Stanford Heuristic Programming Project
Report No. HPP-84-5

March 1984
revised October 1984

Communication and Cooperation

Jeffrey S. Rosenschein
Michael R. Genesereth

COMPUTER SCIENCE DEPARTMENT
Stanford University
Stanford, California 94305

Communication and Cooperation

Abstract

Intelligent agents need to coordinate their actions in pursuit of common goals. When communication is possible, cooperating agents must decide what information to pass in order to agree on a single course of action. This paper outlines several communication strategies (under monotonic and nonmonotonic planning assumptions), proving that some are convergent while others are not. An analysis is also made of the advantages of passing false information.

§1. Introduction

Recent years have seen increasing interest in *Distributed Artificial Intelligence* (DAI) systems, that is, in groups of intelligent agents whose members cooperate in carrying out tasks. Considerable work has gone on in this area, producing a number of tentative approaches to cooperation; notable among these research efforts are Smith and Davis' work on the Contract Net [1], Davis' investigations of Cooperative Problem Solving strategies [2], Georgeff's approach to assuring non-interference among distinct agents' plans [3, 4], and Lesser and Corkill's empirical analyses of distributed computation [5].

Despite some genuine insights that these researchers have gained, however, DAI has lacked much of the formal foundation needed for progress. Recent work by Appelt [6], Moore [7, 8] and Konolige [9, 10, 11, 12] has begun to develop the formal descriptions necessary for one agent to reason about another agent's knowledge and beliefs; this is a key step in the development of successful DAI systems.

This paper begins to lay the groundwork for another aspect of Distributed Artificial Intelligence's foundation; it presents a description and analysis of information passing strategies between intelligent agents. Through use of a formal descriptive language, certain information passing behavior is proven to be convergent. In addition, an analysis is made of the role that can be played by the passing of false information, i.e., information that is logically inconsistent with the beliefs of the sender.

Consider, for example, two individuals who have lost contact with each other in a department store [13]. Both have a common interest in reuniting, but communication

may either be impossible or limited in some way (through the store's paging system, for example). To succeed, the individuals need to use tacit or explicit coordination techniques. Our concern is with the sorts of reasoning these agents would go through in planning their actions; more specifically, we are here concerned with what communication strategies such agents, with common interests, would pursue.

§2. Initial Global Assumptions

We shall consider the case of two agents, each possessing identical logical deduction capabilities and a separate database of propositions that describes the environment. In addition, each of the agents has *meta-level* knowledge regarding the other's database. As an initial assumption, we shall consider all information in our agents' world (including meta-level information) correct; that is, it describes the true state of the environment. Subsequently, we shall remove this assumption, and consider the case where the agents possess incorrect information. However, even under our initial assumption of correct information, base level and meta-level information will sometimes be incomplete.

Both agents are operating under the identical *global utility function*; they are thus motivated by exactly the same "desires" (i.e. have the same goals), and will, to the best of their abilities, cooperate to further global utility. Each agent accepts all facts that are passed by the other.

The agents use standard AI techniques [14, 15] to devise plans; specifically, each goes through a process of *theory extension*. Sets of axioms are derived by an agent such that its database, plus any of these sets, will logically imply the goals; each set prescribes an alternate executable plan of action to achieve these goals (i.e., the plan consists of a set of axioms describing the actions each agent should execute). Note that any plan developed by an agent specifies a global plan of action, that is, it specifies the actions of both agents.

The agents will communicate until they agree upon some one plan; only then will actions be taken that affect the external environment. The agents communicate using a fixed language and a fixed vocabulary. In the case where the agents' communication

overlaps, we assume that there is a way for them to determine whose message originated first (and to resolve ties through some unique ordering).

Convergent information passing strategies are defined to be strategies guaranteed to cause the agents to agree on the execution of a single plan after a finite number of communications. This paper will examine strategies of information passing, and prove that some are convergent.

§3. Notation

Our two agents are labeled agent A and agent B. A's database, consisting of propositions that describe the environment, will be denoted by D_A. Similarly, B's database will be denoted by D_B. A's meta-level propositions (that define A's knowledge of B's database) will be called $D_{B'}$, while B's propositions describing A's database will be called $D_{A'}$. D_A and D_B are assumed to be finite.

It will sometimes be necessary to refer to the total (possibly infinite) set of propositions that describe the environment; we denote this set by D_T. The global set of goals, by definition possessed by both agents, will be called G.

To denote the set of plans provable from a database D and a goal set G we write

$$\Pi_G(D).$$

In this paper we will assume that all plans are consistent with the identical goal set, and will therefore simplify our notation by using $\Pi(D)$ to denote the plan set. Using this convention, $\Pi(D_A)$ is the set of plans provable from A's knowledge of the world; similarly, $\Pi(D_B)$ is the set of plans derivable from B's knowledge. The total set of plans that are provable from the environment is $\Pi(D_T)$. This set of plans is assumed to be non-empty.

Analogously, we introduce the terms $\Pi(D_{A'})$ and $\Pi(D_{B'})$. $\Pi(D_{A'})$ is the set of plans provable from what agent B knows of A's database, and $\Pi(D_{B'})$ is the set of plans provable from what A knows of B's database.

A plan α in, for example, $\Pi(D_A)$ will generally be proved using some proper subset of the agent's database D_A. We will denote this subset of D_A that is sufficient to prove α

by $d_A(\alpha)$. The subset of D_A that is sufficient to prove *all* of the plans in $\Pi(D_A)$ will be denoted by $d_A(\Pi(D_A))$.

Having presented our notation, we now formalize, for future reference, two of the initial global assumptions that were presented in the last section.

Assumption 1. *Both agents' information about the world is correct, but may be incomplete:* $D_A \subseteq D_T$ *and* $D_B \subseteq D_T$.

Assumption 2. *Both agents' information about the other agent's database is correct, but may be incomplete:* $D_{A'} \subseteq D_A$ *and* $D_{B'} \subseteq D_B$.

§4. Convergent Strategies: The Monotonic Case

In this section, we present convergent strategies of information transfer subject to two assumptions not mentioned above. The first is monotonicity of plan generation by agents A and B. Informally, this means that if database D' is a subset of database D, then any plan that an agent generates with D' will also be generated with D. Formally, we write

$$\forall D \forall D' \Big((D' \subseteq D) \supset (\Pi(D') \subseteq \Pi(D)) \Big).$$

The second assumption is that the agents' meta-level information is limited to that expressible by ground clauses. Specifically, we assume that if agent A knows that agent B has a piece of information, A knows that information as well. To simplify matters, we might think of the agent's meta-level propositions as being represented in the same form as its base level propositions; an axiom is then meta-level due to its being in a meta-level database. Thus, direct statements can be made comparing, for example, axioms in D_A and $D_{B'}$. The following formalizes the above assumption.

Assumption 3. *If A knows that B knows some fact, then A knows the fact as well:* $D_{B'} \subseteq D_A$. *Similarly,* $D_{A'} \subseteq D_B$.

As we shall see, this assumption, along with monotonicity, has major implications on our analysis of information passing: they guarantee that if either agent has a non-empty

set of plans that are provable from its database, the two agents will eventually converge and agree on a single plan for execution. Throughout this monotonic section, we assume that $\Pi(D_A)$ and/or $\Pi(D_B)$ is non-empty.

4.1 Consequences of Monotonicity

There are several important consequences in our assuming monotonicity of plan generation. Each of these consequences is directly derivable from the definition of monotonicity and one of the assumptions presented above.

The first consequence is that, because an agent's information is correct (though partial), any plans that are provable from its database are also provable from the true state of the world.

Consequence 1. *In the monotonic case,* $\Pi(D_A) \subseteq \Pi(D_T)$ *and* $\Pi(D_B) \subseteq \Pi(D_T)$.
Substituting into our definition of monotonicity (D_A for D', and D_T for D) we have

$$(D_A \subseteq D_T) \supset (\Pi(D_A) \subseteq \Pi(D_T)).$$

$D_A \subseteq D_T$ is true by Assumption 1. The situation is analogous for $D_B \subseteq D_T$. ■

A second consequence is that any information that an agent has about another agent is "pessimistic," in the sense that the other agent can only possess more plans than the original agent suspects.

Consequence 2. *In the monotonic case,* $\Pi(D_{A'}) \subseteq \Pi(D_A)$ *and* $\Pi(D_{B'}) \subseteq \Pi(D_B)$.
Substituting into our definition of monotonicity ($D_{A'}$ for D', and D_A for D) we have

$$(D_{A'} \subseteq D_A) \supset (\Pi(D_{A'}) \subseteq \Pi(D_A)).$$

$D_{A'} \subseteq D_A$ is true by Assumption 2. The situation is analogous for $D_{B'} \subseteq D_B$. ■

Third, the plans of B that A is aware of are a subset of A's own plans.

Consequence 3. *In the monotonic case,* $\Pi(D_{B'}) \subseteq \Pi(D_A)$ *and* $\Pi(D_{A'}) \subseteq \Pi(D_B)$.
Substituting into our definition of monotonicity ($D_{B'}$ for D', and D_A for D) we have

$$(D_{B'} \subseteq D_A) \supset (\Pi(D_{B'}) \subseteq \Pi(D_A)).$$

$D_{B'} \subseteq D_A$ is true by Assumption 3. The situation is analogous for $D_{A'} \subseteq D_B$. ■

4.2 Information Passing *vs.* Plan Passing

In much of our analysis below, we will be concerned with "information passing," that is, with the transfer of sets of database propositions between agents. In general, these propositions will be sent from one agent to the other so that the latter will have sufficient information to derive plans already generated by the former. Why not instead design the agents differently, allowing the first agent to transfer a single plan, instead of transferring database propositions and forcing the other agent to duplicate deductive steps?

In our scenario, where all propositions are correct and plans are generated monotonically, it may indeed be the case that plan transfer is superior. It is certainly true that passing a plan to the other agent will cause convergence to a correct plan, i.e., one in $\Pi(D_T)$.

Theorem 1. *In the monotonic case, if each agent picks a plan that is provable from its own database and sends it to the other, the agents will converge to a plan in* $\Pi(D_T)$.

Proof: Any plan in $\Pi(D_A)$ is also in $\Pi(D_T)$ (Consequence 1); similarly, any plan in $\Pi(D_B)$ is also in $\Pi(D_T)$. If agent A sends B a plan in $\Pi(D_A)$, B will accept it (by assumption); likewise if B sends A a plan in $\Pi(D_B)$. If both agents send plans, then there is an assumed mechanism for deciding which was sent first (and resolving ties). In any case, both agents will have agreed on a plan in $\Pi(D_T)$ after (at most) one communication act each. ■

The decision as to whether information or plans will be sent in the monotonic case could be based on relative cost. It may be the case that communication is far more expensive than computation, and therefore it is worthwhile to limit communication even at the cost of the receiver's having to rederive old results. In this case, it may be cheaper to transfer a few short axioms than a substantially longer plan. Conversely, if communication is cheap, or the plan length is shorter than the axiom length, plan transfer may be superior.

Another consideration is whether we *want* to allow agents to carry out plans whose execution is not warranted by their own databases (this becomes more of an issue when information cannot be assumed correct or plans are not generated monotonically). In the

sections that follow, we assume that an agent will only execute plans that, along with its own database, imply the goals. Thus, we examine strategies that lead both agents to agree on a single plan that is provable from *both* of their databases and the goal set. Note that this does not limit the agents to pure information transfer, since it will sometimes be possible for one to transfer a plan that is already provable from the other's database.

4.3 Primitive Strategy

We presented above a simple strategy of transferring plans that was guaranteed to cause convergence. There is likewise a simple strategy of information transfer that is guaranteed to allow eventual convergence to a plan in $\Pi(D_T)$: both agents transfer their entire axiom set to the other agent.

Theorem 2. *In the monotonic case, total information transfer allows eventual convergence to a plan in* $\Pi(D_T)$.

Proof: If both agents share their databases, then both A and B end up with identical $D_{A\cup B}$ propositions. Since $D_A \subseteq D_{A\cup B}$ and $D_B \subseteq D_{A\cup B}$, and $\Pi(D_A)$ and/or $\Pi(D_B)$ is non-empty, monotonicity guarantees that $\Pi(D_{A\cup B})$ is non-empty, i.e., A and B will agree on a non-empty set of plans. Since $D_A \subseteq D_T$ and $D_B \subseteq D_T$, $D_{A\cup B} \subseteq D_T$; thus every member of this set ($\Pi(D_{A\cup B})$) is also a member of $\Pi(D_T)$ by monotonicity. Specification of a single plan in this set (if there is more than one) can then occur. If both agents specify different plans at this stage, the first message determines which will be followed. ■

Actually, a transferral of the agents' entire databases is not necessary, since not all of the axioms in a database are relevant to the plans that need to be proven (relative to a particular goal). In place of the entire database, each agent could transfer the subset of its axioms that prove all the goals of which it is aware, e.g., agent A would transfer $d_A(\Pi(D_A))$. This strategy of total relevant information transfer still allows trivial convergence to a plan in $\Pi(D_T)$, and is a more realistic strategy than attempting to transfer all of ones axioms to the other agent. While D_A can be expected to be quite large, $d_A(\Pi(D_A))$ may be a more manageable size.

Note, however, that if the agents only share information that they consider relevant, they may miss some interactions between their data. For example, if A and B both know

that P and Q imply R, but A only knows P and B only knows Q, then neither agent might consider the facts P or Q relevant in isolation (though they are in tandem) and will not send them. As long as $\Pi(D_A)$ or $\Pi(D_B)$ is non-empty, though, this missed conclusion will not matter. If it did for any reason, the agents could define "relevant axioms" as any that appear on the left side of pertinent rules, even if the rules were not used in the derivation because of missing facts. That way, fact interaction will be enabled.

4.4 Sophisticated Strategies

Still, the heavy-handed strategy of total information transfer (or even total relevant information transfer) is far from attractive. While it allows convergence, we might wonder if there are not more elegant strategies that do the same. In fact, such strategies do exist; as we shall see, they depend on the specific view that each agent has of the other's database.

Our task is to show that either agent, having any particular database and information regarding the other's database, can send a plan (or information and a plan) that is both correct and provable from that other's database.

The simplified convergent strategy is as follows. There are two actions that agent A can take (similar rules exist for agent B); they cover all possible cases (as seen from the perspective of agent A with non-empty $\Pi(D_A)$):

1. When $\Pi(D_{B'}) \neq \emptyset$, send a message of the form "I am pursuing plan β" where β is a member of $\Pi(D_{B'})$ (β will also be in $\Pi(D_A)$, from Consequence 3);

2. When $\Pi(D_{B'}) = \emptyset$, send an axiom set i such that $i \subseteq D_A$ and $D_{B'} \cup i \cup \beta \models G$, along with a message "I am pursuing plan β" (such an i will exist since $D_{B'} \subseteq D_A$ and, since $\Pi(D_A)$ is non-empty, there exists a β such that $D_A \cup \beta \models G$; at the very worst, i could simply be the axiom set such that $D_{B'} \cup i = D_A$; note also that, by monotonicity, we are guaranteed that any β chosen will be in $\Pi(D_A)$).

Because A only proposes a plan in $\Pi(D_A)$, we know that the proposed plans are correct (Consequence 1). Because A is pessimistic (Consequence 2), B will always accept the plan

that it proposes. Thus, along with our assumption that B is following analogous rules and that transmissions are ordered, we now have a strategy that converges to a correct plan. Note, however, that a convergent strategy may not exist if both $\Pi(D_A)$ and $\Pi(D_B)$ are empty. If $\Pi(D_A) = \emptyset$, agent A may do best by transferring its entire database D_A to B.

Determining axiom set i

When we discussed the axiom set i above, we simply noted that it existed—we did not mention how A might find the minimal set i and so minimize communication costs. One method would be for A to keep track of the number of axioms not in D_B that are used to generate each of the plans in $\Pi(D_A)$; the plan generation is done such that plans that require fewer non-D_B axioms are proven first (to take care of an infinite $\Pi(D_A)$). A "dovetailing" computation would be performed to avoid being caught in an infinite derivation when attempting to generate any of the plans in $\Pi(D_A)$ [16]. The first plan generated is then the one to propose to B, after sending it the relevant missing axioms i.

Note that this produces a minimum axiom set i from A's perspective. However, A may send facts to B without realizing that B already has them, since A knows only a subset of B's database (e.g., if $D_A = D_B$ but $D_{B'} = \emptyset$, A will transfer a set of axioms to B, when a simple commitment to a plan would suffice). Thus, there may in fact be a smaller set i that would work. Note also that in the discussion above we have not merely shown that strategies of convergence (short of total information transfer) exist—we have shown how an agent can operationally pursue them, a stronger result.

§5. Convergent Strategies: The Nonmonotonic Case

Nonmonotonicity means that, even if a plan and a database D' imply a goal, and there exists a D such that $D' \subseteq D$, the plan and D may not imply the goal. In relaxing the assumption of monotonicity that played such a crucial role in our analysis above, we discover a radically different convergence situation. In particular, we find that there is no strategy of information transfer that is guaranteed to converge onto a correct plan, i.e., a plan in $\Pi(D_T)$, even if $\Pi(D_A)$ and/or $\Pi(D_B)$ are non-empty.

The problem we encounter is the unpredictable nature that information might play on plan generation in the nonmonotonic case. It is possible that B, with partial information, can generate all the plans in $\Pi(D_T)$; after one axiom is received from A, the set of B's plans could become empty, while after receipt of another axiom, the set of plans could become equal to $\Pi(D_T)$ again. It is not difficult to write nonmonotonic rules that give rise to such oscillating behavior, and it makes general analysis of strategies very difficult. Since A cannot be sure of all of B's information, it cannot always avoid these problems.

5.1 Information Passing *vs.* Plan Passing

New arguments appear in favor of information passing over plan passing in the nonmonotonic case. First of all, simply passing a plan from A to B (or vice versa), a strategy that guaranteed convergence to a correct plan in the monotonic case, may fail here.

Theorem 3. *In the nonmonotonic case, if each agent picks a plan that is provable from its own database and sends it to the other, the agents may not converge to a plan in* $\Pi(D_T)$.
Proof: Even if both agents have identical databases, and thus identical plan sets, their plans need not be in $\Pi(D_T)$. Thus they could trivially agree on the communicated plans, and still be agreeing on a plan outside of $\Pi(D_T)$. ■

In addition, there now appear other advantages to information passing over plan passing (besides potential cost advantages). For example, consider the case where A believes that by passing axiom set i to B, the latter will derive some plan, but in fact this is not so (because B has more information that A thinks it has). If B were simply presented with the plan, it would know only that A derived it in good faith, but would not have any idea of the facts that supported it; being presented with some facts that A considered convincing evidence in favor of the plan would allow B to make a more appropriate response.

Unfortunately, however, for arbitrary D_A and D_B there need not be any guaranteed strategy of information transfer that will cause A and B to converge to a correct plan.

Theorem 4. *In the nonmonotonic case, there is no guaranteed strategy of information transfer for arbitrary* D_A *and* D_B *that will cause* A *and* B *to converge to a plan in* $\Pi(D_T)$, *even if* $\Pi(D_A)$ *and* $\Pi(D_B)$ *are non-empty.*

Proof: For many universes D_T, it is possible to find a non-empty subset U of axioms that can be used to derive a set of plans, none of which are in $\Pi(D_T)$; we then assign U as D_A and D_B. Under these circumstances, no matter what information transfer goes on, A and B will not converge to a plan in $\Pi(D_T)$. ■

5.2 Correct Convergence *vs.* Acceptable Convergence

Faced with the prospect that no plan passing or information passing strategies will necessarily converge to plans in $\Pi(D_T)$, we wish to find an alternate definition of convergence, one more reasonable for the nonmonotonic case. One such acceptable definition is that A and B should agree on a plan in $\Pi(D_{A\cup B})$. Though other convergence definitions are plausible (e.g., convergence to plans in $\Pi(D_A) \cup \Pi(D_B)$ or in $\Pi(D_A) \cap \Pi(D_B)$), this definition mirrors a "most-informed" single agent: the agents will accept plans that would also be accepted by a single agent possessing all of their knowledge. We call this correspondence to the single-agent case the "single-agent isomorphism." Intuitively, it is as close as we can come in the nonmonotonic case to having agents converge to correct plans.

There is an aspect of safety in using the "single-agent isomorphism" as the convergence criterion. In the real world, agents may be operating with widely varying models. Unless they are restrained, they may pursue plans that, while warranted by their own knowledge, are not warranted by others' knowledge. By requiring plans to satisfy the union of agents' databases, we are making sure that actions conform to *all* the knowledge that can be brought to bear on the situation. This is a conservative approach.

5.3 Primitive Strategy

Armed with our new definition of "acceptable convergence," there is a primitive information transfer strategy that trivially guarantees it: both agents transfer their entire databases to the other. Then both A and B will have identical $D_{A\cup B}$ facts, and will agree on plans in the set $\Pi(D_{A\cup B})$, if any exist. Note once again that these plans may not be in $\Pi(D_T)$; in fact, before the information transfer both agents might conceivably have possessed plan sets equivalent to $\Pi(D_T)$, only to see their plan sets change when the new information arrived. At least they will now agree on a (possibly empty) set of plans,

isomorphic to what would be generated by a fully-informed single agent.

A variation on this strategy is one of total relevant information transfer, where each agent sends those axioms that it feels are relevant to the plans it has generated (e.g., A sends $d_A(\Pi(D_A))$). It is necessary here, however, to define "relevant" in a slightly broader sense, as we did above. Each agent needs to send *all* facts that appear on the left-hand side of pertinent rules (and that it believes), even if the rules were not used in the agent's actual derivation, since these facts, combined with the other agent's facts, could have unforeseen consequences in the nonmonotonic planning world (allowing new plans to be derived, or disproving old plans that had been derived).

5.4 Sophisticated Strategies

Unfortunately, there is little more that can be presented in the way of other nonmonotonic strategies that guarantee acceptable convergence. There is, however, one special situation that warrants mentioning: the case where we are guaranteed that D_B will always be a subset of D_A. This will be true, for example, when $D_A = D_T$ (the $D_A \subseteq D_B$ case is analogous).

With this design assumption, it may be possible for A to successfully "convince" B of one of the plans in $\Pi(D_A)$. Since $D_B \subseteq D_A$ this will certainly be an acceptable convergence (i.e., $\Pi(D_A) = \Pi(D_{A \cup B})$ under these circumstances); if $D_A = D_T$ this will actually be a correct convergence, that is, a convergence to a plan in $\Pi(D_T)$.

One main assumption must be made: A is assumed to know all the nonmonotonic rules [17, 18] under which B is operating, and the ordering in which B will use them.

Theorem 5. *In the nonmonotonic case, if D_B is guaranteed to be a subset of D_A, and A knows the order and content of B's nonmonotonic rules, there is an acceptable convergent strategy for A.*

Proof: Note that we are not concerned here with $D_{B'}$, but rather with D_B itself; A will effectively have to ignore the contents of $\Pi(D_{B'})$ because, due to insufficient knowledge, it may bear absolutely no relation to $\Pi(D_B)$. However, A *can* assume that any information it has about B is correct. By examining B's nonmonotonic rules in order, A can generate a finite list of axioms that B needs in order to accept any particular plan in $\Pi(D_A)$; any

members of this list that $D_{B'}$ does not indicate B already has are then sent to B. Although B may possess other information, because of the ordering on the nonmonotonic rules it can have no effect on the plan that B derives using the earlier rules; thus the plan will be agreed upon by both agents, and the strategy is convergent. Since this plan is in $\Pi(D_A)$ and $D_{A \cup B} = D_A$, the plan is also acceptable. ■

§6. Agents Possessing Incorrect Information

Initially, we made the assumption that all information possessed by the agents, both about the environment as well as about the other agent's database, is correct. This assumption, while restrictive, might be appropriate for agents whose domain of action is limited. If there is no chance (or extremely little chance) that incorrect information might be introduced into the agents databases, the analysis above is sufficient to categorize available strategies.

However, it is obvious that agents operating in a more complex environment could be expected to hold incorrect beliefs about the world and about each other. Information passing strategies that they employ must take into account the potential inconsistencies between their databases. In this section we consider the implications that incorrect information has on the analysis presented above.

In a general sense, what concerns us is discrepancies between agents' databases, and how to resolve them. The issue of whose information is correct, relative to the outside world, is more difficult. If the discrepancies can be resolved by combining facts from both databases, our agents should adopt appropriate resolution strategies. In some cases, however, there will be no clear way to resolve disagreements over the basic facts; we will touch on this issue only briefly, noting that much work remains to be done on this subject.

6.1 The Sources of Incorrect Information

There are several potential sources for an agent's incorrect information. First, it is possible that the agent was initially configured with incorrect axioms or incorrect rules. Second, an agent may rely on faulty sensor information, and believe some fact to be true of the environment when it is not. Third, a conclusion may have been derived from a non-

monotonic rule, when further information would have shown that the rule was inapplicable under the circumstances.

The first cause listed above, that of an incorrect initial configuration, is difficult to remedy. Particularly if an agent holds a consistent, though spurious, world-view, it may be impossible to correct his assumptions (in fact, it is not at all clear why the corrector can be assumed to have a more accurate set of beliefs, relative to the environment). Arbitration between two distinct, consistent world-views is not our concern. If, on the other hand, an agent's initial incorrect configuration contains internal inconsistencies, or has the potential for containing inconsistencies (e.g., through the introduction of new sensor data), such a problem might be discovered by the agent itself or through the communication of another agent.

The other two possibilities, where faulty sensors or non-monotonic assumptions have introduced errors into a database, are of more interest to us. In the sections below, we examine the immediate consequences that these errors have on information transfer strategies, as well as approaches to resolving the inconsistencies between databases.

6.2 Changed Assumptions, Changed Consequences

In dropping the "correct information" requirement, we are in effect dropping Assumptions 1 and 2, presented above. Assumption 1 stated that both agents' information about the world was correct, though possibly incomplete; Assumption 2 stated that both agents' information about the other agent's database was correct, though possibly incomplete.

The consequence of Assumption 1 (and monotonicity) had been that all plans that an agent proved were guaranteed correct. The consequence of Assumption 2 (and monotonicity) had been that all plans that an agent suspected another agent of having, were, in fact, possessed by the second agent.

The removal of the monotonicity assumption had a deleterious effect on both consequences above, as well as the information passing strategies that these consequences engendered. We have a similar situation if we withdraw Assumptions 1 and 2, even if

monotonicity is maintained. In particular, the plan passing strategy and information transfer strategies presented in the monotonic section above are invalid in the incorrect information case, as they were in the nonmonotonic case.

Lemma 1. *In the incorrect information case, if each agent picks a plan that is provable from its own database and sends it to the other, the agents may not converge to a plan in* $\Pi(D_T)$.

Proof: Same as Theorem 3. ■

Lemma 2. *In the incorrect information case, there is no guaranteed strategy of information transfer for arbitrary* D_A *and* D_B *that will cause* A *and* B *to converge to a plan in* $\Pi(D_T)$, *even if* $\Pi(D_A)$ *and* $\Pi(D_B)$ *are non-empty.*

Proof: Same as Theorem 4. ■

In the nonmonotonic case, we proceeded to define a new type of convergence ("acceptable convergence") and showed that total information transfer allowed it to occur. Here, we have a more vexing problem, since total information transfer (or even total relevant information transfer) may result in an inconsistency in the agents' databases. Clearly, facts cannot be accepted at face value, and there must be an approach to resolving facts that are inconsistent with one another. The following section will examine this issue.

6.3 An Approach to Resolving Information Disagreement

The discovery by an agent that another agent's database disagrees with its own can occur in one of several ways. Most fortuitously, one agent might be pursuing one of the strategies presented above (information transfer), and the second agent simply receives a fact inconsistent with its own database. This may involve a direct contradiction, or it may require a more lengthy deduction to discover (and thus may not be apparent at first).

Another possibility is that agent A has sent B a set of facts which it believes implies a plan when added to B's database. In fact, since A has incorrect information about B, no plan is implied (or a different plan is implied). The discovery may then occur when the plan fails for lack of coordination, or, if the agents are careful, when they compare

prospective actions. In the presence of potentially incorrect information, such plan cross-checking should in general occur.

Let us assume, then, that the discrepancy has been discovered, and that the agents are now interested in proceeding. As an example, let us assume that agent A believes that fact x is true, and that agent B believes that x is false (i.e., $\neg x$ is true). Of course, there would be no problem if B did not believe x or $\neg x$, but our assumption is that there is an active disagreement between the two agents. We are interested in developing an algorithm for A and B to use in resolving their disagreement about fact x.

Actually, the facts x and $\neg x$ may be believed by the agents due to a chain of deductions. Each agent sends "justifications" (in the truth maintenance sense [19]) to the other agent. For example, let us suppose that t and u imply x, and that A believes t and u (and therefore x). Similarly, e and f imply $\neg x$, and B believes e and f (and therefore $\neg x$). Clearly, if the implications being used are correct and/or accepted by both agents, B cannot believe t and u, and A cannot believe e and f. Eventually, a core group of facts is discovered on which the two agents disagree, and which are not themselves based on any other database assertions. The initial problem is thus for the two agents to identify this core group of facts.

In the nonmonotonic case, the identification of the core group proceeds in the same way, though the process may be simplified by the invalidating of certain rules that had been used. Either one of the agents (or both!) will discover that, in the light of their new information, their original conclusions were unwarranted, or they will identify the core group of facts upon which they disagree.

It should be noted that the process of discovering the core group can be simplified by communicating each step of the search. For example, perhaps B believes u but does not believe t. In this case, agent A can restrict its search to the t branch of the deduction; similar pruning can occur at each stage of the tree.

Once the core group has been identified, the agents need to compare the support they have for each conflicting fact. Beyond having arrived at a fact by deduction (which is

irrelevant for the core group of facts), there are three kinds of support for a fact: 1) the fact may be "hard-wired," in the sense that it was built into the original configuration of the system, 2) the fact may be sensor-based, or 3) the fact may have been transmitted by another agent and believed by the receiver.

The resolution of conflicting core facts can be achieved in a variety of ways. There are a variety of heuristics to consider, for example, an ordering of the support categories (e.g., hard-wired facts take precedence over sensor-based facts, sensor-based facts take precedence over transmitted facts). Belief in the ultimate reliability of hard-wired facts may be particularly appropriate if these facts are kept to a reasonable minimum. Transmitted facts can be converted into one of the other two types of facts if the original transmitter can be contacted and traces the fact's support.

The problem of resolving disagreements at the same level may also be treated heuristically, with, for example, certain sensor-based readings considered most reliable (e.g., a satellite robot can be considered to know its own rotation speed reliably). A more difficult problem would be for an agent to deduce a method of convincing the other agent that, for example, a particular sensor has stopped working. Such a strategy might involve creating a situation whereby belief in the faulty sensor's measurements would introduce an inconsistency in the other's database. Although such a deduction is certainly possible, the general problem of designing such a test is very difficult. We look to simpler measures (such as the heuristics mentioned above) to resolve the vast majority of disagreements among agents.

§7. The Limited Utility of Lying

In all of the discussion above, we have only analyzed the transfer of consistent information, i.e., facts that are consistent with the database of the sender. What utility might there be to having agents send each other locally *inconsistent* information, that is, axioms that are inconsistent with the sender's database? Although in some sense the sender believes these inconsistent facts to be false, he may still be able to further global utility by making creative use of them.

Consider, for example, the case where the global utility function specifies that communication between agents A and B should be kept to a minimum. A knows that by transferring a (locally consistent) set of axioms i to B, B will accept some plan in $\Pi(D_A)$; however, there is a much shorter set of (locally inconsistent) axioms f that will also cause B to accept the same plan. Moreover, it can be shown that B's possession of the "false" facts f will have no other effects in the future (e.g., execution of the plan removes those facts from B's database). Under these circumstances, global utility might warrant the passing of locally inconsistent information.

The usefulness of passing incorrect information has been considered for several reasons. For example, the concept of "loosening" honesty requirements so as to achieve common knowledge appears in [20]. The authors show that the practice is actually required under certain circumstances, i.e, when simultaneity among the agents cannot be achieved. In another example, [21] considers the practice of a system "telling white lies" so as to simplify the introduction of a new concept to a user (the way a teacher often does).

In fact, current AI systems have no explicit mechanisms to keep them from communicating inconsistent information; if they can deduce that such communication will help reach their goal state, they will use it. At issue is the desirability of such behavior [22, 23].

7.1 Economic and Specificity Arguments

There are several economic rationales for inconsistent information transfer (IIT) in the presence of a global utility function (the local utility of this practice when local utility functions are involved can be trivially demonstrated, as discussed below). The economic benefits of IIT can take various forms, particularly savings in communication costs and in the costs of deduction for the sender and the receiver of the information.

The actual circumstances when these benefits accrue without potentially catastrophic side-effects are not easily discovered. Particularly when there is uncertainty about future world states, IIT can be a risky business. These unforeseen states might cause the other agent to use the inconsistent information in unexpected ways, lessening global utility.

Consider also the case where the passing of inconsistent information is later discovered by the other agent; there may be a subsequent refusal to accept information, leading to a loss of global utility (though we might choose to design the agents not to react this way).

In addition to economic arguments that can be made for IIT, there is also a "specificity" argument: A may get B to accept some *set* of plans Δ using IIT that B would not specifically accept if it were only passed consistent information (i.e., consistent information might cause it to accept a superset or a subset of Δ). However, there are no plans that B could be made to accept only by IIT. More precisely, we prove the following:

Theorem 6. *No reasonable results (that are discoverable by the sender) can be derived from the passing of locally inconsistent information that cannot be derived from the passing of locally consistent information.*

Proof: A "reasonable" result in the monotonic case is simply making the other agent accept a plan in $\Pi(D_T)$. The only such plans known by the sender are those implied by its own database (all of which are in $\Pi(D_T)$ by Consequence 1). But the whole local database could be sent (which is, of course, locally consistent), causing the receiver to accept the entire set of plans that it implies. Thus, there are no correct plans (of which the sender can be aware) that the receiver could only be convinced of through locally inconsistent information.

In the nonmonotonic case, there is no way for the sender to discover which plans are correct, since it only knows about the plans implied by its own database (which need not be in $\Pi(D_T)$). Thus, we use a different definition of a "reasonable" result. Following our "single-agent isomorphism" argument above, we say that a reasonable plan is one in the set $\Pi(D_{A\cup B})$. Our proof then follows as above. The sender's entire database could be transferred to the receiver; having $D_{A\cup B}$ as its database, it will now accept all plans in $\Pi(D_{A\cup B})$. Thus, there is no $D_{A\cup B}$ consistent plan that the receiver could only be convinced of through locally inconsistent information. ■

§8. Future Research

Strategies in the presence of sophisticated inter-agent knowledge must be examined. In the above work, we have generally restricted ourselves to inter-agent knowledge of the ground clause variety; existentially instantiated knowledge needs to be allowed (agent A knows that B knows some fact, without A itself knowing what the fact is), along with "knowledge about knowledge" (agent A knowing that B knows that A knows something),

disjunctive knowledge (A knowing that B knows fact p or fact q) and "common knowledge" [20].

Future work will examine some consequences of agents having non-identical goals. For example, agent B might ask A for the location of the blue block nearest B (all blue blocks being out of B's sensing range, or out of its line of sight). A might then give the location of another blue block, not the one closest to B but one which A itself desires to see moved. B's desire may be satisfied suboptimally, while A may further its local utility through IIT.

The aim of this investigation is to develop a framework for understanding cooperative behavior. With such a framework in hand, it will be possible to design intelligent agents with an appropriate cooperative bent; machines, working together in practical domains, will flexibly aid one another as they carry out their tasks. Such cooperation is ultimately beneficial since the agents, helping each other, can accomplish their jobs more efficiently (perhaps even accomplish tasks that are impossible without cooperation). The move to distribution of computers has begun; with a better understanding of cooperation, the full power of autonomous agents can be exploited.

References

[1] Davis, R., and Smith, R. G., Negotiation as a metaphor for distributed problem solving, *Artificial Intelligence* **20** (1) (1983) 63–109.

[2] Davis, R., A model for planning in a multi-agent environment: steps toward principles for teamwork, Working Paper 217, MIT AI Lab (1981).

[3] Georgeff, M., Communication and interaction in multi-agent planning, *AAAI-83*, Washington, D.C. (1983) 125–129.

[4] Georgeff, M., A Theory of action for multi-agent planning, *AAAI-84*, Austin, Texas (1984) 121–125.

[5] Corkill, D.D., and Lesser, V.R., The use of meta-level control for coordination in a distributed problem solving network, *IJCAI-83*, Karlsruhe, West Germany (1983) 748–756.

[6] Appelt, D.E., Planning natural language utterances to satisfy multiple goals, Tech Note 259, SRI International, Menlo Park, California (1982).

[7] Moore, R.C., Reasoning about knowledge and action, Tech Note 191, SRI International, Menlo Park, California (1980).

[8] Moore, R.C., A formal theory of knowledge and action, Tech Note 320, SRI International, Menlo Park, California (1984). Also to appear in *Formal Theories of the Commonsense World*, Hobbs, J.R., and Moore, R.C. (Eds.), Ablex Publishing Co. (1984).

[9] Konolige, K., A first-order formalization of knowledge and action for a multi-agent planning system, Tech Note 232, SRI International, Menlo Park, California (1980).

[10] Konolige, K., Circumscriptive ignorance, *AAAI-82*, Pittsburgh, Pennsylvania (1982) 202–204.

[11] Konolige, K., A deductive model of belief, *IJCAI-83*, Karlsruhe, West Germany (1983) 377–381.

[12] Konolige, K., *A Deduction Model of Belief*, Ph.D. Thesis, Stanford University (1984).

[13] Schelling, T.C., *The Strategy of Conflict*, Oxford University Press, New York (1963).

[14] Finger, J. J., and Genesereth, M. R., Residue—A deductive approach to design, Memo HPP-83-46, Stanford Heuristic Programming Project (1983).

[15] Rosenschein, S. J., Plan Synthesis: A logical perspective, *IJCAI-81*, Vancouver, B. C., Canada (1981) 331–337.

[16] Machtey, M., and Young, P., *An Introduction to the General Theory of Algorithms*, North Holland, New York (1978).

[17] Reiter, R., A logic for default reasoning, *Artificial Intelligence* **13**(1, 2) (1980) 81–132.

[18] McDermott, D., and Doyle, J., Non-monotonic logic I, *Artificial Intelligence* **13**(1, 2) (1980) 41–72.

[19] Doyle, J., A truth maintenance system, *Artificial Intelligence* **12** (1979) 231–272.

[20] Halpern, J., and Moses, Y., Knowledge and common knowledge in a distributed environment, *Proceedings of the Third Annual ACM Conference on Principles of Distributed Computing*, Vancouver, British Columbia, Canada (1984).

[21] Swartout, B. R., XPLAIN: A system for creating and explaining expert consulting programs, Information Sciences Institute, ISI Reprint Series, ISI/RS-83-4 (1983).

[22] Bok, Sissela, *Lying: Moral Choice in Public and Private Life*, Vintage Books, New York (1978).

[23] Axelrod, Robert, *The Evolution of Cooperation*, Basic Books, Inc., New York (1984).