Chapter 3

Intentions

I argued in Chapter 1 that intentions are an important scientific abstraction for characterizing agents in multiagent systems. This view is justified by the power with which we, humans, can use concepts such as intentions to understand, predict, and explain the behavior of other humans. The relevant point here is that humans are intelligent beings whose internal physical states we do not have precise knowledge of. I submit that a formalized, though necessarily somewhat restrictive, notion of intentions would prove equally useful in the study of artificial multiagent systems.

In this chapter, I first review the logical and intuitive properties that intentions may, or may not, be taken to have. These properties involve several aspects of our pretheoretic understanding of intentions and constrain how we may formalize them. I then present a formalization of intentions in the framework developed in Chapter 2. Next, I formalize several important properties of intentions, some by imposing additional constraints on models. I then briefly discuss the related concept of desires. I conclude with some general remarks on intentions.

3.1 Dimensions of Variation

Like all commonsense concepts, intentions have several senses or connotations. For the case of humans, especially, many of the intuitions associated with intentions are not always clear, or are mutually contradictory. Intentions are also related to other commonsense concepts, such as desires and hopes. It is common in the philosophical literature, however, to distinguish intentions from these other concepts on the following grounds. An agent's intentions are often taken to be necessarily mutually consistent or, at least, believed to be mutually consistent. And, they are almost always taken to be consistent with the given agent's beliefs. Intentions are also closely related to actions and are taken to be causes of the agent's actions. Indeed, there are several dimensions of variation in the study of intentions. I enumerate and discuss the major ones below with the aim of delineating the issues that are of particular relevance to multiagent systems. I have benefited the most from the work of the philosophers Bratman and Brand for much of this discussion [Bratman, 1987; Brand, 1984].

I-DIM-1. *Propositions versus actions:* Intentions can variously be taken to be towards (a) propositions that an agent is deemed to intend to *achieve*, i.e., achieve a state in which they are true, or (b) actions that an agent is deemed to intend to *perform*. Different natural language examples fit these views to different degrees. This dichotomy is largely irrelevant in the approach taken here. Intentions *per se* are taken to apply to propositions, which makes for a natural discussion of their logical properties. However, since I explicitly consider strategies in this framework, it is possible to obtain the effects of applying intentions to actions. An agent having a certain strategy can be said to intend the abstract action which that strategy denotes. Since the language of strategies allows fairly complex procedures to be described, the present approach can accommodate intentions towards action.

The notion of intending to perform a basic action can also be captured. For example, the condition $\mathsf{A}x\langle a\rangle\mathsf{true}$ denotes that action a is about to be completed on all available scenarios: it holds when the agent has chosen and begun action a, but not yet completed it. That is why it *will* be completed on all possible scenarios. However, it is not clear what this notion of intending basic actions might be used for. This is because in the proposed approach, intentions are supposed to be *abstractions* of agents' states and behaviors. Still, it is good to know that such requirements can be captured here, should they ever be needed for some applications.

I-DIM-2. *Future-directed versus present-directed:* It should be clear that intentions cannot be about past times. In the philosophical literature, they are taken to be either towards future states of the world or future actions, or towards present actions. The terms used here are due to Bratman [1987, p. 4]. Searle uses the term *intention-in-action* to denote the latter sense [1983, p. 106]. Brand uses the terms *prospective* and *immediate* to distinguish them. However, there is general agreement in the recent literature that the former, i.e., prospective, sense is the primary sense of intentions. For the purposes of multiagent systems too, intending to achieve a particular future state of the world is the more useful notion. This is because it is the one that relates intentions to the strategies that agents execute and can be used as a predictor of their behavior.

Indeed, the notion of present-directed intentions is used primarily by philosophers to address some of their concerns about whether a given behavior of an agent is indeed an action. For instance, a popular example is a person's moving his arm: this behavior counts as an action only if he intends it *while* doing it and this intention causes the arm to move. It would not be an action if he intended it before it happened, but it happened because a neurosurgeon sent an electrical pulse on an appropriately connected electrode. Such conundrums are of limited value in computer science, at least at present. A possible use of present-directed intentions is to determine whether a certain putative action was intended. But the relevant aspects of this case are readily subsumed by future-directed intentions: an agent can be said to have intended an action if it was done as part of a prior future-directed intention of his [Brand, 1984, p. 28].

In our formal model, agents are assigned different actions on different periods: if it helps we can think of each of those actions as being intended during the periods in which they are performed. But in the sequel, I shall use the term *intention* exclusively to refer to future-directed intentions.

I-DIM-3. *Intending versus doing intentionally:* Another dimension of variation relevant to intentions is perhaps more useful in computer science. This concerns the difference between intending something and doing it intentionally. The former involves the true intentions or preferences of an agent; the latter applies to actions or states that the agent purposefully performs or brings about, but not with any prior intention to do so [Bratman, 1987, p. 119]. For example, an agent who intends to load paper into a photocopier may have to pick some ream of paper to do so. However, while he picks a specific ream deliberately and knowingly, he may not have intended in advance to pick that particular one. In the proposed approach, the conditions that an agent brings about intentionally are the ones that occur as, possibly contingent, consequences of his following his strategy. Only the conditions that are intended may be used to explain an agent's actions. However, the conditions he brings about intentionally, but does not intend, may be used in understanding or estimating his priorities. I discuss a related point in item I-DIM-6 below.

I-DIM-4. *Satisfiability:* It is usually believed that an agent's intentions are satisfiable in some future state of the world. More weakly, it is assumed that an agent believes that his intentions are satisfiable. For example, Bratman requires that an agent's plan be consistent with his beliefs, assuming that the beliefs are themselves not inconsistent [1987, p. 31]. An agent's plans should be executable if his beliefs are correct. Roughly, the motivation for this is that intentions are taken to apply to agents who are rational in some sense. Rationality is not formally characterized here; however, a constraint is stated in section 3.3 below that captures the requirement of the satisfiability of intentions.

I-DIM-5. *Mutual consistency:* It is commonly suggested that a rational agent's intentions should not preclude each other [Bratman, 1987, p. 31] [Brand, 1984, pp. 125–126]. Brand argues that inconsistency among an agent's intentions would make his mental state too incoherent for him to act.

Mutual consistency follows as a natural consequence of the present approach. If the intentions of an agent are satisfiable, then the strategies assigned by **Y** must be doable, at least on some scenarios. In that case, the intentions of the agent are also mutually consistent. This is because the model considers only those moments and actions that are consistent in the given domain. The above argument applies only if the intentions of one agent are considered. The intentions of different agents may not be mutually consistent. For example, when two agents are playing a zero-sum game, each intends that he win (and the other be forever from prevented from winning that instance of the game). This is allowed by the present approach.

I-DIM-6. *Closure under logical consequence:* In general, if one is talking about human beings, intentions are not closed under logical consequence. This is because the given agent may not have realized the appropriate connection or may have realized it, but does not prefer it, nevertheless. For example, you may intend to be operated on, but even though (let us stipulate) that entails spending a day in a hospital, you may not intend spending a day in a hospital. It is possible to develop formal theories of intentions that preserve this feature, and I have done so in joint work with Nicholas Asher [Singh & Asher, 1993].

However, for the purposes of designing and analyzing multiagent systems, it may be acceptable to let one's theory validate this inference. The reason is similar to those that motivate modal approaches to knowledge, as applied in distributed systems [Chandy & Misra, 1986; Halpern & Moses, 1987]. If one is considering a system from without, one has to consider its possible actions. By definition, a logical consequence of a proposition holds at all moments where that proposition holds. Therefore, it cannot be distinguished from the given proposition on the basis of the results of possible actions for achieving it. Thus there is no principled ground for preventing the given inference.

Newell cites this inference as an example of the profound limitation of the knowledge level. As a result of this inference, the knowledge level cannot aspire to be more than an approximation, and a "radically incomplete" one at that [Newell, 1982, pp. 104–105 and 111]. That is, it may fail to describe "entire ranges of behavior." Robert Boyer has independently argued that the only safe way to use this approach is to check, for each putative claim, whether it improperly relies on the closure inference [Boyer, 1992]. This contrasts with mathematical reasoning in general, where all conclusions of a theory are valid. Here we need an additional filter that looks at the proofs of theorems, rather than the theorems themselves. This restricts the applicability of the proposed approach.

One reason for allowing closure under logical consequence is that sometimes we can take a normative stance towards an (artificial) agent's actions. With such a stance, we can require that an agent have figured out his priorities and decided rationally on a course of action: such an agent is then responsible for whatever actions he performs. Also, a framework that prevents closure under logical consequence can easily become technically intractable. But when one wishes to clarify one's intuitions about complicated concepts such as intentions and to develop tools based on them for designing actual systems, closure under logical consequence remains a shortcoming.

I should note parenthetically that it is possible, in principle, to formally distinguish a proposition from its logical consequences. However, it is not clear if that can be done intuitively acceptably on the basis of possible actions as described here. In any case, logically equivalent propositions cannot be distinguished from each other in any possible worlds framework. Since the intuitive arguments against closure under logical consequence also apply against closure under logical equivalence, I explicitly discuss only the former case.

I-DIM-7. *Closure under believed effects:* It is sometimes argued that, for human beings, intentions are not closed even under beliefs. An agent may intend *p*, believe that *p* necessarily entails *q* and not intend *q*. Bratman gives the example of a strategic bomber who intends to bomb a munitions plant, believes that this will cause the adjacent school to blow up, but nevertheless does not intend to blow up the school [1987, p. 139]. Rao & Georgeff, who agree that this inference is not desirable, have termed it the *side-effect* problem [Rao & Georgeff, 1991a].

I-DIM-8. *Closure under means:* Item I-DIM-7 contrasts with the following claim. An agent who intends *p*, and believes that *q* is a necessary *means* to *p*, should intend *q*. Bratman calls this phenomenon *means-ends coherence* (p. 35). Brand too considers it an essential property of intentions [1984, p. 126]. In this case, it is rational for the agent to intend *q*; indeed he might intend *q* even if it was only of several possible means to *p*. In the case of item I-DIM-7, however, not only does intending the expected side-effects of an intention seem an incorrect description of people's intentions, it is also irrational, since it would only distract an agent from his real intention [1987, p. 142].

I-DIM-9. *Commitment:* A property of intentions that has recently gained acceptance is that usually they involve some measure of commitment on part of the agent [Harman, 1986, p. 94] [Bratman, 1987, ch. 2]. That is, an agent who has an intention is committed to achieving it and will persist with it through changing circumstances. I agree with the usefulness of persistence for the purposes of allowing agents to infer one another's strategies with greater ease and to feasibly coordinate their actions. However, the extent of an agent's persistence with a specific intention is intimately connected with some notion of rationality and involves the costs and benefits of different actions as well as the cost of computation. For example, it is clear that agents should not persist with their intentions forever. If they did, they would end up acting irrationally on intentions that were no longer useful or compatible with their true goals. I do not believe that there is any purely qualitative solution to the problem of when, and for how long, an agent must persist with an intention. However, certain qualitative constraints on intentions and beliefs can be meaningfully stated: an example constraint is described in item I-DIM-11. I have addressed this problem in some related research [1991b; 1991e]; however, I shall not focus on it in this work.

I-DIM-10. *Causation of action:* Another important intuition concerning intentions is that they are causes of actions by agents. This feature is supposed to conceptually differentiate intentions from desires and beliefs. It is not clear how we might use this explicitly in a theory of multiagent systems (recall the discussion in item I-DIM-3 above). However, it can be used to motivate an important constraint on models that intuitively captures a useful property of the architecture of intelligent agents. This is discussed in section 3.3 below.

I-DIM-11. *Intentions must be consistent with beliefs:* Since intentions are somehow related to an agent's rationality, it makes sense to assume that an agent's intentions are consistent with his beliefs about the future. If an agent believes that something is impossible, there is no purpose in his intending it. Similarly, if an agent who has a certain intention later comes to believe that it is, or has become, impossible to achieve, he would do well to drop that intention and concentrate his resources elsewhere. In other words, it should be inconsistent for an agent to intend p and simultaneously believe that p will not occur.

I-DIM-12. *Intentions do not entail beliefs:* While an agent cannot have beliefs that contradict his intentions, it is usually too much to require that an agent believe that he will in fact succeed with whatever intentions he has, or that he will be able to act appropriately for them. In other words, it should be consistent for an agent to intend p and yet not believe that p will occur. This view has been supported by several philosophers, e.g., Bratman [1987, pp. 38]. I return to this point in section 3.3.

I-DIM-13. *Intentions versus beliefs:* Intentions are usually taken to be distinct from beliefs, although they are always taken to be related to them. Allen, however, defines a prior or future-directed intention towards an action as being identical to a belief on part of the agent that he will execute a plan that includes the given action [1984, pp. 145–146]. This definition has some shortcomings. An agent may believe that he will perform a certain action but not be committed to actually performing that action, in the sense of retrying it under appropriate circumstances. Also, it is not clear how an agent's beliefs about the future may actually cause him to act one way or another. One, the agent may have several such beliefs about future happenings and may even intend to prevent some of them. Two, he may intend to do an action, but may not believe that he will necessarily be able to perform it.

> It seems that the key ingredient of adopting a plan or strategy is missing from Allen's definition. It may be possible to define an agent's beliefs about his future actions in such a way as to differentiate them from his other beliefs and to capture many of the important properties of intentions. However, it is not clear what one might gain by that exercise. If one takes the trouble to differentiate the relevant kinds of beliefs from others, one might as well treat them as a distinct concept, and use the term "intentions" to describe them.

Intentions can be assigned several other philosophically motivated properties as well. However, my aim here is to take a minimalist stance, i.e., to study the simplest concept that will suffice for our needs. I submit that for the purposes of multiagent systems, the semantics of intentions should relate them to the strategies of agents and to the actions those agents may possibly perform. Intentions here are assigned on the basis of strategies and are computed in models that consider the possible actions of agents and the possible states of the world. As a result, several of their interesting properties can be derived from model-theoretic constraints on strategies. For example, we can state a constraint on models that ensures that the agents' intentions are satisfiable. Roughly, this constraint says that the strategies assigned to agents at different moments are doable (by them) on at least some scenarios at those moments.

Many intuitive properties of intentions, including their role as causes of action, are properly seen as matters of agent architecture. In the case of causing action, the relevant features of the architecture are the procedures of action selection that an agent employs. Taking strategies as primitive entities not only throws some light on how these properties may be realized in an agent's architecture, but also on how they may be captured in our formal model. Since each intention must be founded on some strategy, we can model the desired property by having agents' strategies restrict their actions appropriately.

For example, we can constrain the selection of actions so that an agent may begin an action only if it, at least potentially, leads to the satisfaction of the current part of the assigned strategy. Thus an agent with an as-yet-unsatisfied strategy must act in a manner that may lead to its satisfaction. There is no guarantee that he would succeed. However, this assumes that his strategy is not impossible to satisfy. The suggested constraint would allow the agent to try different actions on different scenarios, but that is only reasonable: in general, there may be more than one way to satisfy an intention or to carry out a strategy. In Chapter 5, where a constraint on the selection of actions is formally stated, it is somewhat stronger than the one here. There, an agent is required to select actions with which he can force the success of his strategy, assuming, of course, that such actions are available. The technical basis for this is developed in Chapter 4.

Some of the other properties of intentions, e.g., the tendency of agents to persist with them, are not a part of their *semantics*. Rather, these are consequences of constraints on how intentions are updated or agents' strategies revised. These constraints can be motivated on grounds of rationality. They can stated as additional requirements on agents. We can use the semantics developed here to assign meaning to these constraints, to formally infer their properties, and to define a notion of consistency among them.

Further desiderata for a theory of intentions are the following. For a theory of intentions to be of general applicability in computational systems, it should not be committed to a plan-based architecture of intelligent agents. It has recently been argued by several researchers that intelligence is not solely a matter of explicitly representing and interpreting symbolic structures, or at least not necessarily so [Agre & Chapman, 1987]. It would be useful to accommodate the kinds of systems these researchers consider, which are systems that involve a significant reactive component. A good theory of intentions should, however, be compatible with a plan-based architecture. This is because the main intuition behind adopting the intentional stance is that we must proceed without explicit knowledge of the details of the given system's design.

3.2 Intentions Formalized

Perhaps the most basic conception of intentions is to associate them with the preferences of an agent. It is helpful to think of an agent as somehow having "selected" some scenarios as those that he prefers (prefers to realize, as it were). In other words, of all the possible future courses of events, some courses of events are preferred by the agent. These are the ones that correspond to the agent's intentions.

For example, consider Figure 3.1. Assume that $\neg p$ and $\neg q$ hold everywhere other than as shown. Let the agent x (whose actions are written first in the figure) at moment t_0 prefer the scenarios S_1 and S_2. Then, by the informal definition given above, we have that x intends q (because it occurs eventually on both the preferred scenarios) and does not intend p (because it never occurs on S_1). At t_0, x can do either action a or action b, since both can potentially lead to one of the preferred scenarios being realized. Note, however, that if the other agent does action d, then no matter which action x chooses, he will not succeed with his intentions, because none of his preferred scenarios will be

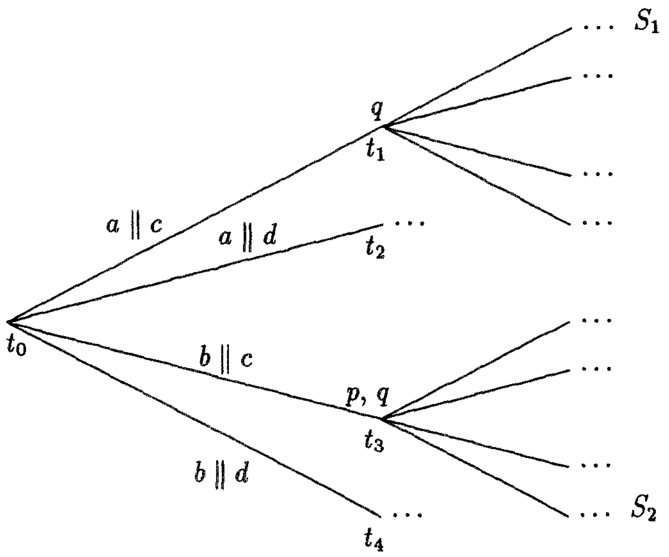

Figure 3.1: Intentions

realized. It is this observation that largely motivates the concept of know-how, which I discuss in Chapter 4.

A standard modal approach would do no more than assume the assignment of preferred scenarios to agents. However, in the approach taken here, I show how the selection of scenarios can itself be founded on the strategies that the agents have. Technically, this has the advantage of grounding claims about agents' intentions in the strategies they follow. It also has the intuitively appealing property that we can apply it to agents who may not be said to have a goal or preference. This is because, as discussed in section 2.5, even an agent who is given as not trying to achieve any goals is following some strategy and, therefore, can be said to have certain appropriate intentions.

Thus we arrive at the following general definition of intentions: *an agent intends all the necessary consequences of his performing his strategy.* Note that only the consequences of the *successful* performance of the strategy are included. There is no guarantee that a given strategy will in fact be successfully performed.

Despite its simplicity, this definition is quite powerful. It considers as intentions only the *necessary* consequences of the performing of the agent's strategy. This is important, because we do not wish to claim that an agent intends even the merely contingent consequences of his performing his strategy. For example, an agent with a strategy for loading paper in a photocopier will have to pick some ream of paper or the other. But he cannot be said to have had the intention of picking the specific one he in fact picks on a particular occasion. This is because he could just as well have picked another one and still satisfied his strategy. This point is related to the discussion in item I-DIM-3 in section 3.1.

3.2.1 Formal Language and Semantics

The formal language of this chapter, $\mathcal{L}^i$, is $\mathcal{L}$ augmented with four operators, $\langle\,\rangle_i$, $\langle\!\langle\,\rangle\!\rangle$, $*$, and I (which stands for $\mathsf{Intends}$).

SYN-13. All the rules for $\mathcal{L}$, with $\mathcal{L}^i$ substituted for $\mathcal{L}$

SYN-14. All the rules for $\mathcal{L}_s$, with $\mathcal{L}^i_s$ substituted for $\mathcal{L}_s$

SYN-15. All the rules for $\mathcal{L}_y$, with $\mathcal{L}^i_y$ substituted for $\mathcal{L}_y$

SYN-16. $p \in \mathcal{L}^i_s$, $x \in \mathcal{A}$, and $Y \in \mathcal{L}^i_y$ implies that $x\langle Y\rangle_i p$ and $\langle\!\langle Y\rangle\!\rangle p \in \mathcal{L}^i_s$

SYN-17. $p \in \mathcal{L}^i_s$, $x \in \mathcal{A}$, and $Y \in \mathcal{L}^i_y$ implies that $x * Y \in \mathcal{L}^i$

SYN-18. $p \in \mathcal{L}^i_s$ and $x \in \mathcal{A}$ implies that $(x \mathsf{I} p) \in \mathcal{L}^i$

It is convenient to define $x[Y]_i p$ as an abbreviation for $x\neg\langle Y\rangle_i \neg p$.

Before giving the formal semantics of the newly introduced operators, it is useful to extend the definition of $[\![\,]\!]$ to apply to strategies. Recall that for a basic action, a, $[\![a]\!]^x$ denotes the set of periods over which a is performed by agent x. Essentially the same intuition is meant to be captured here for the case of strategies. That is, $[\![Y]\!]^x$ will denote the set of periods over which Y is successfully performed by agent x. Thus we have the following definitions. The agent is the same throughout and so is not mentioned.

AUX-5. $[\![\mathbf{skip}]\!] = \{[S;t,t] | t \in \mathbf{T}\}$

The empty strategy is performed on all the trivial periods, i.e., those which consist of just one moment each.

AUX-6. $[S;t,t'] \in [\![\mathbf{do}(q)]\!]$ iff $M \models_{t'} q$ and $(\forall t'' : t \leq t'' < t' \Rightarrow M \not\models_{t''} q)$

The strategy $\mathbf{do}(q)$ is performed over all those periods that begin at any moment where q is achievable and that end at the first occurrence of q after their beginning. This is essentially the property of *action uniqueness*, which was defined as constraint COH-1 in section 2.3, extended to the case of strategies. This extension is only natural, since strategies are abstract actions. A related intuition is that a strategy calls upon an agent to perform some basic actions. There is no reason for the agent to perform a basic action after the relevant strategy has been performed. This is why it makes sense to allow $t = t'$ here, whereas for basic actions, we always have $t < t'$.

AUX-7. $[S;t,t'] \in [\![Y_1;Y_2]\!]$ iff $(\exists t'' : t \leq t'' \leq t'$ and $[S;t,t''] \in [\![Y_1]\!]$ and $[S;t'',t'] \in [\![Y_2]\!])$

In other words, $Y_1;Y_2$ is performed over those periods over which first Y_1 is performed and then, starting at the moment where it ends, Y_2 is performed. Observe that $[\![\mathbf{do}(q);\mathbf{do}(q)]\!] = [\![\mathbf{do}(q)]\!]$. Intuitively, the second $\mathbf{do}(q)$, which is performed in a state in which q holds, does not take any time at all. The first $\mathbf{do}(q)$ may or may not take any time depending on the state where it is performed.

AUX-8. $[S;t,t'] \in [\![\mathbf{if}\ q\ \mathbf{then}\ Y_1\ \mathbf{else}\ Y_2]\!]$ iff $(M \models_t q$ and $[S;t,t'] \in [\![Y_1]\!])$ or $(M \not\models_t q$ and $[S;t,t'] \in [\![Y_2]\!])$

That is, a conditional strategy is performed by performing the substrategy corresponding to the appropriate branch.

AUX-9. $[S;t,t'] \in [\![\mathbf{while}\ q\ \mathbf{do}\ Y_1]\!]$ iff $(t = t'$ and $M \not\models_t q)$ or $(\exists t_0, \ldots, t_n :$ $t = t_0$ and $t' = t_n$ and $(\forall l : 0 \leq l \leq n-1 \Rightarrow ([S;t_l,t_{l+1}] \in [\![Y_1]\!]$ and $M \models_{t_l} q))$ and $M \not\models_{t_n} q)$

That is, an iterative strategy is performed by performing a finite number of iterations of its substrategy. The substrategy is repeated zero or more times until the *first* moment at which the relevant condition does not hold. A consequence of this definition is that $[\![\mathbf{while\ true\ do}\ Y_1]\!]$ is an empty set. And, if $M \models_t q$, then $[\![\mathbf{while}\ q\ \mathbf{do\ skip}]\!]$ is also empty.

Note that the above definitions differ from the corresponding definition for basic actions in at least the following respect: whereas actions always take time, strategies may be trivially performed at a given moment. With these definitions in hand, I can now give the semantics of the operators introduced in this chapter.

SEM-17. $M \models_{S,t} x\langle Y\rangle_i p$ iff $(\exists t' : [S;t,t'] \in \llbracket Y\rrbracket^x$ and $M \models_{S,t'} p)$

SEM-18. $M \models_t x * Y$ iff $\mathbf{Y}(x,t) = Y$

SEM-19. $M \models_t x\langle\!\langle Y\rangle\!\rangle p$ iff $M \models_t \mathsf{A}(x\langle Y\rangle_i \mathsf{true} \rightarrow \mathsf{F}p)$

SEM-20. $M \models_t x\mathsf{I}p$ iff $(\exists Y : M \models_t x * Y$ and $M \models_t x\langle\!\langle Y\rangle\!\rangle p)$

Thus the core semantic definition is that of $x\langle Y\rangle_i p$. Under the above set of definitions, an agent's intentions are determined not only on the basis of the strategy he is currently following, but also on the basis of the actions he can perform on different scenarios. I now state and prove some useful results about auxiliary definition $\llbracket\,\rrbracket$ as applied to strategies, following which I shall consider an axiomatization for intentions.

Lemma 3.1 states that a strategy of the form $\mathbf{do}(q)$ can be begun from any moment before its termination at which it is executing. This corresponds to the similar property of basic actions, which was assumed as coherence constraint COH-3 in Chapter 2.

Lemma 3.1 $[S;t,t'] \in \llbracket \mathbf{do}(q)\rrbracket$ iff $(\forall t_0 : t \leq t_0 \leq t' \Rightarrow [S;t_0,t'] \in \llbracket \mathbf{do}(q)\rrbracket)$

Proof. The right to left direction of the lemma follows by instantiating t_0 as t. For the left to right direction, consider the following. We have that $[S;t,t'] \in \llbracket \mathbf{do}(q)\rrbracket$ iff $M \models_{t'} q$ and $(\forall t'' : t \leq t'' < t' \Rightarrow M \not\models_{t''} q)$. But, since, $t \leq t_0 \Rightarrow (t_0 \leq t'' \Rightarrow t \leq t'')$, the preceding expression implies that $M \models_{t'} q$ and $(\forall t'' : t_0 \leq t'' < t' \Rightarrow M \not\models_{t''} q)$. This holds iff $[S;t_0,t'] \in \llbracket \mathbf{do}(q)\rrbracket$. □

Lemma 3.2 states that, before its termination, an iterative strategy can be begun from any of the intermediate moments at which its substrategy has been executed an integral number of times.

Lemma 3.2 $[S;t,t'] \in \llbracket \mathbf{while}\ q\ \mathbf{do}\ Y_1\rrbracket$ iff $(t = t'$ and $M \not\models_t q)$ or $(\exists t_0, \ldots, t_n : t = t_0$ and $t' = t_n$ and $(\forall l : 0 \leq l \leq n-1 \Rightarrow ([S;t_l,t_{l+1}] \in \llbracket Y_1\rrbracket$ and $M \models_{t_l} q$ and $[S;t_l,t'] \in \llbracket \mathbf{while}\ q\ \mathbf{do}\ Y_1\rrbracket))$ and $M \not\models_{t_n} q)$

Proof. For brevity, let $Y = \mathbf{while}\ q\ \mathbf{do}\ Y_1$. The right hand side of the above expression is stronger than definition AUX-9 (of when $[S;t,t'] \in \llbracket Y\rrbracket$). Hence, the right to left direction follows trivially. Now assume $[S;t,t'] \in \llbracket Y\rrbracket$. Let t_k be any of the intermediate moments, t_0 through t_n. We have $[S;t_n,t_n] \in \llbracket Y\rrbracket$, since $M \not\models_{t_n} q$ holds by the definition. Also, $[S;t_0,t_n] \in \llbracket Y\rrbracket$, by assumption. Consider $0 < k < n$. Then, by definition AUX-9 applied for $[S;t_0,t_n]$, $(\exists t_k, \ldots, t_n : t_k = t_k$ and $t_n = t_n$ and $(\forall l : k \leq l \leq n-1 \Rightarrow ([S;t_l,t_{l+1}] \in \llbracket Y_1\rrbracket$ and $M \models_{t_l} q))$ and $M \not\models_{t_n} q)$. Thus, $[S;t_k,t_n] \in \llbracket Y\rrbracket$. This proves the left to right direction of the lemma. □

Lemma 3.3 $M \models_t \langle Y_1;Y_2\rangle_i p \equiv (\langle Y_1\rangle_i \langle Y_2\rangle_i p)$

Proof. By semantic definition SEM-17, $M \models_t \langle Y_1;Y_2\rangle_i p$ iff $(\exists t' : [S;t,t'] \in \llbracket Y_1;Y_2\rrbracket$ and $M \models_{S,t'} p)$, which by definition AUX-7 holds iff $(\exists t' : (\exists t'' : [S;t,t''] \in \llbracket Y_1\rrbracket$ and $[S;t'',t'] \in \llbracket Y_2\rrbracket)$ and $M \models_{S,t'} p)$. But this reduces to $(\exists t'' : [S;t,t''] \in \llbracket Y_1\rrbracket$ and $M \models_{S,t''} \langle Y_2\rangle_i p)$. Hence, the desired result. □

Lemma 3.4 $[S;t,t'] \in \llbracket Y\rrbracket^x$ iff $(\exists t_1 : [S;t,t_1] \in \llbracket \downarrow_t Y\rrbracket^x$ and $[S;t_1,t'] \in \llbracket \uparrow_t Y\rrbracket^x)$

Proof. The proof proceeds by induction on the structure of strategies. Recall the definitions of $\downarrow$ and $\uparrow$ given in Tables 2.1 and 2.2. The claim holds trivially for strategies of the forms **skip** and **do**(q), since in each case $\uparrow_t Y =$ **skip**. These are the two base cases.

Let $Y = Y_1;Y_2$. By definition AUX-7, $[S;t,t'] \in \llbracket Y_1;Y_2\rrbracket$ iff $(\exists t'' : t \leq t'' \leq t'$ and $[S;t,t''] \in \llbracket Y_1\rrbracket$ and $[S;t'',t'] \in \llbracket Y_2\rrbracket)$. Assume, by the inductive hypothesis, that the given claim holds for Y_1 and Y_2. First, consider the case of $\downarrow_t Y_1 \neq$ **skip**. Then, $[S;t,t''] \in \llbracket Y_1\rrbracket$ iff $(\exists t_1 : [S;t,t_1] \in \llbracket \downarrow_t Y_1\rrbracket^x$ and $[S;t_1,t''] \in \llbracket \uparrow_t Y_1\rrbracket^x)$. Also by definition AUX-7, $[S;t_1,t''] \in \llbracket \uparrow_t Y_1\rrbracket^x$ and $[S;t'',t'] \in \llbracket Y_2\rrbracket$ iff $[S;t_1,t'] \in \llbracket (\uparrow_t Y_1);Y_2\rrbracket$. Thus, $\downarrow_t Y_1 \neq$ **skip** implies that $[S;t,t'] \in \llbracket Y\rrbracket$, which is equivalent to $(\exists t_1 : [S;t,t_1] \in \llbracket \downarrow_t Y\rrbracket$ and $[S;t_1,t'] \in \llbracket \uparrow_t Y\rrbracket)$. If $\downarrow_t Y_1 =$ **skip**, then the claim trivially follows from the inductive hypothesis for Y_2. Thus the given claim holds for strategies of the form $Y_1;Y_2$.

Let $Y =$ **if** q **then** Y_1 **else** Y_2. Let $M \models_t q$. Then, $\downarrow_t Y = \downarrow_t Y_1$ and $\uparrow_t Y = \uparrow_t Y_1$. Since Y_1 is structurally smaller than Y, the desired result holds by the inductive hypothesis. A similar argument applies if $M \not\models_t q$.

Let $Y =$ **while** q **do** Y_1. The case where $M \not\models_t q$ is trivial, since in that case, $t = t'$ and $\downarrow_t Y = \uparrow_t Y =$ **skip**. The case where $M \models_t q$ and $\downarrow_t Y_1 =$ **skip** is also trivial, since then $\llbracket Y\rrbracket = \emptyset$. Now let $M \models_t q$ and $\downarrow_t Y_1 \neq$ **skip**. Then, $\downarrow_t Y = \downarrow_t Y_1$ and $\uparrow_t Y = \uparrow_t Y_1$. Let t_1 be the same as t_1 in the definition of $\llbracket\,\rrbracket$ for **while**. Then, $[S;t,t_1] \in \llbracket Y_1\rrbracket$. Since Y_1 is structurally smaller than Y, by the inductive hypothesis, we have that $(\exists t'' : [S;t,t''] \in \llbracket \downarrow_t Y_1\rrbracket$ and $[S;t'',t_1] \in \llbracket \uparrow_t Y_1\rrbracket)$. By Lemma 3.2 and the choice of t_1, we also have that $[S;t_1,t'] \in \llbracket Y\rrbracket$. Using Lemma 3.3, we obtain that $[S;t'',t'] \in \llbracket (\uparrow_t Y_1);Y\rrbracket$. Thus, $(\exists t'' : [S;t,t''] \in \llbracket \downarrow_t Y\rrbracket^x$ and $[S;t'',t'] \in \llbracket \uparrow_t Y\rrbracket^x)$, which by appropriate relabeling is the present lemma. □

Lemma 3.5 At all moments, t, $M \models_t x\langle Y\rangle_i p \equiv (x\langle \downarrow_t Y\rangle_i x\langle \uparrow_t Y\rangle_i p)$

Proof. The semantic definition of $x\langle Y\rangle_i p$ involves the definition of $\llbracket\,\rrbracket^x$. The proof is a trivial consequence of Lemma 3.4. □

3.2.2 Axioms for Intentions

I now present a sound and complete axiomatization of $\langle Y\rangle_i p$. The agent is not relevant in the following discussion and, therefore, is not mentioned.

Axiom Ax-Int-6 below is a way of relativizing this axiomatization to that of the underlying logic, which includes the boolean operators, i.e., $\wedge$ and $\neg$, the belief or knowledge operator B (or $\mathsf{K_t}$), the temporal operators, i.e., $[\,]$, $\langle\,\rangle$, U, and A, and existential quantification over actions. There are two reasons for relativizing this axiomatization to that of the underlying logic. The first is that we would like to focus attention on the novel contributions here. The other, more subtle, reason is that sometimes no axiomatization may be known for the underlying logic. For example, the logic CTL*, which is decidable, has no known axiomatization [Emerson, 1992]. Axiom Ax-Int-6 can thus be thought of as implicitly invoking an oracle for the underlying logic. This idea is used Chapter 4 also.

Ax-Int-1. $\langle \textbf{skip}\rangle_i p \equiv p$

Ax-Int-2. $\langle Y_1 ; Y_2\rangle_i p \equiv \langle Y_1\rangle_i \langle Y_2\rangle_i p$

Ax-Int-3. $\langle \textbf{if } q \textbf{ then } Y_1 \textbf{ else } Y_2\rangle_i p \equiv (q \rightarrow \langle Y_1\rangle_i p) \wedge (\neg q \rightarrow \langle Y_2\rangle_i p)$

Ax-Int-4. $\langle \textbf{while } q \textbf{ do } Y_1\rangle_i p \equiv (q \rightarrow \langle Y_1\rangle_i \langle \textbf{while } q \textbf{ do } Y_1\rangle_i p) \wedge (\neg q \rightarrow p)$

Ax-Int-5. $\langle \textbf{do}(q)\rangle_i p \equiv (q \wedge p) \vee (\neg q \wedge (\bigvee a : \langle a\rangle \langle \textbf{do}(q)\rangle_i p))$

Ax-Int-6. All substitution instances of the validities of the underlying logic

Theorem 3.6 Axioms Ax-Int-1 through Ax-Int-6 constitute a sound and complete axiomatization of $\langle Y\rangle_i p$ for any model M as described in section 2.1.2.

Proof.

Soundness and Completeness: The proofs of soundness and completeness are developed hand-in-hand. Only formulae of the form $\langle Y\rangle_i p$ are considered here. Construct a model whose indices are maximally consistent sets of sentences of $\mathcal{L}^i$ and $\mathcal{L}^i_s$. The other components of the model, especially, $<$, $\mathbf{B}$, $\mathbf{R}$, and $[\![\,]\!]$ are constrained by the formulae that are true at the different moments and at different scenario and moment pairs. Completeness means that $M \models_{S,t} \langle Y\rangle_i p$ entails $\langle Y\rangle_i p \in (S,t)$ and soundness means that $\langle Y\rangle_i p \in (S,t)$ entails $M \models_{S,t} \langle Y\rangle_i p$.

The proof is by induction on the structure of strategies. It follows as a consequence of the lemmas proved earlier in this chapter. It uses the definition of $[\![\,]\!]^x$ extensively, which is the key primitive in the semantic definition of $\langle Y\rangle_i p$. It also uses the fact that $M \models_{S,t} q$ iff $M \models_t q$, which is a consequence of semantic definition SEM-15 of section 2.1.3, which applies since $q \in \mathcal{L}^i$.

For axiom AX-INT-1, $M \models_{S,t} \langle \textbf{skip}\rangle_i p$ iff $(\exists t' : [S;t,t'] \in [\![\textbf{skip}]\!]$ and $M \models_{S,t'} p)$. But, $[S;t,t'] \in [\![\textbf{skip}]\!]$ iff $t = t'$. Therefore, $M \models_{S,t} \langle\textbf{skip}\rangle_i p$ iff $M \models_{S,t} p$. This accounts for axiom AX-INT-1.

By the definition of $[\![\,]\!]$, $(\exists t' : [S;t,t'] \in [\![Y_1;Y_2]\!]$ and $M \models_{S,t'} p)$ iff $(\exists t' : (\exists t'' : [S;t,t''] \in [\![Y_1]\!]$ and $[S;t'',t'] \in [\![Y_2]\!])$ and $M \models_{S,t'} p)$. Therefore, $M \models_{S,t} \langle Y_1;Y_2\rangle_i p$ iff $(\exists t'' : [S;t,t''] \in [\![Y_1]\!]$ and $(\exists t' : [S;t'',t'] \in [\![Y_2]\!])$ and $M \models_{S,t'} p)$. By induction on the structure of strategies, this is identical to $(\exists t'' : [S;t,t''] \in [\![Y_1]\!]$ and $M \models_{S,t''} \langle Y_2\rangle_i p)$. By the same induction, this holds iff $M \models_{S,t} \langle Y_1\rangle_i\langle Y_2\rangle_i p$. This takes care of axiom AX-INT-2.

Similarly, $[S;t,t'] \in [\![\textbf{if } q \textbf{ then } Y_1 \textbf{ else } Y_2]\!]$ iff $(M \models_t q$ and $[S;t,t'] \in [\![Y_1]\!])$ or $(M \not\models_t q$ and $[S;t,t'] \in [\![Y_2]\!])$. Therefore, $M \models_{S,t} \langle\textbf{if } q \textbf{ then } Y_1 \textbf{ else } Y_2\rangle_i p$ iff $(\exists t' : (M \models_t q$ and $[S;t,t'] \in [\![Y_1]\!])$ or $(M \not\models_t q$ and $[S;t,t'] \in [\![Y_2]\!])$ and $M \models_{S,t'} p)$. Which is identical to $(M \models_t q$ and $(\exists t' : [S;t,t'] \in [\![Y_1]\!]))$ or $(M \not\models_t q$ and $(\exists t' : [S;t,t'] \in [\![Y_2]\!]))$. Thus by induction, the previous expression reduces to $M \models_{S,t} q \wedge \langle Y_1\rangle_i p$ or $M \models_{S,t} \neg q \wedge \langle Y_2\rangle_i p$. But this condition is equivalent to $M \models_{S,t} (q \rightarrow \langle Y_1\rangle_i p) \wedge (\neg q \rightarrow \langle Y_2\rangle_i p)$. This takes care of axiom AX-INT-3.

By Lemma 3.5, $M \models_{S,t} \langle\textbf{while } q \textbf{ do } Y_1\rangle_i p$ iff $M \models_{S,t} \langle\downarrow_t(\textbf{while } q \textbf{ do } Y_1)\rangle_i\langle\uparrow_t(\textbf{while } q \textbf{ do } Y_1)\rangle_i p$. If $M \models_t q$ and $\downarrow_t Y_1 \neq \textbf{skip}$, then this reduces to $M \models_{S,t} \langle\downarrow_t Y_1\rangle_i\langle(\uparrow_t Y_1);(\textbf{while } q \textbf{ do } Y_1)\rangle_i p$. By Lemma 3.3, we obtain $M \models_{S,t} \langle\downarrow_t Y_1\rangle_i\langle\uparrow_t Y_1\rangle_i\langle\textbf{while } q \textbf{ do } Y_1\rangle_i p$. By Lemma 3.5, this reduces to $M \models_{S,t} \langle Y_1\rangle_i\langle\textbf{while } q \textbf{ do } Y_1\rangle_i p$. If $M \models_t q$ and $\downarrow_t Y_1 = \textbf{skip}$, then $M \not\models_{S,t} \langle\textbf{while } q \textbf{ do } Y_1\rangle_i p$, so this case is also taken care of. Lastly, if $M \not\models_t q$, then $\downarrow_t Y = \textbf{skip} = \uparrow_t Y$. Hence, $M \models_{S,t} \langle\textbf{while } q \textbf{ do } Y_1\rangle_i p$ iff $M \models_{S,t} p$. Thus, in all cases, $M \models_{S,t} \langle\textbf{while } q \textbf{ do } Y_1\rangle_i p$ iff $M \models_{S,t} (q \rightarrow \langle Y_1\rangle_i\langle\textbf{while } q \textbf{ do } Y_1\rangle_i p) \wedge (\neg q \rightarrow p)$, as desired.

Finally, $M \models_{S,t} \langle\textbf{do}(q)\rangle_i p$ iff $(\exists t' : [S;t,t'] \in [\![\textbf{do}(q)]\!]$ and $M \models_{S,t'} p)$. If $M \models_t q$, then by definition AUX-6, $t = t'$ and $M \models_{S,t} p$. If $M \not\models_t q$, then by definition AUX-6, $t < t'$. By coherence constraint COH-4, $(\exists a, t_0 : t < t_0 \leq t'$ and $[S;t,t_0] \in [\![a]\!]^x$ and $(\forall t_1 : [S;t,t_1] \in [\![a]\!]^x \Rightarrow t_1 \leq t_0)$ (here x is the given agent, usually elided).

We choose the maximal t_0, so that by coherence constraint COH-5, only a finite number of applications of this axiom will suffice. Using coherence constraint COH-5, for each agent, we can associate with any period the number of actions that are taken by that agent over that period. This number can serve as a metric for mathematical induction, since it decreases over the subperiods of a period and equals zero for the trivial period.

By Lemma 3.1, $[S; t_0, t'] \in [\![\mathsf{do}(q)]\!]$), which entails that $M \models_{S,t_0} \langle\mathsf{do}(q)\rangle_{\imath}p$. Hence, $M \models_{S,t} (\bigvee a : \langle a\rangle\langle\mathsf{do}(q)\rangle_{\imath}p)$. Therefore, $M \models_{S,t} \langle\mathsf{do}(q)\rangle_{\imath}p$ iff $M \models_{S,t} q \wedge p$ or $M \models_{S,t} \neg q \wedge (\bigvee a : \langle a\rangle\langle\mathsf{do}(q)\rangle_{\imath}p)$. □

3.3 Properties of Intentions

Several interesting properties of intentions may be obtained from the definition given above, especially when certain intuitively nice constraints are imposed on the models. Different constraints may be chosen depending on the purpose one has and the precise concept one requires. Certain particularly important constraints are discussed in Chapter 5, where conditions leading to the success of an agent with his intentions are formalized.

I-CONS-1. **Satisfiability:**

$x\mathsf{I}p \rightarrow \mathsf{EF}p$

This says that if p is intended by some agent, then it occurs eventually on some scenario. That is, the given intention is satisfiable on some future. This does not hold in general, since the strategies assigned to the agents may be unexecutable. If a strategy assigned to an agent is unexecutable, then $x\mathsf{I}\,\mathsf{false}$ holds. The simplest such strategy is $\mathsf{do}(\mathsf{false})$. The desired constraint may be expressed as below. It essentially corresponds to the requirement that one of the executions of one of the actions of the agent be on a scenario on which p occurs.

$x * Y \Rightarrow \mathsf{E}x\langle Y\rangle_{\imath}\mathsf{true}$

I-CONS-2. **Temporal Consistency:**

$(x\mathsf{I}p \wedge x\mathsf{I}q) \rightarrow x\mathsf{I}(\mathsf{F}p \wedge \mathsf{F}q)$

This says that if an agent intends p and intends q, then he (implicitly) intends achieving them in some temporal order: p before q, q before p, or both simultaneously. This holds in general because the function $\mathbf{Y}$ assigns exactly one strategy to each agent at each moment. Thus if both p and q, which are scenario-formulae, occur on all scenarios on which that strategy is performed, then they occur in some temporal order on each of those scenarios. The formula $(\mathsf{F}p \wedge \mathsf{F}q)$ is true at a moment on a scenario precisely when p and q are true at (possibly distinct) future moments on the given scenario.

I-CONS-3. **Persistence does not entail success:**

$\mathsf{EG}((x \mathsf{I} p) \wedge \neg p)$ is satisfiable

This is quite obvious intuitively: just because an agent persists with an intention does not mean that he will succeed. Technically, two main ingredients are missing. The agent must know-how to achieve the intended condition and must act on his intentions. I include this here to point out that in the theory of Cohen & Levesque, persistence is sufficient for success [1990, p. 233]. This is a major weakness in any theory of intentions [Singh, 1992a]. Secondly, the need to state the conditions under which an agent can succeed with his intentions is one of the motivations for the concept of know-how, which is formalized in Chapter 4.

I-CONS-4. **Limiting choices:**

I discussed above the causal role that intentions are sometimes taken to have in getting an agent to act. A related idea is that intentions limit an agent's choices [Bratman, 1987, pp. 44–45]. Thus the following constraint, which I do not accept, may seem intuitively quite plausible. At any moment, the only actions that an agent may perform are those which can lead to the satisfaction of the current part of his strategy. There need be no guarantee, however, that any of those actions would ensure success.

$$(\mathbf{Y}(x,t) = Y \text{ and } \downarrow_t Y = \mathbf{do}(q)) \Rightarrow (\forall S \in \mathbf{S}_t, t' \in S, a \in \mathcal{B} : [S;t,t'] \in [\![a]\!]^x \Rightarrow (\exists S'', t'' \in S'' : [S;t,t''] \in [\![a]\!]^x \text{ and } M \models_{S'',t} \mathsf{F}q))$$

This violates the weak determinism constraint, COH-8, of section 2.3. Although it does not validate the statement that intentions entail ability, it validates something almost as unintuitive. Under some additional but fairly weak assumptions, it validates the claim that an agent who intends p is not able to achieve something incompatible with p. For example, assume that $\mathsf{AG}(p \rightarrow \mathsf{AG}p)$ holds in every moment in the model. Then the above constraint ensures that as long as $x * \mathbf{do}(p)$ holds, x is unable to achieve $\mathsf{AG}\neg p$ at t. While it may be worth considering whether an agent with a certain intention will in fact achieve something else, it is surely too strong to require that he loses all ability to do so the moment he adopts that intention. In contrast to constraint I-CONS-1, this applies to all actions of the agent. The acceptable sense of what an agent will achieve in fact is captured by the notion of real scenarios in Chapter 5.

I-CONS-5. **Persist while succeeding:**

If $\downarrow_t Y \neq \textbf{skip}$, then $M \models_t x * Y \rightarrow (\mathsf{A}[\downarrow_t Y]_i(x* \uparrow_t Y)) \wedge \mathsf{A}([\downarrow_t Y]_i\mathsf{true} \rightarrow ((x * Y \vee x* \uparrow_t Y)\mathsf{U}x* \uparrow_t Y))$

This constraint is a possible restriction on the architectures of agents. It requires that agents desist from revising their strategies as long as they are able to proceed properly. If $\downarrow_t Y = \textbf{skip}$, then the strategy is over and this constraint does not apply. Many robots and planners are designed to satisfy this constraint: plans are revised only on failure. The present approach allows the agents to benefit from opportunities that might arise unexpectedly, albeit in a limited way. This is because agents' strategies are composed of abstract actions. As a result, if a condition comes to hold fortuitously, or if a simple course of action becomes feasible because of the actions of other agents, then the given agent can take advantage of these opportunities, without having to revise his strategy.

$(\mathbf{Y}(x,t) = Y$ and $\downarrow_t Y \neq \textbf{skip}$ and $[S;t,t'] \in [\![\downarrow_t Y]\!]^x) \Rightarrow \mathbf{Y}(x,t') = \uparrow_t Y$ and $(\forall t'' : t \leq t'' < t' : \mathbf{Y}(x,t'') = Y)$

I-CONS-6. **Absence of closure under beliefs:**

$x\mathsf{I}p \wedge x\mathsf{BAG}(p \rightarrow q) \wedge \neg x\mathsf{I}q$ is satisfiable

This holds only in general, since intentions are determined independently of the agents' beliefs. As remarked in item I-DIM-7 of section 3.1, closure under beliefs is sometimes called the *side-effect* problem [Rao & Georgeff, 1991a]. By the present result, the proposed theory avoids this problem.

I-CONS-7. **Consistency with beliefs about future possibility:**

$x\mathsf{I}p \wedge x\mathsf{B}\neg\mathsf{EF}p$ is not satisfiable

This holds only in the presence of the following constraint, which can be readily imposed on the models.

$M \models_t x\mathsf{I}p$ implies that $(\exists t' : (t,t') \in \mathbf{B}(x)$ and $M \models_{t'} \mathsf{EF}p)$

I-CONS-8. **Non-entailment of beliefs about future possibility:**

$x\mathsf{I}p \wedge \neg x\mathsf{BEF}p$ is satisfiable

This holds in general. Indeed, it holds on each moment at which the following constraint applies.

$M \models_t x\mathsf{I}p$ implies that $(\exists t' : (t,t') \in \mathbf{B}(x)$ and $M \models_{t'} \neg\mathsf{EF}p)$

Constraints I-CONS-7 and I-CONS-8 are not accurate formalizations of the properties of intentions discussed in items I-DIM-11 and I-DIM-12 of section 3.1. This is because the beliefs involved in those properties are not beliefs about all possible futures, as treated in constraints I-CONS-7 and I-CONS-8, but rather are beliefs about the real future. Therefore, a better formalization of these properties is in terms of the scenarios assigned by **R**, a component of the formal model introduced in section 2.1.2.

I-CONS-9. **Consistency with beliefs about reality:**

$\neg(x\mathsf{I}p \wedge x\mathsf{B}\neg\mathsf{RF}p)$

This property holds in the presence of the following constraint. This constraint may be understood as requiring that an agent with an intention considers at least one alternative in which that intention is realized. One might think of this as the hopeful alternative, which makes it worthwhile for the agent to undertake his intention.

$M \models_t x\mathsf{I}p$ implies that $(\exists t' : (t, t') \in \mathbf{B}(x)$ and $M \models_{t'} \mathsf{RF}p)$

I-CONS-10. **Non-entailment of beliefs about reality:**

$x\mathsf{I}p \wedge x\neg\mathsf{BRF}p$ is satisfiable

This property holds at all those moments at which the following constraint is true. This constraint may be understood as meaning that the agent with an intention considers at least one alternative moment from which that intention is not realized. One might think of this as the cautious alternative, which the agent may believe may be realized unless he exercises his know-how and acts to achieve his intention. Know-how is discussed in Chapter 4 and related to intentions in Chapter 5.

$M \models_t x\mathsf{I}p$ implies that $(\exists t' : (t, t') \in \mathbf{B}(x)$ and $M \models_{t'} \neg\mathsf{RF}p)$

I-CONS-11. **Entailment of belief in possible success:**

$x\mathsf{I}p \rightarrow x\mathsf{BEF}p$

This is the opposite of the property mentioned in item I-CONS-8 above. However, the justification for it is perhaps apparent from the discussion above. Even though an agent may not believe that his intention will succeed, he should surely believe that it may succeed. The incompleteness of one's beliefs about the future should not preclude that. For, an agent would have to be quite irrational if he did not even believe it possible that his intention would succeed but still continued to hold it. This property, which is also validated in the approach of [Singh & Asher, 1993], holds when the following constraint applies.

$$M \models_t x\mathsf{I}p \text{ implies that } (\forall t' : (t,t') \in \mathbf{B}(x) \text{ and } M \models_{t'} \mathsf{EF}p)$$

Postulates I-Cons-9 and I-Cons-10 are jointly termed the *asymmetry thesis* by Bratman [1987, pp. 38]. He argues that they are among the more basic constraints on the intentions and beliefs of rational agents.

3.4 Desires

The concept of desires, which is quite closely related to that of intentions, has been studied extensively in the literature. Earlier works, e.g., [Davidson, 1980], attempted to reduce intentions to desires or to desires combined with beliefs. Such attempts are not considered viable any more [Brand, 1984; Bratman, 1987]. However, desires themselves are still often considered, especially in the computer science literature on the subject [Georgeff, 1987].

Desires are different from intentions in that they have a weaker connection with an agent's rationality. For example, the desires of an agent may be mutually inconsistent and may be inconsistent with that agent's beliefs. Also, an agent who desires something may not desire the necessary means for achieving it. These properties contrast with the relevant properties of intentions, as discussed in items I-Dim-4, I-Dim-5, and I-Dim-8 in section 3.1. The lack of these properties make desires *per se* difficult to relate with an agent's actions. Even the strongest desires of agents, which if unique would trivially be mutually consistent, lack the other two properties and are, therefore, of limited value in the specification of artificial multiagent systems.

We could require that an agent have certain desires in certain conditions, but then we could not use them to constrain his behavior significantly: if the agent's desires are mutually contradictory, or inconsistent with his beliefs, we can hardly state any rationality constraints according to which he might successfully act. The concept of desires could of course be used in general economics-style theories of rationality to relate an agent's desires with his intentions. No such theory is available at present. Thus, given the state of the art, the concept of desires is not of sufficient utility in the study of multiagent systems to merit treatment besides intentions. Therefore, I concentrate my attention on the more restrictive concept of intentions. Ultimately, formalizations such as the present one might shed some light on the form that a theory of desires ought to take for application in multiagent systems.

Sometimes, *goals* are identified with desires. When that is the case, they suffer from all the problems described above. When it is not the case, they are readily subsumed by strategies. This point is developed in greater detail in section 2.5.

3.5 Other Formal Theories of Intentions

In this section, I briefly review the computer science literature on intentions. The concept of goals, which is related to intentions, has been studied since the early days of AI [Georgeff, 1987]. However, much of this work has been architectural in nature. Here I consider only the formal theories of intentions. One of the earliest works on intentions is that of Allen who attempts to reduce intentions to beliefs about future actions [1984, pp. 145–146]. As discussed in item I-DIM-13 in section 3.1 above, this view turns out to be problematic. Allen's main focus in that paper is not on intentions, however, and his paper is better known for its other contributions.

Cohen & Levesque formalize intentions in a framework of discrete linear-time logic with precisely one event between any two successive instants [Cohen & Levesque, 1990]. Each linear fragment of time is isomorphic to the integers and is considered as a possible world. Primitive alternativeness relations are defined for beliefs and goals. A *persistent goal* of an agent is a proposition such that (a) the agent believes it to be false, (b) the agent has it as a goal, and (c) the agent will not give up that goal until he either comes to believe (i) that it is true, or (ii) that it will never become true. Intentions are then defined as special kinds of persistent goals: an agent intends p iff he has a persistent goal for the following condition: (a) he performs an event sequence, e, after which p holds, and (b) before performing e he believes there is an event sequence, e', that he will do immediately and at the end of which p will hold, and (c) he does not have a goal for the negation of the following: e happens followed by p (p. 248).

The nesting of the definitions makes Cohen & Levesque's theory the most complicated of the works on intentions. Their theory also suffers from several conceptual and technical shortcomings. Their definition confuses the semantics of intentions with constraints on intention revision. Thus the concept of intentions is tied to a particular policy of intention revision. A technical shortcoming is that certain properties are stated as "easy to see," but it is possible to construct counterexamples to these properties in the theory itself. I have developed these arguments in greater detail elsewhere [Singh, 1992a]. Cohen & Levesque's success theorem [1990, p. 233] is discussed in Chapter 5.

Seel has proposed a formalization of intentions systems theory [Seel, 1989]. Seel's theory includes a modal logic of beliefs and wants in a framework of discrete linear time. His aim is to capture the essential aspects of certain behavioral experiments. Seel's work shares some intuitions with the present approach. But there are some significant differences.

Seel assumes a set of *world axioms.* Agents are assumed to have perfect memory: they know all ***strict-past*** formulae, i.e., formulae not involving future time operators. The agents' knowledge at any time is given by the world axioms and the strict-past formulae that are true at that time (pp. 22–23). The agents' behavior is also determined by a set of axioms. These, along with world axioms, yield the agents' wants (pp. 25–26). A consequence of these definitions is that knowledge implies wants (p. 28), which means that wants are closed under knowledge (this inference was discussed in item I-CONS-6 above). Another, more troublesome, consequence is that a formula that is wanted in a given state must hold in that state (p. 28). Thus agents are guaranteed to succeed with their wants. This is at odds with our pretheoretic intuitions about wants or intentions. Seel's framework cannot accommodate changing wants, since that would require the axioms describing an agent's behavior to change, and they cannot. But Seel obtains useful results on how agents may acquire knowledge about their environment.

Rao & Georgeff have also recently proposed a theory of intentions [Rao & Georgeff, 1991b; Rao & Georgeff, 1991a]. While their main theory is based on branching-time logic, they seem to be neutral as to the distinction between branching-time and linear-time frameworks: they include formalizations and results for each kind. The best features of Rao & Georgeff's theory are the following: (a) it does not validate closure of intentions under beliefs and (b) it satisfies the asymmetry thesis. These issues were discussed items I-CONS-6, I-CONS-9, and I-CONS-10 in section 3.3 above.

Rao & Georgeff identify goals with desires and give a semantics to goals based on a primitive alternativeness relation. This is problematic. The given semantics ensures that the goals of each agent are mutually consistent. But, as discussed in section 3.4, the main property of desires is that they need not be mutually inconsistent, and indeed often are not. Desires need not even be believed to be consistent by the given agent. And, agents are not constrained by rationality to make their desires consistent. This is one of the major differences between desires and intentions. Thus goals, as formally defined, cannot be identified with desires. Therefore, the formalization of goals as a separate concept from intentions is not well-justified. Rao & Georgeff argue that this allows them to keep an agent's goals (i.e., ends) separate from his intentions (i.e., means) [Rao & Georgeff, 1991a, p. 6]. But this argument is not entirely satisfactory. A rational agent may need to perform means-ends reasoning to an arbitrary level of nesting. A theory should not require a separate concept for each such level of reasoning. Some related aspects of Rao & Georgeff's approach are discussed in Chapter 5.

Werner has also proposed a theory of intentions [Werner, 1991]. According to his definition, an agent, A, intends to X, if (a) A has a general plan to X, and (b) A's actions are guided by that plan (p. 119). A *general plan* is defined as a class of game-theoretic strategies (thus the term *strategy* is used differently than in the present approach). A strategy is a function from the agent's information state to an action. The agent may pick actions according to any of the applicable strategies in the given class. This makes it possible to partially specify actions. Note that this effect is attained more simply in the present approach by allowing abstract specifications, such as **do**(q). Also, the present approach makes the relevant conditions explicit in the conditional and iterative forms.

Although Werner's models involve time, his formal language does not. He gives no postulates relating intentions and beliefs, or intentions and time. A counterintuitive component of his definitions is that they do not distinguish between past, present, and future. Thus a strategy may be for a condition in the past. This makes it possible for agents to intend past conditions, which is problematic if intentions are meant to lead to action. More significant, perhaps, is the following problem. By the given definitions, agents cannot revise their strategies within the model. Entire histories are compared with strategies to test for compatibility. Thus if an agent is in the same information state at two points in a history, he must behave the same way at both, relative to whatever strategy he may have. Consequently, the important notions of persistence of intentions or how they are updated, which Werner also considers as important (p. 119), cannot be studied within his own framework.

Werner requires that if an agent intends something, then he must believe that he can achieve it. He also states that an intention presupposes the corresponding ability (pp. 110, 119). These requirements are unnecessarily strong. Agents often intend to achieve conditions of which they cannot guarantee the success. Indeed, they may not even believe that they will succeed with their intentions. This point was discussed in items I-CONS-8 and I-CONS-10 in section 3.3. Some other aspects of Werner's approach are discussed in Chapter 5.

The above approaches are all modal in nature and are based on possible worlds models. In joint work with Nicholas Asher, I have also developed a different logic of intentions [Singh & Asher, 1993]. That logic is based on Kamp's Discourse Representation Theory [Kamp, 1984] and seeks to be cognitively more accurate than any of the modal approaches. For example, it rightly invalidates the inference that an agent's intentions must be closed under logical equivalence. However, as discussed in item I-DIM-6 of section 3.1 above, doing so takes us away from the intentional stance and into the design stance. This makes it harder to apply that theory in the specification of multiagent systems. I return to this point in Chapter 7.

3.6 Philosophical Remarks

In the definitions given above, an agent's intentions are determined on the basis of his strategies and on how he may act on them in a formal model consisting of possible states of the world. Thus an agent's (possible) actions are significant to the process of assigning intentions to him. In this sense, the proposed approach is *pragmatist*, as that term is defined by Stalnaker [1984, pp. 15–19]. Pragmatism is the philosophy behind the logics that are based on "possible worlds" models. Such models arise in a number of formal theories besides the one proposed here, e.g., in the logic of knowledge of [Halpern & Moses, 1987] and in dynamic logic as surveyed in [Kozen & Tiurzyn, 1990].

The main technical consequence of considering possible states of the world is that agents' intentions are closed under logical equivalence. However, agents can have intentions they do not act on and, of course, those that they act on but fail with. While the agents' intentions are not associated with their actual actions, they are associated with their possible actions. Thus if an agent intends p, he automatically intends q, where q is logically equivalent to p. As discussed in item I-DIM-6 in section 3.1, while this is not always desirable, it is quite all right for many of our purposes.

Another point that is sometimes raised is about whether agents can really have intentions. An alternative view, seemingly quite plausible when one is talking of artificial entities, is that they do not have any intentions of their own, but merely reflect their designer's "intent." Leaving issues of philosophy aside, there are technical and pragmatic reasons for not taking such a view seriously in one's theorizing:

- Intelligent systems are conceivable that have no unique designer or, at least, whose intentions can be attributed to no unique designer. Examples of such systems include markets, which have evolved into complex entities and of which we can speak in intentional terms.

- The designer's intent would involve *types* of states (e.g., "in conditions of heavy local load, the agent should intend to obtain another agent's assistance"). However, the intentions of the agents, as we need that concept, involve *tokens* or specific conditions (e.g., "the agent intends now to request its nearest neighbor's help").

- The designer usually is not around to supervise the functioning of the given system. If a system can be considered autonomous, it can, and must, be ascribed intentions of its own. The states of the environment an agent faces can include those that were not anticipated by the designer. Additionally, if the designer's intent really mattered, no system would ever perform incorrectly. Therefore, we must evaluate agents by their possible actions and validate claims about their intentions accordingly.

3.7 Conclusions

Strategies allow us to succinctly describe the relevant aspects of the agents' design. These strategies yield the intentions of the agents in a simple and direct way. This is quite important. Usually, modal approaches to intentions simply postulate a primitive alternativeness relation that captures the relevant dispositions of the given agents. However, it is not clarified how such a relation may be implemented in the agent's design. When strategies are used as proposed here, this connection is at once rigorous and obvious. It also allows us to state natural model-theoretic constraints that capture important properties of the architectures of our agents.

The definition of intentions proposed here differs from most other approaches in at least one other respect: the proposed definition is separated in the logic from beliefs and knowledge. Clearly, an agent may adopt or update his intentions based on his knowledge. I do not prevent this. However, the concept of intentions is formalized so that the fact of whether or not an agent intends a particular proposition is largely independent of what he believes or knows (modulo the constraints introduced in section 3.3). A consequence of this separation is that certain counterintuitive inferences (pertaining to the entailment of beliefs or ability) that the other approaches validate are quite naturally prevented here.

The proposed approach also highlights the additional assumptions or constraints needed to ensure the success of an agent's intentions. To be assured success, an agent must have the relevant skills or basic actions as well as possess the knowledge needed to select his actions and to identify the conditions he intends to achieve. These are necessary, but not sufficient, conditions. The above prerequisites may be bound and studied together as one concept: know-how. This is the subject of the next chapter.