Chapter 7

Conclusions and Future Work

I have presented formalizations of the concepts of know-how and intentions as they are needed in the study of multiagent systems. These formalizations capture many of our pretheoretic intuitions about these concepts. For example, an agent's intending something does not entail that he knows how to achieve it; and, his knowing how to achieve something clearly does not entail that he should intend it. The agents' intentions, along with their knowledge and know-how, constrain the actions they choose to perform. Agents who know how to achieve what they intend can succeed provided they (a) act on their intentions, (b) exercise their know-how, and (c) persist with their intentions long enough.

The formalizations of know-how and intentions were developed in a model of actions and time that is among the more general ones considered in computer science. It unifies temporal and dynamic logics by relating actions and time. Further, it allows several agents to act concurrently and asynchronously. Thus, it admits games with simultaneous moves, i.e., those in which there is no notion of turn-taking. The concept of strategies is revived and formalized as a means for abstractly specifying the behavior of agents. Strategies are then used to unify intentions and know-how and to state constraints on the selection of actions by agents.

Intentions and know-how can be used also to assign a formal semantics to communications among agents. Such a semantics for communications is proposed. This semantics brings to fore the role of communications in influencing the actions of the participating agents. A communication is satisfied only if it has the requisite effects on the intentions of the agents involved and if those agents have the necessary know-how and act appropriately. The unified theory of intentions, know-how, and communications can be used to formalize constraints on the behavior of agents. These constraints are both abstract and in terms closer to the requirements of the ultimate human users of the multiagent system being designed.

Perhaps the key limitation of the approach proposed here results from its being an *intentional stance* or *knowledge level* approach. This causes it to be a "radically incomplete" approximation in Newell's term [1982, p. 111]. In other words, this approach does not faithfully model the limitations of reasoning of real-life systems. As Newell himself argues, there is still much to be said in favor of a knowledge level approach. It provides us with a set of abstractions for artificial systems, a set of abstractions that we use with great success in our daily lives. As long as a designer of multiagent systems is sensitive to the limitations of his agents, he can indeed use the proposed approach effectively.

One way to avoid the abovementioned limitation is to consider one of the representational approaches to beliefs and intentions. Such approaches explicitly consider the representations that agents may have and the computations they may engage in. Many of these approaches, the so-called sentential ones [Konolige, 1986], do not associate any real semantics to the elements of their formal languages. This is a major shortcoming, especially when one wishes to use the resulting theory in developing a methodology for designing distributed systems. Some hybrid approaches exist that seek to avoid the problems of both the proposed framework and the sentential approaches [Fagin & Halpern, 1988; Singh & Asher, 1993]. Such approaches are technically more complex than either. By definition, they require some specification of the representations and the reasoning processes of the agents. Thus they belong in what Newell calls the *symbol level* [1982, p. 95]. Not only are they technically more complex, the hybrid approaches also require a more detailed specification before they can be put to use. Such specifications may not be available when an existing system is being analyzed. And, they may not even be available in the early stages of the design process when the basic behavioral requirements for the desired system are being determined. Thus, while an extension of the present framework to the symbol level would be desirable, the present framework itself may still be needed.

There are also a number of technical problems merit a lot of attention, but which I was not able to properly address in this monograph. Foremost among these is the incredibly hard problem of constructing models. The present work assumes that a model of some sort has been constructed and described, so that further reasoning can proceed in it. It appears likely that good models cannot be automatically generated for all but the most trivial application domains. However, much work is required in developing tools that assist in the formulation of models. Work in the broad area of problem solving in artificial intelligence will be especially pertinent here.

Another problem has to do with how to do with how agents should

revise their intentions and beliefs. There has been much work recently on belief revision [Gärdenfors, 1988]. The problem is technically harder when intentions are brought into the fold. However, in an intuitive sense, it might even become simpler because beliefs by themselves give a gloriously incomplete picture of an intelligent agent. Also, the technical results are likelier to be more natural when intentions are included.

Purely qualitative solutions to the revision problem, while possibly of some value, are not likely to be highly accurate. This is because intentions are a matter of what an agent really wants to achieve and reflect the agent's preferences as to how much effort and resources he would assign for what conditions. An instance of this is the problem of how long an agent should persist with an intention: clearly, changing one's intentions extremely frequently or never are both likely to prove inefficient and irrational. Thus, in some sense, the problem is necessarily one of probabilistic or utility-theoretic reasoning. I have studied some of these problems elsewhere, but the solutions are far from complete [1991b; 1991e]. In these works, I also argue that intentions are valuable to limited agents, those who are rational, but are not perfect Bayesian reasoners (and none are). More precise analyses that take the cost of thinking or computation into account are needed.

Another set of related ideas that I mentioned only briefly in this monograph have to do with the structural aspects of multiagent systems. A multiagent system can itself function as a single entity in another, larger, system. This is extremely common in human organizations. If we allow this, we are faced with the problem of defining the know-how and intentions of the system as a whole, so that our approach can apply at the next level up, in terms of the semantics of communications as well as other general constraints. I have explored some of these issues in past research [1991a; 1991c]. That work is not nearly as technically sophisticated as I would like, but it contains some interesting ideas. For example, the structure of a group is seen as embodied in the constraints on the communications among the different members. An open problem is to see how the semantics of communications proposed here would and should interact with the definitions of structured groups.

The theory developed in this monograph provides a framework for specifying multiagent systems, but it is clearly not sufficient in itself. To be practical, any approach to designing multiagent systems must include actual design rules or heuristics that apply in a wide variety of cases. Of course, the theory can be used to formalize the design rules, to show their correctness, and to elucidate the assumptions under which they are most naturally applicable. These are all valuable functions. However, actually coming up the rules is no trivial matter either. A detailed *design theory* is urgently needed.

Besides the uses of intentions and know-how that are of primary interest here, they have some applications in other problem domains. Notable among these are cognitive modeling, plan recognition, and machine learning of how to act. For example, it would be fruitful to design families of algorithms or even instances of them that are phrased directly in terms of the primitives formalized here. Such algorithms could be used for solving any of a variety of problems in multiagent systems, for example, those pertaining to coordination, cooperation, or collective learning. Indeed, the formalization of the contract net protocol present in Chapter 6 can be seen as taking a step in this direction. Other connections, especially those with machine learning, must be explored in greater technical detail, however.

Yet another major piece of remaining work is to develop algorithms for checking the satisfiability of specifications in proposed formal language, or some subset thereof. It is likely to be significantly more tractable to check the satisfiability relative to a given design: this kind of *model checking* has been used to great profit for the case of temporal logics [Burch *et al.*, 1990]. For this reason, the development of efficient model checking algorithms for the proposed framework would be of great use.

One can think of many more challenging research projects in multiagent systems. I only hope that I have convinced the reader that the study of formal methods in multiagent systems is not only of great practical significance, but also promises to be an exciting area for further research.