A Representationalist Theory of Intention

Kurt Konolige*
Artificial Intelligence Center
SRI International
Menlo Park, CA 94025

Martha E. Pollack+
Dept. of Computer Science
Univ. of Pittsburgh
Pittsburgh, PA 15260

Abstract

Several formalizations of cognitive state that include intentions and beliefs based on normal modal logics (NMLs) have appeared in the recent literature. We argue that NMLs are not an appropriate representation for intention, and provide an alternative model, one that is representationalist, in the sense that its semantic objects provide a more direct representation of cognitive state of the intending agent. We argue that this approach results in a much simpler model of intention than does the use of an NML, and that, moreover, it allows us to capture interesting properties of intention that have not been addressed in previous work

1 Introduction

Formalizations of cognitive state that include intentions and beliefs have appeared in the recent literature [Cohen and Levesque, 1990a; Rao and Georgeff, 1991; Shoham, 1990; Konolige and Pollack, 1989]. With the exception of the current authors, these have all employed *normal modal logics* (NMLs), that is, logics in which the semantics of the modal operators is defined by accessibility relations over possible worlds. This is not surprising, since NMLs have proven to be a powerful tool for modeling the cognitive attitudes of belief and knowledge. However, we argue that intention and belief are very different beasts, and that NMLs are ill-suited to a formal theory of intention.

We therefore present an alternative model of intention, one that is representationalist, in the sense that its semantic objects provide a more direct representation of cognitive state of the intending agent. We argue that this approach results in a much simpler model of intention than does the use of an NML, and that, moreover, it allows us to capture interesting properties of intention that have not been addressed in previous work Further, the relation between belief and intention is mediated by the fundamental structure of the semantics, and is independent of any particular choice for temporal operators or theory of action. This gives us a very direct, simple, and semantically motivated theory, and one that can be conjoined with whatever temporal theory is appropriate for a given task.

In the next section (Section 2), we make the case for a representationalist theory of intention. Section 3 constitutes the technical heart of our paper: there we develop our formal model of intention. Finally, in Section 4, we draw some conclusions and point the way toward further development of our logic of intention.

'Supported by the Office of Naval Research under Contract No. N00014-89-C-0095.

Supported in part by a National Science Foundation Young Investigator's Award (IRI-9258392), by the Air Force Office of Scientific Research (Contract F49G20-92-J-0422), and by DARPA (Contract F30602-93-C-0038).

2 The case for representationalism

As we noted above, NMLs have been widely and successfully used in the formalization of belief. It is largely as a result of this success that researchers have adopted them in building models of intention. However, we argue in this section that these logics are inappropriate to models of intention:

- The semantic rule for normal modal operators is the wrong interpretation for intention This rule leads to the confusion of an intention to do ϕ with an intention to do any logical consequence of ϕ, called the *.side-effect problem* [Bratman, 1987]. A simple and intuitively justifiable change in the semantic rule makes intention side-effect free (and nonnormal).
- Normal modal logics do not provide a means of relating intentions to one another. Relations among intentions are necessary to describe the means-end connection between intentions.

NMLs are closed under logical consequence: given a normal modal operator L, if $L\phi$ is true, and $\phi \models \psi$, it follows that $L\psi$ is true When *L* represents belief, consequential closure can be taken to be an idealization: although it is obviously unrealistic in general to assume that an agent believes all the consequences of his beliefs, it is reasonable to assume this property of an ideal agent, and this idealization is acceptable in many instances.

However, consequential closure *cannot* be assumed for intention, even as an idealization. It is clear that an agent who intends to perform an action usually does not intend all the consequences of that action, or even all the

consequences he anticipates. Some of the consequences are goals of the agent, while others are "side effects" that the agent is not committed to.[1]

Because NMLs are subject to consequential closure, and intention is not, several strategies are used to make the logics side-effect free. They all involve relativizing the side-effects of intentions to believed consequences. The thesis of *realism* is that all of an agent's intended worlds are also belief worlds [Cohen and Levesque, 1990a], that is, a rational agent will not intend worlds that he believes are not possible. Given the realism thesis, whenever the agent intends a and believes $a \supset b$, he will also intend 6. Cohen and Levesque [Cohen and Levesque, 1990b] adopt the realism thesis, and rely on claims about way an agent may change his beliefs about the connection between an intended proposition and its consequences to make their theory side-effect free. In their case, an agent who always believes that $a \supset b$ is always true will incur the side-effect problem when intending a. Also, any analytic implication (i.e., when $a \supset b$ must be true in all possible futures) will cause problems. Two special cases are abstractions (e.g., making a dinner is an abstraction of making a spaghetti dinner) and conjunctions (intending $a \wedge b$ implies intending a and intending 6 separately).

Rao and Georgeff [Rao and Georgeff, 199l] point out that by relaxing realism, intentions can be made side-effect free. *Weak realism* is the thesis that at least one intended world is a belief world. There can thus be intention worlds that are not belief worlds. Now, even though the agent believes $a \supset b$, b is not an intention, because there is an intended world in which a is true but not b. Weak realism seems inherently less desirable than realism (how is it possible for an agent to intend worlds he does not believe possible?), and it is still not fully side-effect free, since it is closed under conjunctions and abstractions.

These problems do not mean we have to abandon possible worlds. In fact, with the right semantics, possible worlds are an intuitively satisfying way of representing future possibility and intention for an agent. We note that intentions divide the possible futures into those that the agent wants or prefers, and those he does not. Consider the diagram of Figure 1 The rectangle represents the set of possible worlds W. The *scenario* for a proposition a is the set of worlds in W that make a true: the shaded area in the diagram. An agent that has a as an intention will be content if the actual world is any one of those in the shaded area, and will be unhappy if it is any unshaded one. The division between wanted and unwanted worlds is the important concept behind scenarios. For example, consider another proposition b that is implied by a (for concreteness, take a = "1 get my tooth filled," and b — "1 feel pain") If we just look

*For example, an agent may intend to go to the dentist to get his tooth filled, believing that he will feel pain as a consequence, without being committed to feeling the pain. If he discovers that the dental work is painless, he will not seek to experience the pain nonetheless. See Bratman [Bratman, 1987] and Cohen and Levesque [Cohen and Levesque, 1990b] for further discussion.

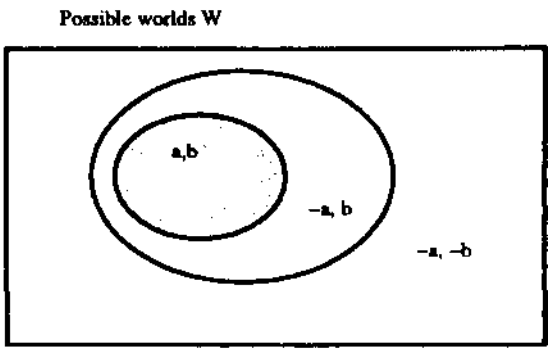

Figure 1: A Venn diagram of two scenarios

at interpretations within the shaded area, a and 6 both hold, and so cannot be distinguished. But the complement of these two propositions is different. A world in the area $\neg a, b$, in which the agent feels pain but does not have his tooth pulled, is an acceptable world for the intention 6, but not for a. So the interpretation rule for intention must take into account the complement of the intended worlds. As we will see in Section 3, this makes intention a nonnormal modal operator. It also makes it side-effect, abstraction, and conjunction free, whether we choose realism or weak realism.

The representationalist part of the model comes in representing the mental state of the agent using scenarios. *Cognitive structures,* containing elements representing intentions and the relationship among intentions, are used for this purpose.

3 Cognitive structures

Our model of intention will have two components: possible worlds that represent possible future courses of events, and *cognitive structures,* a representation of the mental state components of an agent. We introduce complications of the model in successive sections. To begin, we define the simplest model, a static representation of primary or "top-level" intentions. Primary intentions do not depend on any other intentions that the agent currently has*

3.1 Possible Futures

The concept of intention is intimately connected with choosing among course of future action. In the model, courses of action are represented by possible worlds. Each possible world is a complete history, specifying states of the world at all instants of time. We assume there is a distinguished moment *now* in all worlds that

[2]This is a bit of an overstatement, since an agent's intentions change over time, and an intention that begins life as primary may later also be used in support of some other intention. In such cases we say that the intention has been *overloaded.* Overloading is a cognitively efficient strategy for an agent to employ [Pollack, I99I]. For the moment, however, we will not worry about primary intentions that later are overloaded.

is the evaluation point for statements.[3]

The set of possible worlds is W. For each world $w \in W$, there is an evaluation function that determines the value of sentences in a language $\mathcal{L}$, which refer to states of the world or actions that take place between states of this world. For any sentence ϕ of $\mathcal{L}$, $w(\phi)$ is the truth-value of ϕ.

To talk about contingent and necessary facts $\mathcal{L}$ is extended to $\mathcal{L}_\Box$, which includes the modal operators $\Box$ and $\Diamond$. The possibility operator $\Diamond$ expresses the existence of a world with a given property. $\Diamond\phi$ says that there is a world (among W) for which ϕ is true. Its semantics is:

Definition 3.1

$$w, W \models \Diamond\phi \quad \textit{iff} \quad \exists w' \in W.\ w', W \models \phi\ .$$

$\Diamond$ is used to specify the background of physically possible worlds under which reasoning about intention takes place, and will be important in describing the structure of a given domain. The necessity operator $\Box\phi$ is defined as $\neg\Diamond\neg\phi$.

3.2 Belief and primary intentions

We begin by defining concept of scenario.

Definition 3.2 *Let W be a set of possible worlds, and ϕ any sentence of $\mathcal{L}$. A* scenario for ϕ *is the set*

$$M_\phi = \{w \in W \mid w, W \models \phi\}.$$

A scenario for ϕ identifies ϕ with the subset of W that make ϕ true.

A cognitive structure consists of the background set of worlds, and the beliefs and intentions of an agent.[4]

Definition 3.3 *A* cognitive structure *is a tuple $(W, \Sigma, \mathcal{I})$ consisting of a set of possible worlds W, a subset of W (Σ, the beliefs of the agent) and a set of scenarios over W ($\mathcal{I}$, the intentions of the agent).*

We extend $\mathcal{L}_\Box$ to $\mathcal{L}_I$ by adding the modal operators B for belief and I for intentions. The beliefs of an agent are taken to be the sentences true in all worlds of Σ.[5] For simplicity, we often write Σ as a set of sentences of $\mathcal{L}_\Box$, so that M_Σ is the corresponding possible worlds set.

Definition 3.4

$$\langle W, \Sigma, \mathcal{I}\rangle \models B(\phi) \quad \textit{iff} \quad \forall w' \in M_\Sigma.\ w', W \models \phi, \textit{ i.e., } M_\Sigma \subseteq M_\phi\ .$$

The beliefs of an agent are always possible, that is, they are a subset of the possible worlds. This also means that an agent cannot be wrong about necessary truths. A more complicated theory would distinguish an agent's beliefs about what is possible from what is actually possible. The key concept is that intentions are represented with respect to a background of beliefs about possible courses of events (represented by $\Diamond$), as well as beliefs about contingent facts (represented by B). Stated in $\mathcal{L}_I$, the following are theorems:

$$\begin{aligned} &B(\phi) \supset \Diamond\phi \\ &B(\Box\phi) \equiv \Box\phi \end{aligned} \qquad (1)$$

Of course, beliefs about contingent facts can still be false, since the real world does not have to be among the believed ones. The B operator represents all futures the agent believes might occur, including those in which he performs various actions or those in which he does nothing. The beliefs form a background of all the possibilities among which the agent can choose by acting in particular ways.

The third component of a cognitive structure for an agent, an intention structure, is a set of scenarios M_ϕ. Intuitively, an agent's intention structure will include one scenario for each of his primary intentions. We write $\mathcal{I}$ as a set of sentences of $\mathcal{L}_\Box$, where each sentence ϕ stands for its scenario M_ϕ.

Definition 3.5

$$\langle W, \Sigma, \mathcal{I}\rangle \models I(\phi) \textit{ iff } \exists \psi \in \mathcal{I} \textit{ such that } M_\psi \textit{ is a scenario for } \phi, \textit{ i.e. } M_\psi = M_\phi.$$

This definition bears an interesting relation to the semantics of normal modal operators. Each primary intention (i.e., each element of $\mathcal{I}$) acts like a separate modal operator. A normal modal operator I_ψ for the element M_ψ would be defined using:

$$\langle W, \Sigma, \mathcal{I}\rangle \models I_\psi(\phi) \quad \text{iff} \quad M_\psi \subseteq M_\phi\ ,$$

just as for belief. The semantic rule for I is similar, but uses equality between the scenarios instead of subset, so that the worlds *not* in M_ψ must satisfy $\neg\phi$. By identifying intentions with scenarios, we explicitly encode in the semantics the distinction between preferred and rejected possible worlds. If we were to use the weaker form of the semantic rule for $I(\phi)$ (i.e., $M_\psi \subseteq M_\phi$), then there could be some world w which satisfies ϕ but is not a world satisfying the agent's intention. This is contrary to our reading of intention as a preference criterion dividing possible worlds. [6]

From this formal definition, it is easy to show that $I(\phi)$ will hold just in case ϕ is equivalent to some proposition $\psi \in \mathcal{I}$, given the background structure W.

Proposition 3.1 *For any structure $\langle W, \Sigma, \mathcal{I}\rangle$,*

$$\langle W, \Sigma, \mathcal{I}\rangle \models I(\phi) \quad \textit{iff} \quad \exists \psi \in \mathcal{I}.\ W \models \Box(\phi \equiv \psi)\ .$$

3 This definition of possible worlds is the one usually used in the philosophical literature, but differs from that of Moore in [Moore, 1980], where possible worlds are identified with states at a particular time.

4In this paper, we deal only with the single agent case, and thus we neither explicitly indicate the name of the (unambiguous) agent associated with any cognitive structure, nor include an agent argument in our intention or belief predicates.

5This enforces the condition of logical omniscience [Levesque, 1984J on the agent's beliefs, which is not a realistic assumption. We could chose a different form for beliefs, say a set of sentences of $\mathcal{L}_I$ that is not closed with respect to consequence; but it would obscure the subsequent analysis.

6Our semantics is also equivalent to the minimal model semantics of Chellas [Chellas, 1980]. In the minimal model semantics, the accessibility relation is from a world to a set of sets of worlds, i.e., a set of propositions. As Chellas shows, such logics are nonnormal, and the simplest system, E, contains only the inference rule $\phi \equiv \psi / I\phi \equiv I\psi$.

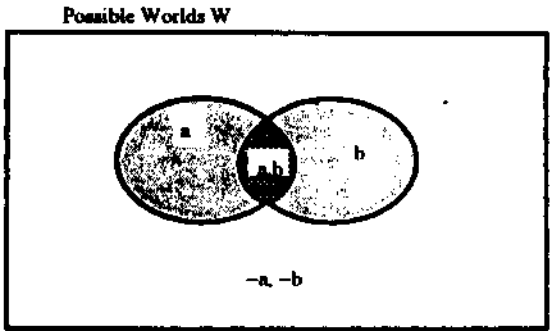

Figure 2: A Venn diagram of conjunctive scenarios

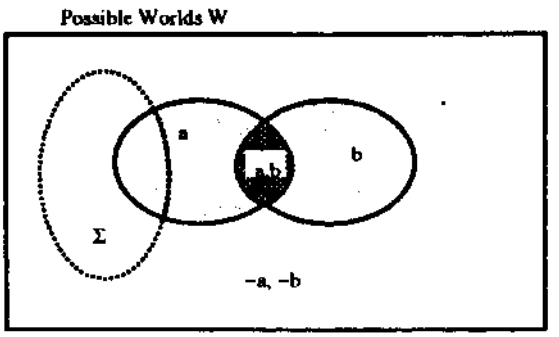

Figure 3: A Venn diagram of belief and intention.

The *I* operator is true precisely of the individual top-level intentions the agent has. It is not subject to closure under logical consequence or under the agent's beliefs. To see this, consider the cognitive structure $\langle W, \Sigma, \{a\}\rangle$, i.e., the agent has the single intention to perform *a* Assume that *a* logically implies 6, but not the converse, i.e.,

$$W \models \Box(a \supset b) \wedge \Diamond(b \wedge \neg a) .$$

Then $M_a \neq M_b$, because there is a world in which *b* is true but *a* is not From the semantics of J, we have

$$\langle W, \Sigma, \{a\}\rangle \models I(a) \wedge \neg I(b)$$

This shows that *I* is not closed with respect to valid consequence. To distinguish the intention of a from its necessary consequence b, there must, be at least one possible world in which *b* is true but *a* is not. As a particular instance of this, our theory does not equate an intention to perform a conjunction with a conjunction of intentions Assume that the set of possible worlds distinguishes *a* arid *b*, i.e., $W \models \Diamond(a \wedge \neg b) \wedge \Diamond(\neg a \wedge b)$. Now consider two agents: the first has the single primary intention $a \wedge b$, and the second has exactly the two primary intentions *a* and 6. Then:

$$\begin{aligned} \langle W, \Sigma, \{a \wedge b\}\rangle &\models \quad I(a \wedge b) \wedge \neg I(a) \wedge \neg I(b) \\ \langle W, \Sigma, \{a, b\}\rangle &\models \quad I(a) \wedge I(b) \wedge \neg I(a \wedge b) \end{aligned}$$

The reason for this is clear from the diagram of Figure 2. The scenario $M_{a\wedge b}$ excludes all interpretations outside of the overlap area in the figure; hence it is not equivalent to M_a, for which a perfectly acceptable world could contain *a* arid $\neg b$; nor is it equivalent to M_b

On the other hand, taking the two scenarios M_a and M_b singly, acceptable worlds are in the respective regions *a* and *b*. Thus the most acceptable worlds are in the overlap region. However, if one of the goals becomes impossible, say a, then any world in *b* is acceptable, unlike the case with the conjunctive scenario $M_{a\wedge b}$.

A similar story can be told for side effects and abstraction. The ability to distinguish between an intention and its side effects, abstractions, and conjunctions is basic to the semantics given in Definition 3.5, and does not require any further axioms or stipulations, nor any commitment to a particular temporal logic

An alternative to the reading of "intention" as separate primary intentions is the reading as conjoined intention, i.e., "ϕ is intended if it is the intersection of worlds of some set of primary intentions." We use the operator $I^*(\phi)$ for this reading.

Definition 3.6

$\langle W, \Sigma, \mathcal{I}\rangle \models I^*(\phi)$ *iff* $\exists J \subseteq \mathcal{I}$ *such that* M_J *is a scenario for* ϕ, *i.e.*, $M_J = M_\phi$.

I^* can be characterized by the following axioms.

Proposition 3.2 *The following are theorems of* $\mathcal{L}_I$.

$$\begin{aligned} &I(a) \supset I^*(a) \\ &I^*(a) \wedge I^*(b) \supset I^*(a \wedge b) \end{aligned}$$

So I^* sanctions the conjoining of separate intentions, but not the splitting of an intention that is a conjunction.[7]

3.3 Rationality constraints: intention and belief

So far we have not related the agent's intentions to his beliefs. Consider the diagram of Figure 3, for which the cognitive structure is $\langle W, \Sigma, \{a, b\}\rangle$. The agent's two intentions are jointly possible, since the overlapping area contains at least one world in which they both hold. However, based on the contingent facts of the situation, the agent does not believe that they will actually occur, since his beliefs, given by the set Σ, fall outside the overlap area. A rational agent will not form intentions that he does not believe can be jointly executed. Further, intentions should be nontrivial, in the sense that the agent intending ϕ should not believe that ϕ will occur without the intervening action of the agent. To enforce rationality, we define the following conditions on cognitive structures.

Definition 3.7 *A cognitive structure* $\langle W, \Sigma, \mathcal{I}\rangle$ *is* admissible *iff it is* achievable:

$$\exists w \in \Sigma.\ \forall \phi \in \mathcal{I}.\ w \in M_\phi$$

and nontrivial:

$$\forall \phi \in \mathcal{I}.\ \exists w \in \Sigma.\ w \notin M_\phi .$$

This condition leads immediately to the following consequences.

[7] In terms of Chellas' minimal models, the semantics of I^* is for models that are closed under intersection. This makes sense: I^* represents any intersection of intentions.

Proposition 3.3 *These sentences are valid in all admissible structures.*

$\neg I(a \wedge \neg a)$	*Consistency*
$I(a) \wedge I(b) \supset \Diamond(a \wedge b)$	*Joint Consistency*
$I^*(a) \supset \Diamond a$	
$I^*(a) \supset B\Diamond a$	*Realism*
$I(a) \supset \neg B(\neg a)$	*Epistemic Consistency*
$I(a) \wedge I(b) \supset \neg B(\neg(a \wedge b))$	*Joint Epistemic Consistency*
$I^*(a) \supset \neg B\neg(a)$	
$I(a) \supset \neg B(a) \wedge \neg B(\neg a)$	*Epistemic Indeterminacy*

A rational agent, characterized by achievable structures, does not believe that his joint intentions represent an impossible situation: this is the theorem of Joint Epistemic Consistency. This theorem can be stated using either reading of intention.

In addition, the nontriviality condition on models means that the agent does not believe that any one of his intentions will take place without his efforts (Epistemic Indeterminacy). Recall that the *B* operator represents all futures the agent believes might occur, including those in which he performs various actions or does nothing. The beliefs form a background of all the possibilities among which the agent can choose by acting in particular ways. If in all these worlds a fact ø obtains, it does no good for an agent to form an intention to achieve ø, even if it is an action of the agent, because it will occur without any choice on the part of the agent. So, for example, if the agent believes he will be forced to act at some future point, perhaps involuntarily (e.g., by sneezing), it is not rational for the agent to form an intention to do that.

Note that in our logic, the realism thesis is expressed using beliefs about what is possible. This is because we distinguish beliefs about contingent facts ("Nixon was president") from the background possibilities an agent believes could occur, but haven't or won't. Realism follows directly from Joint Consistency and the simplifying assumption (1) that all worlds *W* are possibilities for the agent.

In this logic, we are deliberately leaving the temporal aspects vague until they are necessary. At this level of abstraction, different kinds of goals can be treated on an equal basis. For example, goals of prevention, which are problematic for some temporal logic accounts of intention, are easily represented. For an agent to prevent a state *p* from occurring, he must believe both *p* and -p to be possible at some future state The agent's intention is the scenario consisting of worlds in which *p* is always true.

3.4 Relative intentions

As we discussed earlier, one of the primary characteristics of intentions is that they are structures: agents often form intentions relative to pre-existing intentions. That is, they "elaborate" their existing plans. There are various ways in which a plan can be elaborated. For instance, a plan that includes an action that is not directly executable can be elaborated by specifying a particular way of carrying out that action; a plan that includes a set of actions can be elaborated by imposing a temporal order on the members of the set; and a plan that includes an action involving objects whose identities are so far underspecified can be elaborated by fixing the identities of one or more of the objects. As Bratman [Bratman, 1987, p.29] notes, "[p]lans concerning ends embed plans concerning means and preliminary steps; and more general intentions ...embed more specific ones." The distinction between these two kinds of embedding recurs in the AI literature. For instance, Kautz [Kautz, 1990] identifies two relations: (1) *decomposition*, which relates a plan to another plan that constitutes a way of carrying it out (means and preliminary steps), and (2) *abstraction*, which relates a specific plan to a more general one that subsumes it. It is useful to have a term to refer to the inverse relation to abstraction: we shall speak of this as *specialization*.

Both kinds of elaboration are represented in the cognitive structure by a graph among intentions. The graph represents the means-ends structure of agent intentions. For example, suppose the agent intends to do a by doing b and c. Then the cognitive structure contains the graph fragment $M_b, M_c \rightarrow M_a$. As usual, in the cognitive structure we let the propositions stand for their associated scenarios.

Definition 3.8 *An elaborated cognitive structure consists of a cognitive structure and an embedding graph $\rightarrow$ among intentions: $(W, \Sigma, \mathcal{I}, \rightarrow)$. The graph is acyclic and rooted in the primary intentions.*

Remarks. The reason we need both primary intentions and the graph structure is that, while every root of the graph must be a primary intention, primary intentions can also serve as subordinate intentions. Consider the masochistic agent with a tooth cavity: he both intends to feel pain, and intends to get his tooth filled. His cognitive structure would be:

$$\{W, \{a \supset b\}, \{a, b\}, a \rightarrow b\} .$$

Also note that a scenario of the graph may serve to elaborate more than one intention; Pollack [Pollack, 1991] calls this overloading.

The embedding graph $\rightarrow$ is the most strongly representationalist feature of the model. It represents the structure of intentions in a direct way, by means of a relation among the relevant scenarios. A normal modal logic is incapable of this, because its accessibility relation goes from a single world (rather than a scenario) to a set of possible worlds.

In the language, $\mathcal{L}_I$ is extended to include a modal operator $By(\alpha; \beta_1, \cdots \beta_n)$, where the β_i together are an elaboration of α.

Definition 3.9

$$(W, \Sigma, \mathcal{I}, \rightarrow) \models By(\alpha; \beta_1, \cdots \beta_n) \quad \textit{iff} \quad \beta_1, \cdots \beta_n \rightarrow \alpha$$

For rational agents, intention elaborations will have the same properties vis-a-vis belief as top-level intentions. So, in admissible structures we insist on the condition that any scenario of $\rightarrow$ is part of the achievable and nontrivial intentions.

Definition 3.10 *A cognitive structure $(W, \Sigma, \mathcal{I})$ is* admissible *iff it is* achievable*:*

$$\exists w \in \Sigma. \forall \phi \in (\mathcal{I} \textit{ and } \rightarrow). w \in M_\phi$$

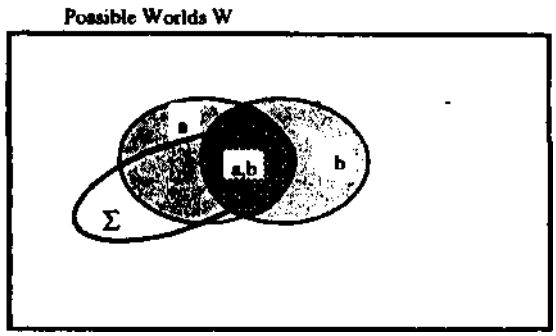

Figure 4: Means-ends intentions and belief.

and nontrivial.

$$\forall \phi \in (\mathcal{I} \text{ and } \rightarrow). \exists w \in \Sigma. w \notin M_\phi .$$

This semantic constraint has the immediate consequence that all By-arguments are conjoined intentions, and share in all their properties.

Proposition 3.4

$$\models By(\alpha; \beta_i) \supset I^*(\alpha) \wedge I^*(\beta_1) \wedge \cdots \wedge I^*(\beta_n)$$

But there is an additional constraint on the elaboration of intentions, having to do with their means-end relation. An agent should believe that if the elaboration is achieved, the original intention will be also. Consider the diagram of Figure 4, in which the agent has the intention to achieve a by achieving b; for concreteness, take the example of calling the telephone operator by dialing 0. There can be possible worlds in which b does not lead to a: for example, in using the internal phone system of a company. The correct rationality condition for an agent is that he believe, in the particular situation at hand, that achieving 6 will achieve a. This is represented by the set Σ of belief worlds, in which b $\supset$ a holds.

We call a model embedded if it satisfies this constraint on belief and intention structure.

Definition 3.11 A cognitive structure is embedded iff whenever $b_1 \cdots b_n \rightarrow a$, $\bigcap_{i=1}^{n} M_{b_i} \subseteq M_a$.

It can be easily verified that this condition leads to the following theorem.

Proposition 3.5 In all embedded cognitive structures $\langle W, \Sigma, \mathcal{I}, \rightarrow \rangle$,

$$\langle W, \Sigma, \mathcal{I}, \rightarrow \rangle \models By(\alpha; \beta_1, \cdots \beta_n) \supset B(\beta_1 \wedge \cdots \wedge \beta_n \supset \alpha) .$$

While the embedding graph semantics is simple, it leads to interesting interactions in the statics of intention and belief. For example, in plan-recognition it can be used to determine if a recognized plan is well-formed. It is also critical to the theory of the dynamics of intention and belief. We have a preliminary theory of this dynamics expressed as a default system.

4 Conclusion

We have concentrated on the static relation between intention and belief, and shown how the relationship between these two can be represented simply by an appropriate semantics. The static formalism is useful in task such as plan recognition, in which one agent must determine the mental state of another using partial information.

More complex applications demand a dynamic theory, which is really a theory of belief and intention revision. The formalism of cognitive structures can be extended readily to time-varying mental states, by adding a state index to the model. However, the theory of revision is likely to be complicated, even more so than current belief revision models [Gardenfors and Makinson, 1990], and will probably involve elements of default reasoning.

References

[Bratman, 1987] Michael E. Bratman. Intention, Plans and Practical Reason. Harvard University Press, Cambridge, MA, 1987.

[Chellas, 1980] B. F. Chellas. Modal Logic: An Introduction. Cambridge University Press, 1980.

[Cohen and Levesque, 1990a] Philip R. Cohen and Hector Levesque. Intention is choice with commitment Artificial Intelligence, 42(3), 1990.

[Cohen and Levesque, 1990b] Philip R. Cohen and Hector Levesque. Intention is choice with commitment. Artificial Intelligence, 42(3), 1990.

[Gardenfors and Makinson, 1990] P. Gardenfors and D. Makinson. Revisions of knowledge systems using epistemic entrenchment. In M. Vardi, editor, Theoretical Aspects of Reasoning about Knowledge. Morgan Kaufmann, 1990.

[Kautz, 1990] Henry A. Kautz. A circumscriptive theory of plan recognition. In Philip R. Cohen, Jerry Morgan, and Martha E Pollack, editors, Intentions in Communication. MIT Press, Cambridge, MA, 1990.

[Konolige and Pollack, 1989] Kurt Konolige and Martha Pollack. Ascribing plans to agents: Preliminary report. IJCAI, Detroit, MI, 1989.

[Levesque, 1984] Hector J Levesque. A logic of implicit and explicit belief. AAAI. University of Texas at Austin, 1984.

[Moore, 1980] Robert C. Moore. Reasoning about Knowledge and Action. PhD thesis, MIT, Cambridge, MA, 1980.

[Pollack, 1991] Martha E. Pollack. Overloading intentions for efficient practical reasoning. Nous, 1991.

[Rao and Georgeff, 1991] Anand S. Rao and Michael P. Georgeff. Modelling rational agents within a bdi-architecture. KR9I, Cambridge, MA, 1991.

[Shoham, 1990] Yoav Shoham. Agent-oriented programming. Technical Report STAN-CS-90-1335, Stanford University, Palo Alto, CA, 1990.