Asymmetry Thesis and Side-Effect Problems in Linear-Time and Branching-Time Intention Logics

April, 1991

Technical Note 13

By:

Anand S. Rao
Australian Artificial Intelligence Institute

Michael P. Georgeff
Australian Artificial Intelligence Institute

This paper is to appear in the *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence, IJCAI-91*, Sydney, published by Morgan Kaufmann Publishers, Mountain View, California, 1991
This research was partly supported by a *Generic Industry Research and Development Grant* from the Department of Industry, Technology and Commerce, Australia, and partly by the Australian Civil Aviation Authority.

Abstract

In this paper, we examine the relationships between beliefs, goals, and intentions. In particular, we consider the formalization of the *Asymmetry Thesis* as proposed by Bratman [1987]. We argue that the semantic characterization of this principle determines if the resulting logic is capable of handling other important problems, such as the side-effect problem of belief-goal-intention interaction. While Cohen and Levesque's [1990] formalization faithfully models some aspects of the asymmetry thesis, it does not solve all the side-effect problems; on the other hand the formalization provided by Rao and Georgeff [1991] solves all the side-effect problems, but only models a weak form of the asymmetry thesis. In this paper, we combine the intuition behind both these approaches and provide a semantic account of the asymmetry thesis, in both linear-time and branching-time logics, for solving many of these problems.

1 Introduction

Formalizations of intentions and their relationships with other propositional attitudes such as beliefs and goals have received increased attention in recent years [Cohen and Levesque, 1990; Konolige and Pollack, 1990; Werner, 1990; Rao and Georgeff, 1991; Konolige, 1991]. Some of these formalizations have been influenced by the philosophical work of Bratman [1987]. He argues, convincingly, that intentions involve a characteristic form of commitment, play a distinct role in practical reasoning, and are not reducible to beliefs and desires (or goals).

According to Bratman, it is irrational for an agent to intend to do an act a and at the same time believe that he will not do a. However, it is rational for him to intend to do a and not believe that he will do a. In other words, it is irrational for an agent to have beliefs that are inconsistent with his intentions, but perfectly rational to have incomplete beliefs about his intentions. Bratman refers to these two principles of *intention-belief consistency* and *intention-belief incompleteness* as the *asymmetry thesis*. One can also extend the asymmetry thesis to the relationship between intentions and goals, and goals and beliefs. Thus, it is reasonable to require a rational agent to have *intention-goal consistency* and *goal-belief consistency*, but allow *intention-goal incompleteness* and *goal-belief incompleteness*.

The way in which the relationships between beliefs, goals, and intentions are captured can have a significant impact on the design of a rational agent. In particular, if not represented properly, it can lead to the *side-effect problem* and the *transference* problem. The side-effect problem has received a great deal of attention in the literature [Allen, 1990; Bratman, 1987; Cohen and Levesque, 1990; Konolige and Pollack, 1990; Rao and Georgeff, 1991; Konolige, 1991]. It can be stated as follows: an agent who intends to do a should not be forced to intend to do b, no matter how strongly he believes that doing a will force him to do b. Bratman [1987] provides the example of a strategic bomber[1] who intends to bomb a munitions factory and also believes that doing so would kill all the children in a nearby school. In this case, one can argue that the strategic bomber does not intend to kill the children in the school but brings it about as a side-effect of bombing the munitions factory. The same principle extends to the relationship between goals and beliefs.

A related problem is the problem of transference. An agent who believes that the formula ϕ will be inevitably true some time in the future should not be forced to have a goal to achieve ϕ nor be forced to intend it. For example, an agent believing that "it is inevitable that the sun will rise in the east tomorrow morning" should not be required to have this condition as a goal nor to intend it.

Cohen and Levesque [1990] were the first to formalize some of these ideas. They present a possible-worlds model for beliefs and goals. Each possible world is a *time-line* representing a sequence of events, temporally extended infinitely into the past and the future. Formulas are evaluated with respect to a given world and an index into the course of events defining the world. Accessibility relations $\mathcal{B}$ and $\mathcal{G}$ are relations between the world at an index to a set of worlds or courses of events. These worlds are called belief-accessible and goal-accessible worlds, respectively. Intuitively, an agent believes a proposition in a world at a particular index if and only if the proposition is satisfied in all the belief-accessible worlds. A similar relationship holds between goals and goal-accessible worlds.

In the Cohen–Levesque formalism, one would intuitively expect the goal-accessible worlds to be some subset of the agent's belief-accessible worlds. This constraint, called *realism* by Cohen and Levesque [1990], ensures that the worlds chosen by an agent are not ruled out

[1] A more sensitive reader can consider the example of a person intending to water rose plants, without intending to water the weeds at the base of the rose plants, even though he strongly believes that watering the rose plants will result in watering the weeds.

by his beliefs. This constraint also realises some aspects of Bratman's asymmetry thesis. However, as we shall see later, it also leads to certain problems concerning the side effects of actions (as observed by Cohen and Levesque [1990] and Allen [1990]). Additionally, it is unsatisfactory in that any beliefs about the future thereby become adopted as goals.

Elsewhere [Rao and Georgeff, 1991], we have provided an alternative possible-worlds formalism where each world is a branching-time structure with a single past and multiple futures. Accessibility relations $\mathcal{B}$, $\mathcal{G}$, and $\mathcal{I}$ are used to represent the beliefs, goals, and intentions of the agent, respectively.

As with Cohen and Levesque's formalism, each possible world represents, according to the agent, the way the world could turn out to be. However, it differs in that the branches within each of these possible worlds represent the *choice* available to the agent in determining what actions to perform. Thus the formalism distinguishes between the choice available to the agent (represented by the branching structure within each possible world) and the chance (or lack of knowledge of the agent) concerning in which world he is possibly situated.

In our approach, the notion of realism is captured by requiring that for every belief-accessible world there exists a goal-accessible world that is a sub-world of that belief-accessible world. However, there can be goal-accessible worlds that do not have corresponding belief-accessible worlds. A similar relationship holds between goal-accessible worlds and intention-accessible worlds. Thus, moving from belief to goal to intention worlds amounts to successively pruning the paths of the time tree; intuitively, to making increasingly selective choices about one's future actions.

Thus stated, this property turns out to be a somewhat stronger notion of realism than that used by Cohen and Levesque. It essentially states that an agent can only have a goal towards some proposition if he believes that, *no matter how the world turns out*, he has the *option* of eventually achieving that goal. While the agent may contemplate possible failure along the way, he believes that he can eventually recover from such failures and ultimately achieve his goals. A similar constraint applies to the relationship between goals and intentions. This restriction on goals and intentions is desirable when one wants to ensure that a system (agent) will only adopt goals or intentions towards ends over which it has control. However, it is too strong for modeling rational agents. We thus call this constraint *strong realism* [Rao and Georgeff, 1991].

The realism constraint for linear-time intention logic proposed by Cohen and Levesque can be weakened to a constraint that requires only that the intersection of belief and goal-accessible worlds be non-empty, i.e. there is *at least* one world common to the belief and goal-accessible worlds. We shall call this the *weak-realism* constraint for linear-time intention logic.

Similarly, we can weaken the strong-realism constraint for branching-time intention logic to a weak-realism constraint. In particular, instead of requiring that for every belief-accessible world there exists a corresponding goal-accessible world, we simply require that there exists *at least one* belief-accessible world with a corresponding goal-accessible world (as before, this goal-accessible world must be a sub-world of the belief-accessible world).

In this paper, we shall show that, although the weak-realism constraint appears extremely weak and inadequate, it is all that is needed to satisfy all aspects of the asymmetry thesis and to avoid the side-effect and transference problems. We also show that, by adding this semantic constraint to the formalism proposed by Cohen and Levesque, some of the stronger side-effect problems in their logic can be avoided. However, even with this semantic constraint, the strong case of side-effects between intentions and goals cannot be avoided. If intentions are defined as basic entities, irreducible to the other basic attitudes of belief and desire, this side-effect can also be avoided.

2 Belief-Goal-Intention Interaction

In this section, we formally define some of the properties discussed above. We characterise these principles for linear-time intention logics. Although we provide a language for expressing these principles and also talk about satisfiability or validity of these principles with respect to a model, we do not provide a specific model in this section. In other words, any model (be it possible-worlds [Cohen and Levesque, 1990; Rao and Georgeff, 1991], situation-semantics [Werner, 1990] or representationalist [Konolige and Pollack, 1990]) has to satisfy at least these principles. In later sections we shall examine specific models.

The language we use to capture these properties for any linear-time model is as follows.[2] $\mathsf{BEL}(\phi)$, $\mathsf{GOAL}(\phi)$, and $\mathsf{INTEND}(\phi)$ denote the belief, goal, and intention in ϕ, where ϕ is a first-order formula.[3] In addition to the above, temporal formulas $\Box\phi$ (always) and $\Diamond\phi$ (sometimes) are also defined.

Asymmetry Thesis

Bratman argues that it is irrational for an agent to intend do an action and also believe that he will not do it. Thus he does not allow *intention-belief inconsistency* (BI-ICN). On the other hand, he does allow a rational agent to intend to do an action but not believe that he will do it. Thus *intention-belief incompleteness* (BI-ICM) is allowed. These two principles put together is called the *Asymmetry thesis*. More formally,

(BI-ICN) $\not\models \mathsf{INTEND}(\phi) \wedge \mathsf{BEL}(\neg\phi)$
(BI-ICM) there exists a model M such that $M \models \mathsf{INTEND}(\phi) \wedge \neg\mathsf{BEL}(\phi)$.

The asymmetry thesis can be extended to hold between intentions and goals, and goals and beliefs as well. That is we must not allow intention-goal (GI-ICN) and goal-belief inconsistency (BG-ICN), whereas we should allow intention-goal (GI-ICM) and goal-belief incompleteness (BG-ICM).

Side-Effect-Free Principle

The belief-intention side-effect-free principle states that, if an agent intends ϕ, he should not be forced to intend a side-effect ψ, no matter how strong the belief about $\phi \supset \psi$. As described by Cohen and Levesque, the strength of the belief could be either one of the following:[4] $\mathsf{BEL}(\phi \supset \psi)$, $\mathsf{BEL}(\Box(\phi \supset \psi))$, or $\Box\mathsf{BEL}(\Box(\phi \supset \psi))$. The strongest side-effect-free principle is stated as:

(BI-SE3) there exists a model M such that
$M \models \mathsf{INTEND}(\phi) \wedge \Box\mathsf{BEL}(\Box(\phi \supset \psi)) \wedge \neg\mathsf{INTEND}(\psi)$.

Substituting the second conjunct of BI-SE3 by the weaker forms of beliefs, yields BI-SE1 and BI-SE2. The side-effect problem exists, not only between intentions and beliefs, but also between intentions and goals, and goals and beliefs. Thus analogous to (BI-SE1) – (BI-SE3), we have (GI-SE1) – (GI-SE3) and (BG-SE1) – (BG-SE3).

[2] The language we use here is essentially that of Cohen and Levesque [1990] with some additions from [Rao and Georgeff, 1991]. However, we do not require beliefs, goals, and intentions to be treated as modal operators in this section.

[3] For the sake of simplicity we have dropped the agent argument from all these propositional attitudes.

[4] In reality, there are nine different cases, namely $\mathsf{BEL}(\gamma)$, $\mathsf{BEL}(\Diamond\gamma)$, $\mathsf{BEL}(\Box\gamma)$ $\Diamond\mathsf{BEL}(\gamma)$, $\Diamond\mathsf{BEL}(\Diamond\gamma)$, $\Diamond\mathsf{BEL}(\Box\gamma)$, $\Box\mathsf{BEL}(\gamma)$, $\Box\mathsf{BEL}(\Diamond\gamma)$, $\Box\mathsf{BEL}(\Box\gamma)$, where γ is $\phi \supset \psi$.

Non-Transference Principle

One aspect of the non-transference principle states that no matter how strongly an agent believes in a proposition, he should not be forced to adopt it as a goal. This non-transference principle can be stated as follows:

(BG-NT) there exists a model M such that $M \models \mathsf{BEL}(\phi) \wedge \neg\mathsf{GOAL}(\phi)$.

This transference problem exists, not only between beliefs and goals, but also between goals and intentions and beliefs and intentions. Thus analogous to (BG-NT), we have (GI-NT) and (BI-NT).

A rational agent is one who satisfies all the above principles. More formally,

Proposition 1 *: The necessary conditions for an agent to be called rational are as follows:*
(a) the asymmetry thesis principles BI-ICN, BG-ICN, and GI-ICN regarding consistency and BI-ICM, BG-ICM, and GI-ICM regarding incompleteness are satisfied;
(b) all the side-effect-free principles BI-SE1 – BI-SE3, BG-SE1 – BG-SE3, and GI-SE1 – GI-SE3 are satisfied; and
(c) all the non-transference principles BI-NT, BG-NT, and GI-NT are satisfied.

3 Linear-Time Intention Logic

In this section we present a linear-time intention logic which has all the desirable properties of Proposition 1. We take Cohen and Levesque's [1990] logic as the starting point for our logic and make two major modifications: (a) in addition to the accessibility relations $\mathcal{B}$ and $\mathcal{G}$ for beliefs and goals, we introduce the relation $\mathcal{I}$ for intentions; and (b) instead of the realism constraint between belief-accessible and goal-accessible worlds, we introduce the weak-realism constraint between belief- and goal-, goal- and intention-, and belief- and intention- accessible worlds.

Similar to Cohen and Levesque's logic, we consider a possible-worlds model where each possible world is a sequence of events, temporally extended infinitely in past and future. Formulas are evaluated with respect to an interpretation M, a variable assignment v, a given world w, and an index t into the course of events defining the world. The interpretation M is a fairly standard possible worlds structure with accessibility relations $\mathcal{B}$, $\mathcal{G}$ and $\mathcal{I}$ that map a world at an index to a set of worlds. We use B_t^w to denote the set of belief-accessible worlds from world w and index t, i.e., $B_t^w = \{\ w' \mid <w, t>\mathcal{B}w'\}$. The sets G_t^w and I_t^w are defined likewise. Except for intentions, our semantics is identical to that of Cohen and Levesque. While Cohen and Levesque define intentions in terms of persistent goals, we define intentions in the same way as beliefs and goals, using the intention-accessibility relation $\mathcal{I}$.

The weak-realism constraint (WC) requires that there be at least one world common to belief- and goal-accessible worlds, and similarly for belief- and intention-accessible worlds and goal- and intention-accessible worlds. More formally,

(WC-BG) $G_t^w \cap B_t^w \neq \emptyset$ (i.e., $\mathcal{G} \cap \mathcal{B} \neq \emptyset$)
(WC-GI) $I_t^w \cap G_t^w \neq \emptyset$ (i.e., $\mathcal{I} \cap \mathcal{G} \neq \emptyset$)
(WC-BI) $I_t^w \cap B_t^w \neq \emptyset$ (i.e., $\mathcal{I} \cap \mathcal{B} \neq \emptyset$).

The above semantic constraints correspond to the following weak-realism axioms (WA):

(WA-BG) $\models \mathsf{GOAL}(\phi) \supset \neg\mathsf{BEL}(\neg\phi)$
(WA-GI) $\models \mathsf{INTEND}(\phi) \supset \neg\mathsf{GOAL}(\neg\phi)$
(WA-BI) $\models \mathsf{INTEND}(\phi) \supset \neg\mathsf{BEL}(\neg\phi)$.

Weak Realism: $\mathcal{G} \cap \mathcal{B} \neq \emptyset$

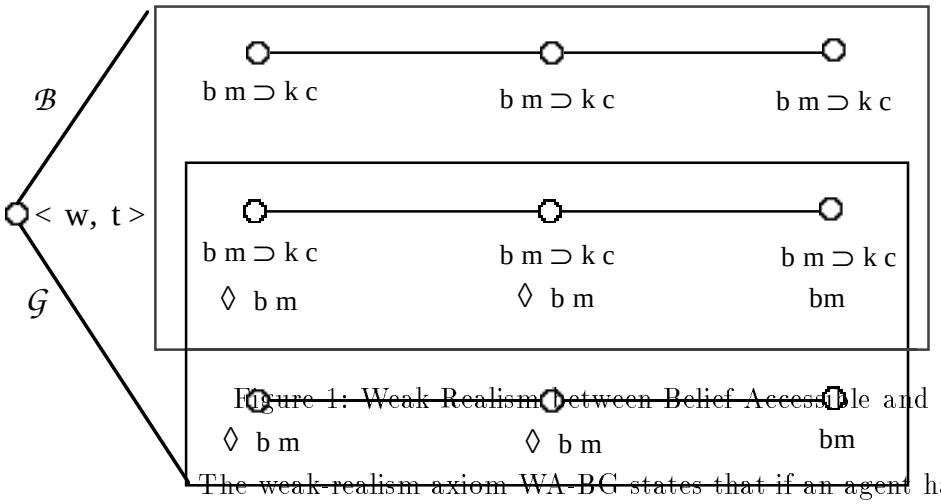

Figure 1: Weak Realism between Belief-Accessible and Goal-Accessible Worlds

The weak-realism axiom WA-BG states that if an agent has the goal ϕ then he will not believe in the negation of ϕ. If the formula ϕ is $\Diamond p$ then the above axiom states that if the agent has chosen the goal to achieve p in the future then he will not believe that it is impossible to achieve p.

One can impose a somewhat stronger constraint; namely, that there be at least one world that is common to belief-, goal-, and intention-accessible worlds. Formally, this translates to $\mathcal{B} \cap \mathcal{G} \cap \mathcal{I} \neq \emptyset$. This stronger constraint implies the above weak-realism constraints but not vice versa.

The semantic constraint of weak-realism is shown in Figure 1 for the strategic bomber example. Note that we have two belief-accessible worlds and two goal-accessible worlds, but only one world is both belief- and goal-accessible. The agent in world w at index t believes always that bombing the munitions factory (bm) will result in killing the children (kc), i.e. $\mathsf{BEL}(\Box(bm \supset kc))$ is satisfiable in w at t. Also, the agent has the goal to eventually bomb the munitions factory, i.e., $\mathsf{GOAL}(\Diamond bm)$ is satisfiable in w at t. However, the formula $\mathsf{GOAL}(\Box(bm \supset kc))$ is not satisfiable because there is a goal-accessible world that is not a belief-accessible world. Intuitively, the goal-accessible worlds which are not belief-accessible worlds are those possible worlds that may turn out to be the real world but are not strong enough to be considered as belief-accessible worlds. For example, the strategic bomber may consider the possibility of the children being moved from the school in which case the implication will not hold [Bratman, 1987]. However, this possibility is not strong enough for him to consider it to be a belief-accessible world. On the other hand, if such a scenario did arise (namely, the children being moved from the school), he would still like to bomb the munitions factory and hence makes it one of his goal-accessible worlds.

To summarise, we shall adopt the following set of axioms and inference rules for our logic: (a) weak-S5 (or KD45) axioms and belief-necessitation inference rule for beliefs; (b) K and D axioms and goal-necessitation inference rule for goals; (c) K and D axioms and intention-necessitation inference rule for intentions; (d) S5 (or KT45) axioms and $\Box$-necessitation inference rule for $\Box$; and (e) the three weak-realism axioms (WA-BG, WA-BI, WA-GI) connecting beliefs, goals, and intentions. The class of models which correspond to the above axiom system can be easily constructed. Apart from the standard constraints on belief, goal, and intention relations (e.g., serial, transitive, and euclidean for $\mathcal{B}$, and serial for $\mathcal{G}$ and $\mathcal{I}$) we also have the three weak-realism constraints mentioned above. We shall refer to the Linear-Time Intention Logic with the above axiom system and class of models as

LITIL-W-BGI.[5]

Now we want to show that LITIL-W-BGI satisfies Proposition 1. The various inconsistency properties of beliefs, goals, and intentions are reformulations of the weak-realism axioms. The incompleteness principles and different versions of the non-transference principles hold because each relation is not a subset of any other relation. For example, the intention-belief incompleteness principle holds because (a) there exists a model where all intention-accessible worlds satisfy ϕ and (b) there is a belief-accessible world which is not an intention-accessible world in which ϕ is not satisfied. The non-transference principle is essentially the reverse of this; by substituting intentions with beliefs and vice versa in the above argument we can show that belief-intention transference does not arise.

The different versions of the side-effect-free principles are satisfied for the same reason as the non-transference principles. Consider the case of the intention-belief side-effect-free principle BI-SE2. No matter how strong the beliefs of an agent are, one can postulate an intention-accessible world where ϕ is true but $\phi \supset \psi$ is false, and hence the agent does not intend ψ. A similar argument holds for the side-effect-free principles involving the other propositional attitudes.

In the case of the strategic bomber, the agent is not required to have the goal that it is always the case that bombing the munitions factory will result in the children being killed, even though he might have such a belief. Therefore, the agent is not forced to have the goal that eventually the children will be killed nor is he forced to intend to kill the children.

Note that the above line of reasoning does not appeal to possible changes of belief and hence can be applied to the stronger version, namely the BI-SE3 principle. This is illustrated in the following example by Allen [1990]. Consider an agent who, for the sake of winning a bet, intends to drink a full bottle of wine within the next five minutes and always believes that it is always true that drinking for five minutes will cause him to get drunk. By a similar line of reasoning to that above, the agent can intend to drink the entire bottle of wine without intending to get drunk. Thus having intended the primary action the agent is not forced to intend one of its side-effects no matter how strong his beliefs about these side-effects. These results are summarised in Figure 3. The symbols Y and N denote the satisfaction and non-satisfaction of the principles, respectively.

4 Other Linear-Time Intention Logics

In this section we consider two other linear-time intention logics – Cohen and Levesque's logic, and a modified version of Cohen and Levesque's logic with the weak-realism constraint between beliefs and goals.

Using beliefs, goals, and actions as basic entities, Cohen and Levesque [1990] define the notion of persistent goals and intentions. An agent has a *persistent goal* or $\mathsf{PGOAL}(\phi)$ if and only if the agent currently believes $\neg\phi$, has the goal to eventually make ϕ true, and maintains this goal until he either comes to believe in ϕ or comes to believe that ϕ is impossible. They capture the notion of intention as a special type of persistent goal. More specifically, an agent intends an action a if and only if he has a persistent goal to have done the action a and, until he has done it, maintains his belief that he is doing it. Although the above notion of intention is based on a fanatical commitment by the agent, Cohen and Levesque also define other notions of intention based on less severe forms of commitment, namely, relativized commitment. For the purposes of this paper, however, we shall not be concerned

[5] W indicates weak-realism and BGI indicates that there are independent relations for Beliefs, Goals, and Intentions.

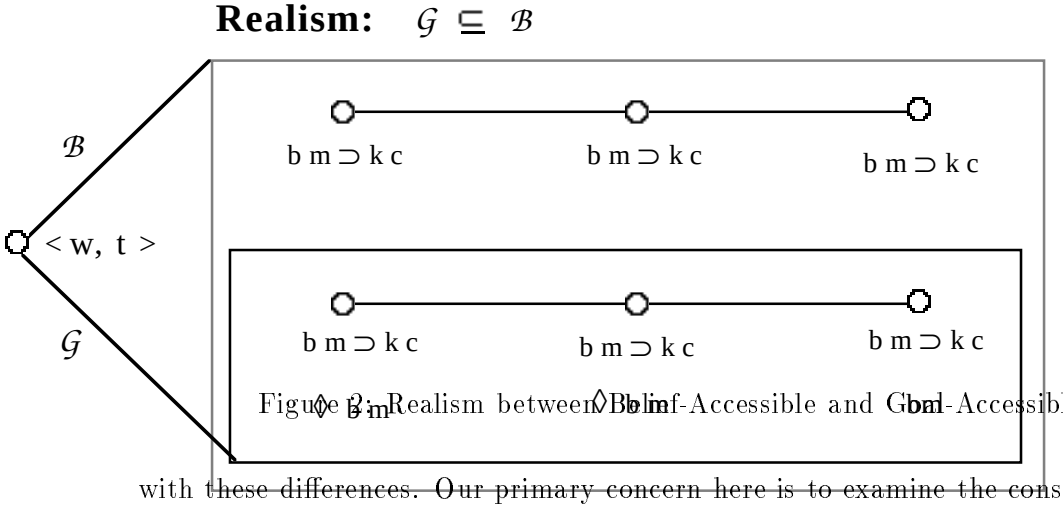

Figure 2: Realism between Belief-Accessible and Goal-Accessible Worlds

with these differences. Our primary concern here is to examine the constraints imposed on beliefs and goals, and see how these affect the properties discussed earlier.

As they define persistent goals and intentions in terms of the beliefs, goals, and actions of the agent, Cohen and Levesque have only two relations: the belief-accessibility relation $\mathcal{B}$ and the goal-accessibility relation $\mathcal{G}$. The realism constraint (RC) and its corresponding axiom (RA) (Proposition 3.26 of [Cohen and Levesque, 1990]) between beliefs and goals are as follows:

(RC-BG) $G_t^w \subseteq B_t^w$ (i.e., $\mathcal{G} \subseteq \mathcal{B}$)
(RA-BG) $\models \mathsf{BEL}(\phi) \supset \mathsf{GOAL}(\phi)$.

The axiom states that if an agent believes in ϕ he also has it as a goal. In other words, if ϕ is taken to be $\Diamond p$, the axiom states that, if the agent believes that eventually p will be true, he will adopt it as a goal. Note that the realism axiom implies the weak-realism axiom but not vice versa.[6]

The object of an intention as defined by Cohen and Levesque can only be an action formula. However, as intentions are just a special type of persistent goal and the object of a persistent goal can be a well-formed formula, we shall treat $\mathsf{PGOAL}(\phi)$ as being synonymous with $\mathsf{INTEND}(\phi)$ as discussed in Section 2. Hence, we shall consider all the properties of Section 2 with INTEND substituted by PGOAL.

We shall refer to Cohen and Levesque's logic as LITIL-R-BG. A summary of some of the important properties satisfied by LITIL-R-BG are shown in Figure 3. All the incompleteness principles are satisfied by LITIL-R-BG. They follow directly from the realism axiom and the definition of persistent goals. As shown in Figure 3, the belief-goal inconsistency principle is satisfied by LITIL-R-BG, but the goal-intention and belief-intention inconsistency principles are not satisfied. This is because, by the definition of persistent goals, if the agent has a persistent goal towards ϕ, he believes in $\neg\phi$. (Consequently, by the realism axiom, he must have the goal that $\neg\phi$.) However, this should not be viewed as a problem that needs to be fixed; it is a consequence of their definition of persistent goals.

The results for the non-transference principles are exactly the opposite and are given in Figure 3. The realism axiom forces belief-goal transference. However, as Cohen and Levesque note, the agent may adopt the goal while reluctantly believing that, if he comes to change his beliefs about the inevitability of ϕ in the future, he might drop the goal. The lack of a

[6] From the realism axiom $\mathsf{BEL}(\neg\phi) \supset \mathsf{GOAL}(\neg\phi)$ and the D-axiom $\mathsf{GOAL}(\neg\phi) \supset \neg\mathsf{GOAL}(\phi)$, we have $\mathsf{BEL}(\neg\phi) \supset \neg\mathsf{GOAL}(\phi)$. Taking the contrapositive of this, we obtain the weak-realism axiom $\mathsf{GOAL}(\phi) \supset \neg\mathsf{BEL}(\neg\phi)$.

similar axiom between goals and persistent goals avoids goal-intention and belief-intention transference.

We illustrate the side-effect-free principle for LITIL-R-BG using the strategic bomber example as shown in Figure 2. As before, the formulas $\mathsf{BEL}(\Box(bm \supset kc))$ and $\mathsf{GOAL}(\Diamond bm)$ are satisfiable in w at t. The realism constraint requires that all goal-accessible worlds be belief-accessible, which forces $\mathsf{GOAL}(\Box(bm \supset kc))$ also to be satisfiable. However, the bomber need not have the goal to kill the children because at the next time point (or some time in the future before he bombs the munitions factory) he can change his beliefs such that the implication does not hold. Consequently, he also need not intend to kill the children. Thus both BG-SE2 and BI-SE2 principles are satisfied. However, as noted by Cohen and Levesque [1990] and Allen [1990], the same reasoning does not apply to the stronger SE3 principles for non-trivial cases.

A similar result holds for the side-effects involving intentions and goals, i.e., GI-SE1 and GI-SE2 are satisfied, but GI-SE3 is not satisfied. As beliefs imply goals, all three side-effect problems involving goals and beliefs exist, i.e., BG-SE1, BG-SE2, and BG-SE3 are not satisfied. A summary of the strongest form of the side-effect-free principles, namely SE3, is shown in Figure 3. The table considers only the non-trivial cases of the SE3 principle; the trivial cases of SE3 are satisfied by all the logics.

Next we consider Cohen and Levesque's logic with the realism constraint replaced by the weak-realism constraint. We shall refer to this logic as LITIL-W-BG. The properties satisfied by LITIL-W-BG are summarised in Figure 3. The important points to note are that LITIL-W-BG does not suffer from belief-goal transference nor does it have belief-goal or belief-intention side-effect problems. The same line of reasoning as in Section 3 can be applied to prove these results. Note that the intention-goal side-effect problem remains in LITIL-W-BG. This is because intentions are defined in terms of beliefs and goals. This manifests itself when we consider the side-effect-free principles regarding intentions and goals. We cannot use the same reasoning as above to solve the strong case of intention-goal side-effect problem GI-SE3 and we still have to appeal to the change in goals from one time point to another to solve the weaker intention-goal side-effect problem GI-SE2.

While the desirability of the GI-SE3 principle may be debatable, we have provided the constraint WC-GI that satisfies this principle. One can either adopt this constraint or not independent of whether one wants to introduce intentions as basic entities or define them in terms of beliefs and goals.

However, we believe that the intention relation plays an important role in means-end reasoning. In essence, the goals of an agent represent his pro-attitudes and his intentions represent the refinement of these goals into realizable means (or conduct-controlling pro-attitudes [Bratman, 1987]). In particular, this distinction allows an agent to maintain his goals, even if his intentions towards the means for achieving these goals fail. If intentions are defined in terms of goals, such means-end reasoning is not transparent and can only be captured by resorting to various book-keeping mechanisms.

5 Branching-Time Intention Logic

The language we use for branching-time intention logic is a CTL^* [Emerson and Srinivasan, 1989] branching-time logic within a possible-worlds framework (see [Rao and Georgeff, 1991] for more details). In addition to the operators of linear-time intention logic, the branching-time logic has two additional operators: $\mathsf{inevitable}(\phi)$, meaning that in all future paths ϕ is true; and $\mathsf{optional}(\phi)$, meaning that in at least one future path ϕ is true. Well-formed formulas that contain no positive occurrences of inevitable (or negative occurrences of optional) outside

Logic	*ICN Principle*			*ICM Principle*			*SE3 Principle*			*NT Principle*		
	BG	GI	BI	BG	GI	BI	BG	GI	BI	BG	GI	BI
LITIL-W-BGI	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
LITIL-R-BG	Y	N	N	Y	Y	Y	N	N	N	N	Y	Y
LITIL-W-BG	Y	N	N	Y	Y	Y	Y	N	Y	Y	Y	Y

Figure 3: Principles satisfied by Linear-Time Logics

the scope of belief, goal, or modal operators will be called O-formulas and will be denoted by α.

An ideal theory of rational agency, in the case of branching-time intention logic, should also satisfy Proposition 1, namely, the asymmetry thesis, side-effect-free and non-transference principles.

Now we briefly describe the strong-realism constraint introduced elsewhere [Rao and Georgeff, 1991]. First, we define the notion of a sub-world. Intuitively, a sub-world is a sub-tree of a given world with the same truth assignment and accessibility relations. More formally, we say that w' is a sub-world of w, denoted by $w' \sqsubseteq w$, iff (a) the index points of w' are a subset of the index points of w; (b) the same events occur between two index points in w' and w; (c) the assignment of truth values for predicate symbols for w' and w are identical; and (d) the accessibility relations for w' and w are also identical. Sometimes, we shall also say that w is a super-world of w'. The strong-realism constraint (SC) (between belief and goal-accessible worlds) requires that for every belief-accessible world there exists a goal-accessible world that is a sub-world of the belief-accessible world. A similar constraint holds for goal- and intention-accessible worlds.

(SC-BG) $\forall w' \in \mathcal{B}_t^w \; \exists w'' \in \mathcal{G}_t^w$ such that $w'' \sqsubseteq w'$ (denoted by $\mathcal{B}_t^w \subseteq_{super} \mathcal{G}_t^w$)
(SC-GI) $\mathcal{G}_t^w \subseteq_{super} \mathcal{I}_t^w$.

The strong-realism constraint is equivalent to the following two axioms (as before, α is taken to be an O-formula):

(SA-BG) $\models \mathsf{INTEND}(\alpha) \supset \mathsf{GOAL}(\alpha)$
(SA-GI) $\models \mathsf{GOAL}(\alpha) \supset \mathsf{BEL}(\alpha)$.

In other words, even if the agent intends optionally to do an action, he should have a goal that optionally he is going to do the action, and also believe that he will optionally do it. Thus the intention-goal, goal-belief, and intention-belief incompleteness principles are not satisfied. However, as we have shown previously [Rao and Georgeff, 1991], none of the side-effect problems arise with this semantic constraint. This is because there are goal-accessible worlds that are not belief-accessible and there are intention-accessible worlds that are not goal-accessible. Also, none of the transference problems arise for the same reasons. We refer to this Branching-Time Intention Logic with the above strong realism axioms and other axioms given elsewhere [Rao and Georgeff, 1991] as the BRITIL-S-BGI system. Some of the properties of BRITIL-S-BGI are summarised in Figure 4.

The weak-realism constraint (WC) (between belief and goal-accessible worlds) states that there is at least one belief-accessible world such that there exists a goal-accessible world that

is a sub-world of this belief-accessible world. A similar constraint holds between goal and intention-accessible worlds, and belief and intention-accessible worlds.

(WC-BG) $\exists w' \in \mathcal{B}_t^w$ such that $\exists w'' \in \mathcal{G}_t^w$ and $w'' \sqsubseteq w'$ (denoted by $\mathcal{B}_t^w \cap_{super} \mathcal{G}_t^w \neq \emptyset$)
(WC-GI) $\mathcal{G}_t^w \cap_{super} \mathcal{I}_t^w \neq \emptyset$
(WC-BI) $\mathcal{B}_t^w \cap_{super} \mathcal{I}_t^w \neq \emptyset$.

The weak-realism axioms for branching-time intention logic are the same as the weak-realism axioms for linear-time intention logic.

We can show that the strong-realism axioms imply the weak-realism axioms but not vice versa. With the semantic constraint of weak-realism we can satisfy the asymmetry thesis and also avoid the side-effect and transference problems. We shall refer to this branching-time intention logic with the above weak-realism constraints and axioms as the BRITIL-W-BGI system. The properties of this system are summarised in Figure 4.

Logic	*ICN Principle*			*ICM Principle*			*SE3 Principle*			*NT Principle*		
	BG	GI	BI	BG	GI	BI	BG	GI	BI	BG	GI	BI
BRITIL-S-BGI	Y	Y	Y	N	N	N	Y	Y	Y	Y	Y	Y
BRITIL-W-BGI	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Figure 4: Principles satisfied by Branching-Time Logics

6 Conclusion

Bratman [1987] and others [Bratman *et al.*, 1988; Cohen and Levesque, 1990] have stated certain properties that are desirable for the design of rational agents. This paper formalizes some of these properties and examines different logics that satisfy some or all of these properties.

The primary contribution of this paper is in defining the semantic constraint of weak-realism that has all the desirable properties for both linear-time and branching-time intention logics. Replacing the realism constraint in Cohen and Levesque's logic with the weak-realism constraint allows us to avoid all cases of intention-belief side-effect problems without appealing to changing beliefs or goals, which was one of the main criticisms of the formalism by Allen [1990]. However, if intentions are modeled as independent entities, the intention-goal side-effect problem is also avoided. This problem remains in a formalism that treats intentions as being definable in terms of beliefs and persistent goals.

We have also shown how the strong-realism constraints used in our branching-time intention logic [Rao and Georgeff, 1991] can be weakened to satisfy additional properties; namely, the incompleteness principles.

References

[Allen, 1990] J. Allen. Two views of intention: Comments on Bratman and on Cohen and Levesque. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge, Ma., 1990.

[Bratman *et al.*, 1988] M. E. Bratman, D. Israel, and M. E. Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355, 1988.

[Bratman, 1987] M. E. Bratman. *Intentions, Plans, and Practical Reason*. Harvard University Press, Massachusetts, 1987.

[Cohen and Levesque, 1990] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(3), 1990.

[Emerson and Srinivasan, 1989] E. A. Emerson and J. Srinivasan. Branching time temporal logic. In J. W. de Bakker, W.-P. de Roever, and G. Rozenberg, editors, *Linear Time, Branching Time and Partial Order in Logics and Models for Concurrency*, pages 123–172. Springer-Verlag, Berlin, 1989.

[Konolige and Pollack, 1990] K. Konolige and M. Pollack. A representationalist theory of intention. Technical Report (to be published), SRI International, Menlo Park, California, 1990.

[Konolige, 1991] K. Konolige. Intention, commitment and preference. Technical Report (to be published), SRI International, Menlo Park, California, 1991.

[Rao and Georgeff, 1991] A. S. Rao and M. P. Georgeff. Modeling rational agents within a BDI-architecture. In J. Allen, R. Fikes, and E. Sandewall, editors, *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*. Morgan Kaufmann Publishers, San Mateo, 1991.

[Werner, 1990] E. Werner. Cooperating agents: A unified theory of communication and social structure. In L. Gasser and M. N. Huhns, editors, *Distributed Artificial Intelligence: Volume II*. Morgan Kaufmann Publishers, 1990.