based Joint Intentions

Consider the following assumptions on belief and precondition.

(A1) $bel(i, \Phi \wedge \Phi') \leftrightarrow bel(i, \Phi) \wedge bel(i, \Phi')$
(A2) $bel(i, \Phi) \leftrightarrow bel(i, bel(i, \Phi))$
(A3) $bel(i, bel(j, \Phi \wedge \Phi')) \leftrightarrow bel(i, bel_j(\Phi) \wedge bel(j, \Phi'))$
(A4) For all I, x, if $val_I(x, \Phi \leftrightarrow \Phi')$ then, for all Φ'': $val_I(x, precon(\Phi, \Phi'') \leftrightarrow precon(\Phi, \Phi''))$.

(A1) and (A2) are present, for instance, in the modal system[21] S4, (A3) is a natural 'lift' of (A1), which would follow from (A1) if i would believe $bel(j, \Phi \wedge \Phi')) \leftrightarrow bel(j, \Phi) \wedge bel(j, \Phi')$ under the commonly accepted rule

(A5) $bel(i, \Phi) \wedge bel(i, \Phi \rightarrow \Phi') \rightarrow bel(i, \Phi')$.[22]

(A4) links precondition to logical implication: equivalent propositions are equivalent as preconditions.

In the following lemmas we restrict our presentation to the simplest case of just two persons: i, j. If in mutual belief more than two iterations are considered then (A1) - (A3) have to be supplemented by corresponding assumptions for higher levels in order to preserve the following results. The lemmas are formulated with respect to a given structure $x = \langle J, A, S, B, \leq, \neg, \|, ;, bel, int, can, pres, com, precon \rangle$ allowing for plan-based joint intentions, and a fixed subset $I \subseteq J$.

Lemma 1. If (A1) and (A2) hold in x then, for all $\Phi \in S$ and $i \in I$, if $bel(i, mubel(I, \Phi))$ then $bel(i, \Phi)$.

The proofs of the lemmas and the theorem of this section are given in the appendix.

Lemma 2. If (A1), (A2), and (A4) hold in x then, for all $\Phi, \Phi' \in S$, the following are equivalent:

[21] Compare, for instance, (Chellas 1980).
[22] Sec, e.g. (Cohen and Levesque 1990), p. 231.

(i) $precon([bel(i, mubel(I, \Phi)) \wedge bel(i, \Phi)], \Phi')$
(ii) $precon(bel(i, mubel(I, \Phi)), \Phi')$.

Lemma 3. If (A1), (A2), and (A3) hold in x then, for all $\Phi, \Phi' \in S$:
$mubel(I, \Phi \wedge \Phi') \leftrightarrow mubel(I, \Phi) \wedge mubel(I, \Phi')$.

Lemma 4. If (A1) and (A2) hold in x then for all I, p, k: $we\text{-}intends(k, p, I)$ iff
(1) $int(k, hispart(k, p))$ and
(2) $bel(k, mubel(I, feas(p)))$.

We now consider a postulate with more far-reaching implications.

(P1) For all $k \in I$ and $a \in A$: $bel(k, int(k, a)) \rightarrow int(k, a)$.

(P1) says that individual k's belief about her intentions is correct. Though requiring a certain amount of rationality, this postulate is 'empirically' rnuch weaker than other assumptions used in modal logics, like for instance the basic postulate '$bel(k, \Phi)$, for every valid sentence Φ'. For later reference, we also state

(P2) For all $k \in I$ and $a \in A$: $int(k, a) \rightarrow bel(k, int(k, a))$.

Lemma 5. If (A1) - (A4) and (P1) hold in x and p satisfies JI^*-1 then
$mubel(I, \forall k \in I(we\text{-}intends(k, p, I)))$ implies $\forall k \in I(we\text{-}intends(k, p, I))$.

That is, mutual belief in we-intentions for all members already implies that such we-intentions are present.

Putting together these propositions we obtain the following, simpler characterization of plan-based joint intention.

Theorem 1. Let $x = \langle J, A, S, B, \leq, \neg, \|, ;, bel, int, can, pres, com, precon\rangle$ be a structure allowing for plan-based joint intentions which satisfies (A1) - (A4) and (P1), and let $p \in B$.

(a) In $x, i_1 ..., i_n$ have the plan-based joint intention to perform p iff
(1) for all i: $i \in I \leftrightarrow \langle i, p\rangle \in Dom(hispart)$
(2) $mubel(I, \forall k \in I(we - intends(k, p, I)))$
(3) $\forall k \in I(precon(bel(k, mubel(I, feas(p))), int(k, hispart(k, p))))$, where, for all $k \in I, we\text{-}intends(k, p, I)$ holds iff
(a) $int(k, hispart(k, p))$
(b) $bel(k, mubel(I, feas(p)))$
(c) $precon(bel(k, mubel(I, feas(p))), int(k, hispart(k, p)))$.

(b) Condition (2) in (a) is equivalent to the conjunction of the following two clauses:
$(JI1)$ $mubel(I, \forall k(int(k, hispart(k, p))))\wedge$
$(JI2)$ $mubel(I, \forall k(bel(k, mubel(I, feas(p))))))$.

In T1-b the we-intentions have been 'calculated' with the effect that mutual beliefs prevail. For a joint intention to exist there have to be mutual beliefs about (1) everybody's intending to do his part of the joint action, (2) everybody's believing that the joint action opportunities obtain, and (3) everybody's having his belief in the mutual belief that the joint action opportunities obtain

as a precondition of his intention to do his part. Condition (3) of Tl-a may be further analyzed as follows.

Lemma 6. If (A1)-(A5) and for all Φ, Φ', Φ_1: $precon(\Phi, \Phi_1) \wedge precon(\Phi', \Phi') \leftrightarrow precon(\Phi \wedge \Phi')$ holds in x, if $I = \{i, j\}$, and if f stand for '$feas(p)$' and h_k for '$int(k, hispart(k, p))$', respectively, then the expression

$(**)$ $\quad \forall k \in I$: $precon(bel(k, mubel(I, f)), h_k)$

is equivalent to the conjunction of the following

(1) $precon(b, f, h_i)$ (2) $precon(b_{ij}, f, h_i)$ (3) $precon(b_{iji}, f, h_i)$
(4) $precon(b_j, f, h_j)$ (5) $precon(b_{ji}, f, h_j)$ (6) $precon(b_{jij}, f, h_j)$

Now in a final step we define models in which a plan-based joint intention exists for the members of a subset I of individuals. We consider a possibly larger system of individuals (a structure x allowing for plan-based joint intentions), in which there is a distinguished subset I of persons which jointly intend. The joint intention for members of I involves three assumptions which must hold in a structure x allowing for plan-based joint intentions. First, we assume that requirements (A1) - (A4) and (P1) hold. Second, we assume that there is a joint plan p, a plan involving more than one individual (D6-2 below), and third, we state the conditions of Theorem 1 to define that the members of I with respect to the joint plan p have the joint intention to perform p.

D6 x is a model with joint intention for I iff there exist $J, A, S, B, \leq, \neg, \|, ;, bel, int, can, pres, com, precon$ such that
$x = \langle J, A, S, B, \leq, \neg, \|, ;, bel, int, can, pres, com, precon\rangle$ is a structure allowing for plan-based joint intentions and there exists a formal plan $p \in B$ such that

(1) assumptions (A1) - (A4) and (P1) above are satisfied
(2) $hispart(-, p)$ is defined for at least two individuals and for all $k \in J$: $\langle k, p\rangle \in Dom(hispart) \leftrightarrow k \in I$
$(J1)$ $mubel(I, \forall k \in I(int(k, hispart(k, p))))$
$(J2)$ $mubel(I, \forall k \in I(bel(k, mubel(I, feas(p)))))$
$(J3)$ $\forall k \in I$ $(precon(bel(k, mubel(I, feas(p))), int(k, hispart(k, p))))$.

Condition (2) says that p is a joint plan involving exactly the members of I.

5. The Check of Plan-Based Joint Intentions

On the basis of the previous considerations we now can turn to a closer investigation of how to check whether a plan-based joint intention is present or not. Our goal here is rather modest and practical. We want to find conditions which in applications – for instance on the computer – have to be checked in order to conclude that a joint intention is present. These conditions are not sufficient for joint intentions. They are sufficient only if further assumptions can be taken for granted. We aim at finding such 'other' assumptions which, in an 'ordinary' situation may be assumed to hold, and which therefore in applications may be

given default values. As there are rnany different possibilities of how a joint intention may develop we concentrate on one ideal-type which seems to represent the simplest case. Other patterns can be analyzed analogously. In the simplest case there is only one plan which is put forward by one individual, the other individuals having no plans of their own in the beginning. In this case the build-up of a joint intention may be– at least conceptually – broken down into the following phases.

(1) *Plan Formation*: One individual develops a joint plan and proposes this to other individuals of which she thinks they can perform some of the planned atomic actions occurring in it.

(2) *Negotiation*: The individuals which learned about the plan negotiate about which parts to perform or about changing the plan in order to achieve it's goal in a different way. 'Negotiation' actually is not a very adequate label because this kind of interaction offen involves dependence relations[23] and may range from 'real' negotiation among peers to the exertion of coercion, threat, or brute force, heavily involving a social, institutionalized background.[24] In such social settings the actors still are autonomous insofar as – ultimately – they cannot be forced to submit; they may 'prefer' to suffer heavy sanctions instead. Negotiation does not imply that the plan is accepted. Rather, from each person's point of view negotiation is part of the process of deliberation of whether to engage in the plan or not.[25] The result of this phase is that each of the persons has the plan and hispart of the plan clearly before her eyes. Although in reality there may be differences of perception we may assume for the present analysis that all persons at the end of this phase think about the same plan, and have consistent ideas about their individual parts. This marks – at least conceptually – the beginning of the phase of the build-up of a plan-based joint intention in which we are interested here.

(3) *Build-Up of a joint intention*: In this phase the mere idea of the plan which circulated on the cognitive level before must lead to changes of individual beliefs and intentions which, if 'successful', amount to the build-up of a joint intention. D6 teils us what is the result of the process. According to D6-J2 the individuals must come to believe that the plan is feasible, and that all others believe so, resulting in a mutual belief in the plan's feasibility. Second, according to D6-J1 everybody must form the intention to perform his part of the plan and the belief that everybody eise has formed such an intention, so that mutual belief about these intentions results. Third, according to D6-J3 everybody's intention to perform his part must in part be due to the mutual belief in the plan's being feasible. Of course, this is the idealized picture, but we believe that the complicated structure made explicit may well serve as a basis for less complete but more practical accounts of how the presence of a joint intention is, or can be checked. We may imagine that the build-up whose result we have clearly before

[23] See (Castelfranchi, Miceli and Cesta 1992).
[24] Compare (Balzer 1990) and (Balzer 1993).
[25] This holds at least when engagement is not coerced.

us admits of degrees, and that for each individual there is a threshold beyond which the individual may be said to have sufficient beliefs and intentions to be ready to accept the plan.

(4) *Agreement*: When all individuals have passed their threshold they can agree to perform the plan jointly. Of course, there are many forms of agreement, in particular there are cases in which not all the individuals reach their threshold but nevertheless the plan's execution is started.[26] The individuals agree explicitly or – in dependence of appropriate social rules operating in the background – implicitly to take part in the joint action. Each individual accepts the plan and we-intends to perform the joint action as planned. The build-up of their plan based joint intention is completed, and they can proceed to phase.

(5) *Implementation*: which is of no concern to us here.

Let us look more closely on phase (3) in the light of conditions $(J1)$ to $(J3)$ of D5. Again, we restrict ourselves to the case of just two individuals $I = \{i, j\}$, and we abbreviate $bel(k, w)$ by $b_k w$, $bel(k, bel(j, w))$ by $b_{kj} w$ etc. Writing out *mubel* in each condition and applying (A1) and (A2) we obtain conjunctions of expressions beginning with strings of belief operators: $b_i b_j b_i w$ (that is, $b_{iji} w$) etc. We want to Sec which of these conjuncts are really important, and how conjuncts from the different clauses $(J1)$ - $(J3)$ are related to each other. Thus we are not concerned with logical transformations but with the rather practical problem of obtaining a set of individual beliefs which under plausible conditions are sufficient for the existence of a joint intention. We proceed in two steps, first looking at each of clauses $(J1)$ - $(J3)$ in isolation, and then study the relations of conjuncts 'across' $(J1)$ - $(J3)$.

To deal with $(J1)$ let us write h_i for $int(i, hispart(i, p))$, eliminate *mubel*, and apply (A1). We end up with a conjunction of the following clauses

$$b_i h_i, b_j h_i, b_{ii} h_i, b_{ij} h_i, b_{ji} h_i, b_{jj} h_i, b_i h_j, b_j h_j, b_{ii} h_j, b_{ij} h_j, b_i h_j, b_{jj} h_j.$$

Repeated indices can be dropped by (A2), so, eliminating repetitions, we remain for one individual, say i, with the following:

$$(1)\ b_i h_i \quad (2)\ b_i h_j \quad (3)\ b_{ij} h_i \quad (4)\ b_{ij} h_j$$

Obviously, these propositions decrease in importance from left to right. So it is natural to check whether (1) – (4) hold in just this order. An arbitrary member i of I will first check whether he intends to do his part, and believes so (1), only then he will look at j and check whether j intends to do his part, that is, check (2). And only after this is done i will turn to (3) and (4). This procedure is represented in Fig.1.

[26] This is stressed in (Tuomela 1995).

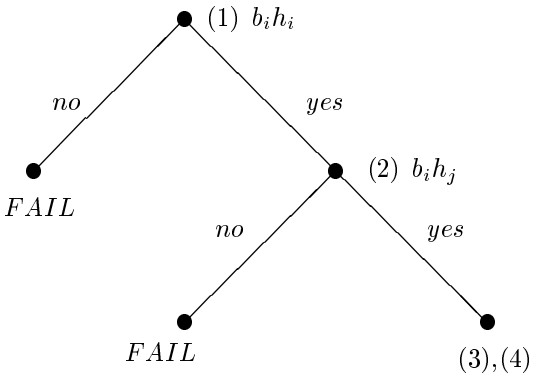

Fig. 1 An order of checks

At node (1), i checks whether b_ih_i. If this does not hold we come in the 'no'-branch; i has no intention to do his part because if he had he would believe so (by P2). This means that i does not want to take part in the joint action at all, and the joint intention fails. If b_jh_i holds we follow the 'yes'-branch to node (2) which is checked next. If b_ih_j does not hold then i does not believe that j intends to do his part, and no joint intention is developing. If b_ih_j holds we come to the next node, at which (3) and then (4) are checked.

Now in the check of (3) we can assume that b_ih_j and b_ih_j both hold, for otherwise this Checkpoint would not have been reached. Suppose (3) would not hold, that is, not: $b_{ij}h_i$. i might not believe that b_jh_i, that is, i might doubt whether j believes that he, i, will do his part. However, in the present case where b_ih_i is already established i may feel that he will be able to change j's belief in this respect, and make j believe that he, i, really intends to do his part. Moreover, i has already come to believe that j intends to do her part; so the issue of i's believing b_jh_i or not is not of major importance anyway. This suggests to assume $b_{ij}h_i$ by default, and to introduce a corresponding rule

(R1) If $b_ih_i \wedge b_{ij}h_i$[27] then $b_{ij}h_i$.

Note that if i is in doubt about h_j, that is, if b_ih_j is not present in i's knowledge base, consideration of b_jh_i may make a difference. If in this case i also fails to believe b_jh_i there is little hope for joint intention. We think, however, that clause (2) really has priority in the order of checks so that doubt about (2) cannot be overruled by (3). (4) finally may be reduced to clause (2), b_ih_j, on the assumption on i's side that j is 'normal' in believing what he intends. Then h_j will lead to b_jh_j, and this, via A5 and b_ih_j to $b_{ij}h_j$. A corresponding rule would be

(R2) If b_jh_j and b_ih_j then $b_{ij}h_j$.

Note that, in contrast to (R1), this rule requires access to j knowledge base. Its application is not under the complete control of i. Rather, the premiss b_jh_j must be 'inferred' or obtained in some other way by i.

Two special cases may be noted. First, when i is the person who originally

[27] This rule is informally contained in the participants' agreement to accept and carry out a joint plan.

proposes the plan then i will from the beginning intend to do his part. In this case the check of (1) can be omitted, and we may pass on to (2) immediately. In the second case, j is the person who proposed the plan. In this case it may be assumed by default that h_j holds and also that $b_i h_j$ holds. That is, in (2) no check is necessary. If (2) is reached we may pass on directly to (3).

Next look at $(J2)$. We write f for $feas(p)$, and we assume that all iterated versions of (A1) - (A3) hold, that is, beliefs distribute under $\wedge$, and 'double' beliefs of a person contract to 'single' ones. X-ing out (J2), omitting repetitions, and stating only beliefs of i (there is perfect symmetry in i and j) yields a conjunction of the following

(5) $b_j f$ (6) $b_{ij} f$ (7) $b_i j i f$ (8) $b_{ijij} f$ (9) $b_{ijiji} f$

As before, the order from (5) to (9) seems to be most natural (see Fig. 2.)

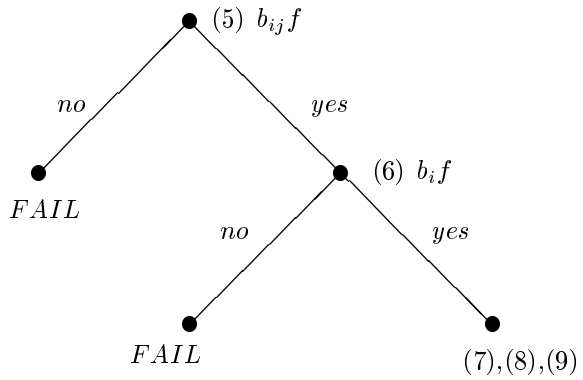

Fig. 2 An order of checks

Person i will first try to get clear about p's feasibility (f) herself, and check at node (5). If i fails to come to establish $b_i f$ then i will have no interest in executing p and no joint intention including i will arise. Only if $b_i f$ is established i will go on (through the yes-branche Fig. 2) to node (6) and check whether j believes in f. If this check has a positive result, $b_{ij} f$, the person may pass on to (7), (8), (9). (7) may be read as stating that j has a correct perception of i's belief (from $\ll i$'s perspective). Though i might believe that j has a wrong perception here, it is not easy to see how this might affect both persons' 'basic' beliefs that p is feasible which already are established by (5) and (6). This suggests to assume (7) by default. What can happen in (8) is that i may be wrong in believing that j has the right perception of him as far as f is concerned. Again, it is difficult to see how this could affect the established, 'basic' beliefs in p's being feasible, and we may assume that (8) also holds by default. And, having become tired by (7) and (8), we assume so for (9).

Things get more complicated if the check at node (6) has a negative result, that is, if i does not believe that $b_j f$. If j does not believe p to be feasible (and i believes this) j is unlikely to participate in the joint action. i's clear cut reaction would be to eliminate j which, in our idealized setting means that no

joint intention will build up. However, there may be inconsistencies here with the data about intentions. Under $J1$ $b_i h_j$ may have been established, so i believes that j intends to do her part. This is in conflict with, say, $b_i(\neg b_j f)$. If intentions are primary with respect to beliefs we could apply the rule

(R3) If $b_i h_j$ then $b_{ij} f$,

and by this overrule the presence of $b_i(\neg b_j f)$. But beliefs come in degrees, and $b_i h_j$ may have a small degree in contrast to $b_i(\neg b_j f)$. In general, therefore, it may not be good to apply rule R3, and there Secms to be no uniquely distinguished way of proceeding in this case.

Again, the two special cases should be mentioned. If i is the proponent of the plan then i will believe in f from the beginning, and we may pass from node (5) to node (6) by default. If j is the person who proposed the plan (6) may be taken to hold by default (assuming that i knows who proposed the plan). In any case, a check of (5) and (6) Secms to be sufficient for $J2$.

Under the assumptions of lemma 6 which are by no means extravagant ($J3$) reduces to the six conjuncts stated in Lemma 6. Consider the first three of them which are dealing with individual i.

(10) $precon(b_i f, h_i)$ (11) $precon(b_{ij} f, h_i)$ (12) $precon(b_{iji} f, h_i)$.

Again, the order from (10) to (12) seems natural, but now there are difficulties with the content. If person i wants to get clear about, say, (10) she has to introspect and find out how her intention h_i did develop, and whether, in particular, her belief in p's being feasible was a precondition here. We may again draw a graph like in Fig. 2 above, and have the person running through nodes (10) to (12). If all nodes yield a positive result, all conditions for joint intention are satisfied.

But what if a node fails. What if, say, i by introspection finds that his intention h_i did not develop 'partly because of' $b_i f$? Ideally, according to D6 then there is not joint intention. As before, this may conflict with other beliefs whose presence was already established. Suppose that i checked conditions $J1$. $J2$ and $J3$ in that order. Having arrived at the present stage means that $b_2 h$. Has come out positivdy in the context of $J1$. That is, i intends to do his part, and believes this. But now in $J3$ he recognizes that his intention has developed for 'other' reasons. and not at all 'because of' $b_i f$ and the other beliefs mentioned in (11) and (12). It seems that in this situation the intention already present has a much heavier weight than the way it was 'caused', and this extends to clauses (11) and (12) as well. This suggests 'overruling' (10) – (12) by the other beliefs, if these are present. If we stick to the given order it is difficult to Sec how a failure of (10) - (12) might occur if $J1$ and $J2$ have been checked with positive result, and we propose to assume that (10) to (12) hold by default in this case, that is, if the conditions are checked in the order followed here, and $J3$ is reached after all.

This does not mean that (10) to (12) could be given up as constituents of the conceptual analysis. If these conditions are not satisfied then the mutual beliefs about the plan's feasibility and about others' individual intentions are

completely irrelevant to the development of each individual's intention to perform his part. Though this may perfectly well happen we would not like to speak of a joint intention in such a case. In real-life applications it is unlikely that all beliefs required under $J1$ and $J2$ will be present, and nonetheless (10) to (12) be false at the same time. There seems room for further discussion, though.

As an alternative to assuming (10) - (12) by default we might distinguish between an individual's weak and full we-intention. i fully we-intends only if conditions (10) - (12) are all satisfied, otherwise i's we-intention is weak.

Let us turn now to the relations of conjuncts across ($J1$) - ($J3$). First, a general principle for individual intention is that what is intended must be believed to be possible. In the present context this principle establishes a link between the inten tions to do hispart, and the beliefs in the plan's feasibility. Individual i will intend to do his part only if he believes that p is feasible, and feasibility is not only a matter of i's part of p. This shows that condition ($J2$) should be checked first. If beliefs in p's feasibility are not present there is no point in checking whether the actors intend to do their parts.

Second, we can no longer look at just one person. Assuming that the individuals can communicate with each other there is the possibility of consistency check. If, say, i falls to have $b_{ij}f$ in his knowledge base he still may try to find out whether b_jf by means of communication or other kinds of check. If b_jf is present in j's knowledge base, and if i recognizes this then there is good reason for i to develop the missing belief $b_{ij}f$ and store it in his knowledge base.

A final question is about the role of the precondition relation. As this relation refers to (mutual) beliefs in f and intentions of the form h_i there is little reason to check the preconditions of such beliefs or intentions. That is, only after ($J1$) and ($J2$) have led to sufficiently positive results a check of ($J3$) can be started. As just discussed, even if the precondition relations do not hold we may speak at least of a weak joint intention. Including a check of ($J3$) therefore may at best lead to distinguishing between a weak and a full joint intention, given that ($J1$) and ($J2$) came out positively.

Combining these three features into one 'full' check for individual i which now has to take into account also parts of j's analogous procedure we obtain a process as depicted in Fig. 3. Figure 3 concentrates on individual i and includes only those parts of j's process which are necessary for a full account on i's side. As before, the exits from a node may have two values yes or no (indicated by y and n), corresponding to whether the proposition stated at the node is found in the individual's knowledge base or not. The diagonal arrows represent 'answers' of j to previous 'questions' of i (represented by the two left-right arrows to j's 'side').

Individual i first checks whether she and j believe p to be feasible. If the second check has a negative result, i 'asks' j, and in case of a positive result proceeds to check her intention h_i. If the answer is negative horizotal arrow to the feit leads to ultimate failure n. If b_ih_i is present then next b_ih_j is checked. Again, in the negative case i 'asks' j, and proceeds in case she gets a positive result. Otherwise, the diagonal arrow to the right indicates ultimate failure. Next, the preconditions are checked. If all of them hold we arrive at i's full

we-intention. Otherwise, only a weak we-intention may be stated.

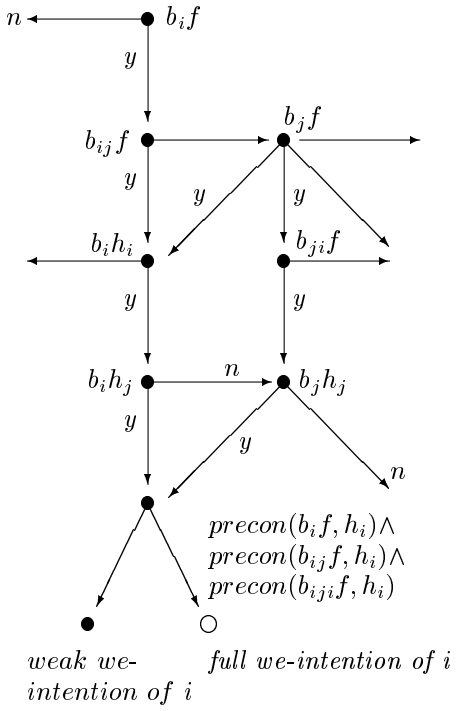

Fig. 3 Weak and full joint intention

Of course, the 'rules' (R1) - (R3) as well as other transitions we termed 'plausible' in this discussion cry out for a more systematic treatment in a logical frame binding together beliefs and intentions. In such a more comprehensive frame the 'freely floating' modal principles (A1) - (A3) or possibly other, more appropriate such principles would find their natural place. We hope to provide such a treatment in the future.

Finally, we want to point out – as rightly stated by an anonymous referee – that the present analysis does not show how the individual beliefs and intentions necessary for a joint intention come about. Intuitively, the suggestion of a plan must bring about appropriate changes of the individuals' beliefs and intentions. We mention two points that are relevant here. First, having the above process of check in mind a person proposing a plan might try to change other persons' internal states in the appropriate way and order. This may be feasible in simple situations. Second, and more realistically, however, it seems that there must be some 'social glue' which connects the agents in the right way and leads to an implicit or explicit agreement. Having a joint plan in the full sense is equivalent to such an agreement.[28] When the agents have internalized the concept of agreement they understand that they are obliged to carry out what they have agreed

[28] This is elaborated in (Tuomela 1995), Chap. 3.

on: they are committed. This is largely independent of their wants, but they must have an intention to carry out their parts. Properly accepted agreements entail intentions for appropriate actions. In the plan-language we can say that the agents all intend to carry out the plan jointly, and as a 'conceptual-causal' consequence that they intend to perform their parts of the joint action. The right understanding of the situation has the causal consequence that they form the intention to do their part (in part) because of the joint intention. The participants are jointly committed to carry out the joint action, and each individual is committed to carry out his part. Thus we can say that each individual's primary responsibility (and commitment) is to perform his part, whereas his secondary responsibility is to see to it, jointly with the others, that they indeed perform the joint action. The secondary commitment is not an intention to perform the joint action in question but it entails an intention to perform additional contributing actions when needed. We can thus say that in the full-blown acceptance of a joint plan the general idea of agreement-making – sufficiently broadly understood – gives the social glue for and explains how the intentions to perform parts of the joint action come about.

Acknowledgements

We are indebted to C. Castelfranchi and N. R. Jennings for helpful remarks on an earlier draft, and to the referees of this journal as well as those of the Modelage group for their comments.

Appendix

We abbreviate $mubel(I, feas(p))$ by B and $hispart(i, p))$ by h_i.

Proof of Lemma 1:

Suppose that $bel(i, mubel(I, \Phi))$ which, by D1 means that $bel(i, bel(i, \Phi) \wedge bel(i, bel(i, \Phi)) \wedge ...)$. By (A1) this implies $bel(i, bel(i, \Phi)) \wedge ...$, from which we obtain $bel(i, bel(i, \Phi))$ and so, by (A2), $bel(i, \Phi)$.

Proof of Lemma 2:

By Lemma 1, $bel(i, mubel(I, \Phi)) \rightarrow bel(i, \Phi)$ and so $bel(i, mubel(I, \Phi)) \leftrightarrow (bel(i, mubel(I, \Phi)) \wedge bel(i, \Phi))$, from which the lemma follows immediately by (A4).

Proof of Lemma 3:

This follows by applying (A1) - (A3) to the l.h.s.

Proof of Lemma 4:

By Lemma 1, WI^*-2 is implied by WI^*-3 and thus may be dropped.

Proof of Lemma 5:

Let $I = \{i, j\}$ and write h_k for '$int(k, hispart(k, p))$' and B for $mubel(I, feas(p))$'. By Lemma 4, the definition of $we\text{-}intends(i, p, I)$ 'reduces' to conditions (1) and (2) as stated in Lemma 4 which we will take for a definition of $we\text{-}intends(i, p, I)$ in the following. By D1-1, the assumption implies

$bel(k, \forall l \in I(we\text{-}intends(l, p, I)))$, for all $k \in I$, that is,
(1) for all $k \in I$: $bel(k, we\text{-}intends(i, p, I) \wedge we\text{-}intends(j, p, I))$. Now let $k \in I$ be given. From (1) and (A1) we obtain: $bel(k, we\text{-}intends(i, p, I)) \wedge bel(k, we\text{-}intends(j, p, I)))$, which, by the definition of $we\text{-}intends(i, p, I)$ (Lemma 4) yields: $bel(k, h_i \wedge bel(i, B)) \wedge bel(k, h_j \wedge bel(j, B))$. By A1 this yields
(2) $bel(k, h_i) \wedge bel(k, bel(i, B)) \wedge bel(k, h_j) \wedge bel(k, bel(j, B))$.

Without loss of generality suppose that $k = i$. Then (2) yields
(i) $bel(k, h_k)$ and (ii) $bel(k, bel(k, B))$.
From (i), the definition of h_k, and P1 we obtain:
(3) h_k, that is, $int(k, hispart(k, p))$. From (ii) and (A2) we obtain
(4) $bel(k, B)$. Now (3) and (4) in terms of Lemma 4 just say that $we\text{-}intends(k, p, I)$.
As k was arbitrary, we have proved: $\forall k \in I(we\text{-}intends(k, p, I))$.

Proof of Theorem 1:

(a) By Lemma 5, JI^*-3 can be dropped, by Lemma 1, WI^*-2 can be dropped, and by Lemma 2, is equivalent to clause (3) in T1-a.
(b) $mubel(I, \forall k \in I(we\text{-}intends(k, p, I)))$ iff (by Lemma 3)
$\forall k \in I(mubel(I, (we\text{-}intends(k, p, I)))$ iff (by the definition of $mubel$)
$\forall k \in I(mubel(I, (h_k \wedge bel(k, B))))$ iff (by Lemma 3)
$\forall k \in I(mubel(I, h_k) \wedge mubel(k, bel(k, B)))$ iff (by logics)
$\forall k \in I(mubel(I, h_k) \wedge \forall k \in I(mubel(I, bel(k, B))))$ iff (by Lemma 3)
$mubel(I, \forall i h_i) \wedge mubel(I, \forall i \in I(bel(i, B)))$. These are just the two clauses stated in T1-b.

Proof of Lemma 6: (0) by D1 is equivalent to
$\forall k \in I$: $precon((b_k(b_i f \wedge b_{ji} f \wedge b_{ij} f \wedge b_{ij} f \wedge ....)), h_k)$, and this, by A1 and A4, with
$\forall k \in I$: $precon((b_{ki} f \wedge b_{kii} f \wedge b_{kii} f \wedge ...), h_k)$, which by $(*)$ is equivalent to
$\forall k \in I$: $precon(b_{ki} f, h_k) \wedge precon(b_{kj} f, h_k) \wedge ... \wedge precon(b_{kjj} f, h_k)$. Setting k to i and j yields the following

(7) $precon(b_{ii} f, h_i)$	(8) $precon(b_{ij} f, h_i)$	(9) $precon(b_{iii} f, h_i$
(10) $precon(b_{iij}, h_i)$	(11) $precon(b_{iji} f, h_i)$	(12) $precon(b_{ijj} f, h_i)$
(13) $precon(b_{ji} f, h_j)$	(14) $precon(b_{jj} f, h_j)$	(15) $precon(b_{jii} f, h_j)$
(16) $precon(b_{jij} f, h_i)$	(17) $precon(b_{jji} f, h_i)$	(18) $precon(b_{jjj} f, h_j)$

Now by A1, $b_{ii} f \leftrightarrow b_i f$, so by A4: $precon(b_{ii} f, h_i) \leftrightarrow precon(b_i f, h_i)$. By the same argument clause (9) reduces to (7), and may be omitted. Similarly, clause (18) reduces to (14), and this, in turn, to $precon(b_j f, h_j)$.

By A2, $b_{iij} f \leftrightarrow b_{ij} f$, so by A4, (10) reduces to $precon(b_{ij} f, h_i)$ which is the same as (8) and thus may be omitted. By A2, $b_{jj} f \leftrightarrow b_j f$, so by A5: $b_i(b_{jj} f) \leftrightarrow b_i(b_j f)$, and with A4 (12) reduces to $precon(b_{ij} f, h_i)$ which, being identical with (8), also may be dropped. In a similar way (15) and (17) reduce to (13), and may be dropped. The remaining list of conjuncts contains exactly (1) - (6).

References

1. R. Audi (1982) Believing and Affirming, *Mind* 91, 115-20.

2. Th. Ballmer and W. Brennenstuhl (1981) *Speech Act Classification*, Springer, Berlin.

3. W. Balzer (1990) A basic model for social institutions, *Journal of Mathematical Sociology* 16, 1-29.

4. W. Balzer (1993) *Soziale Institutionen*, de Gruyter, Berlin.

5. M. Bratman (1993) Shared intention, *Ethics* 104, 97-113.

6. M. Brecher (1977) Toward a theory of international crisis behavior, *International Studies Quarterly* 21, 39-73.

7. C. Castelfranchi, M. Miceli and A. Cesta (1992) Dependence relations among autonomous agents, in Ref. 25, 215-27.

8. B. F. Chellas (1980) *Modal Logic*, Cambridge University Press, Cambridge/ Mass.

9. P. R. Cohen and H. J. Levesque (1990) Intention is choice with commitment, *Artificial Intelligence* 42, 213-61.

10. M. Colombetti (1993) Formal semantics for mutual belief, *Artificial Intelligence* 62, 341-53.

11. R. Fagin and J. Y. Halpern (1988) Belief, awareness, and limited reasoning, *Artificial Intelligence* 34, 39-76.

12. L. Gasser (1991) Social conceptions of knowledge and action: DAI foundations and open Systems semantics, *Artificial Intelligence* 47, 107-38.

13. D. Harel (1984) Dynamic logic, in :*Handbook of Philosophical Logic*, Vol. II, eds. D. Gabbay and F. Guenthner, Reidel, New York, 497-604.

14. K. Konolige, (1986) *A Deduction Model of Belief*, Pitman, London.

15. H. Levesque, P. Cohen and J. Nunes (1990) On acting together, *Proc. Eighth National Conf. on Artificial Intell.*, Vol. I, MIT Press, Cambridge/Mass., 94-9.

16. D. K. Lewis (1973) *Counterfactuals*, Cambridge University Press, Cambridge.

17. A. S. Rao, M. G. Georgeff and A. E. Sonenberg (1992) Social plans: A preliminary report, in Ref. 25, 57-76.

18. G. Sandu and R. Tuomela (1996) Joint action and group action made precise, *Synthese*, 105, 319-45.

19. J. R. Searle (1990) Collective intentions and actions, in *Intentions in Communication*, eds. P. R. Cohen et al., MIT Press, Cambridge/Mass., 401-15.

20. M. P. Singh and N. M. Asher (1993) A logic of intentions and beliefs, *Journal of Philosophical Logic* 22, 513-44.

21. R. Tuomela (1984) *A Theory of Social Action*, Reidel, Dordrecht.

22. R. Tuomela and K. Miller (1988) We-intentions, *Philosophical Studies* 53, 115-137.

23. R. Tuomela (1995) *The Importance of Us*, Stanford University Press, Stanford/Calif.

24. R. Tuomela (1996) Collective goals and Cooperation, forthcoming in: Proc. KCS'95 congress held in San Sebastian, Spain, Kluwer, Dordrecht.

25. E. Werner and Y. Demazeau (1992) *Decentralized A.I.-3*, North-Holland, Amsterdam.

26. M. J. Wooldridge and N. R. Jennings (1994) Formalizing the cooperative problem solving process, in *Proc. ISth Int. Workshop on Distributed AI*, Lake Quibault WA, 403-17.