Testing Collaborative Strategies by Computational Simulation: Cognitive and Task Effects

To Appear in Knowledge Based Systems, March, 1995

Marilyn A. Walker
Mitsubishi Electric Research Labs*
201 Broadway, Cambridge, Ma. 02139
walker@merl.com

Abstract

A theory of communication between autonomous agents should make testable predictions about which communicative behaviors are collaborative, and provide a framework for determining the features of a communicative situation that affect whether a behavior is collaborative. The results presented here are derived from a two-phase empirical method. First, we analyze a corpus of naturally-occurring problem-solving dialogues in order to identify potentially collaborative communicative strategies. Second, we experimentally test hypotheses that arise from the corpus analysis in Design-World, an experimental environment for simulating dialogues. The results indicate that collaborative behaviors must be defined relative to the cognitive limitations of the agents and the cognitive demands of the task. The method of computational simulation provides an additional empirical basis for theories of human computer collaboration.

1 Introduction

A theory of communication between autonomous agents should make testable predictions about which communicative behaviors are collaborative. Thus this paper has three inter-related goals: (1) to present and argue for an empirical method of testing a theory of collaborative communication; (2) to use this method to show that agents' cognitive limits and the cognitive demands of the task determine whether the communicative strategies tested are collaborative, and thus (3) to support the claim that

*This research was partially funded by ARO grant DAAL03-89-C0031PRI and DARPA grant N00014-90-J-1863 at the University of Pennsylvania and by Hewlett Packard, U.K.

cognitive capabilities and processes must be modeled as part of designing systems for human-computer collaboration.

A theory of collaborative communication must first of all have a way of identifying and representing communicative strategies. Strategies for agents to communicate with humans are best identified from the strategies used by human conversants in joint problem solving dialogues. The strategies to be described below are based on an analysis of two corpora: (1) 55 problem solving dialogues, constituting 5 hours of a radio talk show for financial advice [24, 37, 33]; and (2) 16 dialogues collected to explore the role of a shared visual environment when two agents collaborate on the design of the floor plan of a two room house [38, 39]. We will refer to the first corpus as the financial advice corpus and the second corpus as the design corpus.

Second, we must be able to generate hypotheses about whether and why a strategy is collaborative. The fact that a strategy occurs in human-human dialogue does not guarantee that it is generally collaborative. For example, consider utterance (17) in Dialogue excerpt 1 from the financial advice corpus, in which H makes an inference explicit.

(1) (15) H: Oh no. *I R A's were available as long as you are not a participant in an existing pension*
(16) J: Oh I see. Well I did work, *I do work for a company that has a pension*
(17) H: ahh. THEN YOU'RE NOT ELIGIBLE FOR EIGHTY ONE

In Dialogue 1, utterance (15) asserts a biconditional inference rule, (16) instantiates one of the premises of this rule, and (17) realizes an inference that follows from (15) and (16), for the particular tax year of 1981, by the inference rule of MODUS TOLLENS. There is no evidence that J would not have made this inference if H hadn't made it explicit. H's assertion violates a commonly assumed conversational rule of not asserting facts that are already mutually believed. H's conversational strategy may also be inefficient because it increases the number of utterances in the dialogue [11, 20, 9]. On the other hand H's strategy of making inferences explicit might well be collaborative, if, for example, H could say 17 and J could process it with less effort or in less time than if H left J to make the inference realized by 17 on his own. Clark's model of conversation suggests such a view of collaboration with the notion of conversants achieving LEAST COLLABORATIVE EFFORT [7, 4].

The difficulty with identifying a collaborative strategy is due to the fact that no-one has provided a definition of COLLABORATIVE or of COLLABORATIVE EFFORT. Collaborative problem solving involves at least three processes: (1) retrieval processes necessary to access previously stored beliefs in memory; (2) communicative processes related to generating and interpreting utterances; and (3) inferential processes that operate on beliefs stored in memory and those communicated by other agents. Collaborative effort must include the costs for both agents for all of these processes. This suggests the definition of collaborative effort below:

COLLABORATIVE EFFORT =
(the total cost of communication for both agents)
+ (the total cost of inferences for both agents)
+ (the total cost of retrievals for both agents)

Then a COLLABORATIVE communicative strategy is one that improves the combined PERFORMANCE of **both** agents in a dialogue, where performance is the difference between a measure of the quality of the problem solution and COLLABORATIVE EFFORT.

With these definitions of COLLABORATIVE and COLLABORATIVE EFFORT, we are closer to testing the hypothesis that utterance 17 in the naturally-occurring dialogue above is collaborative even though it increases the cost of communication because it reduces the cost of inference. Below we will refine our definitions of performance and collaborative effort so that we can test this hypothesis.

This brings us to a third requirement for a theory of collaborative communication: a way of testing hypotheses about collaborative strategies. In naturally-occurring dialogues, we often cannot tell why a strategy was used in a particular situation, and whether it is generally collaborative. We need to compare strategies according to the collaborative effort involved in different situations. If a strategy is collaborative, we must be able to specify whether this is due to properties of the conversants such as their inferential capability, or due to the difficulty of the task, to what has been said previously in the conversation, or to the domain.

I will argue for a method of developing a well specified theory of collaborative communication that consists of corpus analysis coupled with theory abstraction and computational simulation. This method is illustrated by the results of experiments in Design-World, a computational dialogue simulation environment. In Design-World, we vary agents' communicative behaviors and other features of the communicative situation and then measure the effect of these variations on how well the agents perform a simple task. The remainder of the paper presents the method and empirical results and discusses the implications of the results and the method for a theory of collaborative communication.

2 Corpus Analysis

The first step of the method is the identification of potentially collaborative communicative strategies. These were developed from an analysis of naturally-occurring problem-solving dialogues from the financial advice corpus and the design corpus.

A critical observation is that the agents in these problem solving dialogues do not always behave **optimally** if we consider the least number of utterances criterion mentioned above. One way in which apparently non-optimal behavior is exhibited is in the production of INFORMATIONALLY REDUNDANT UTTERANCES, IRUs, in which agents tell other agents facts that are mutually believed or that are inferrable from mutual beliefs. Because IRUs are so prevalent and apparently non-optimal, the corpus analysis focused on strategies that include IRUs.[1] For example, in Dialogue 1, utterance (17) is an IRU. Similarly, in (26a) in Dialogue excerpt 2, H repeats the causal inference that one course of action will lead to a gain of 400 dollars, before considering potential costs:

(2) (20) H: Right. The maximum amount of credit that you will be able to get will be 400 *that they will be able to get will be 400 dollars on their tax return*
(21) C: *400 dollars for the whole year*?
(22) H: *Yeah it'll be 20%*
(23) C: *um hm*
(24) H: Now if indeed they pay the $2000 to your wife, that's great.
(25) C: um hm
(26a) H: SO WE HAVE 400 DOLLARS.
(26b) Now as far as you are concerned, that could cost you more. Remember, you're gonna have two thousand dollars worth of income. What's your tax bracket?
(25) C: Well, I'm on pension Harry, and uh and my wife hasn't worked at all and .....

The information in 26a was already known to C based on the information previously asserted and accepted in utterance 20 . . . 23. Thus 26a is an IRU.

The same strategy of making inferences explicit occurs in the design corpus. In utterance 107 in Dialogue 3, S makes an inference explicit that follows from the description of the sofa given in 100 . . . 106:

(3) (100) R: .....And now my final shape is a sofa.
(101) S: Right.
(102) R: Which is um rectangular. It's 12 x 8 -
(103) S: uh huh -
(104) R: but there's a chunk missing out of it, er another rectangular of 9 x 4.
(105) S: Sounds like mine.
(106) R: It's not symmetrical.
(107) S: Right. THERE'S A 1 AND A 2.
(108) R: Yes.
(109) S: Ok......

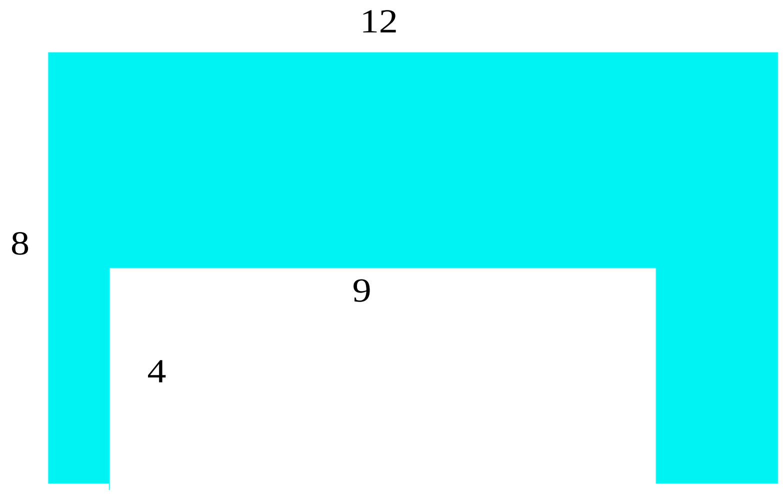

Figure 1: Non Symmetrical Sofa used in Design Corpus

The engineers who served as subjects in the design dialogues presumably are capable of calculating that a 12 x 8 rectangle, missing a 9 x 4 rectangular chunk, leaves a nonsymmetrical shape, with one piece that measures 1 foot and another that measures 2, as shown in figure 1.[2] However S makes this inference explicit.

All of these dialogues generalize to any situation in which propositions P and Q have been discussed in a dialogue by agents A and B, and both A and B know the inference rule of $(P \wedge Q \rightarrow R)$. When should one of the agents make the inference of R explicit?

The corpus analysis supports four observations: (1) The IRU that realizes R is a CONCLUSION that follows as an inference from a set of PREMISES. Thus, the informational relation between the premises realized by other utterances in the dialogue and the IRU is PREMISE:CONCLUSION [21]; (2) It is plausible that the other conversant had already inferred or could have deduced R; (3) A speaker who makes R explicit makes a communicative **choice** to do so.

So why does an agent A choose to realize R explicitly in a conversation with an agent B? Plausible hypotheses suggested by the corpus and research on human memory and inferencing are:

1. HYPOTH 1: Even when both agents **know** the premises for an inference, these premises must be accessible in working memory to be used in making an inference [26, 19]. Thus agent A might make an inference explicit because either the major or some of the minor premises are not accessible.

2. HYPOTH 2: Even when agent B can in principle retrieve the premises for an inference into working memory, retrieval would take longer and/or require more

[1] The corpus analysis classifies IRUs into three types: Attitude, Attention and Consequence. We will only be concerned with Consequence IRUs here.

[2] All the piece measurements were in full feet.

cognitive effort than the effort involved in processing the utterance that makes an inference explicit.

3. HYPOTH 3: Agent A makes the inference R explicit because R should be accessible to B as a premise for another inference.
4. HYPOTH 4: Agent A makes the inference R explicit because it is important for agent B to **remember** R. This hypothesis relates IRUs to human cognitive limitations and the well-known effects of recency and frequency, since repetition of any known fact improves an agent's capability to remember that fact [2].
5. HYPOTH 5: Agent A makes an inference explicit to **demonstrate** to agent B that agent A made the inference [7, 15].
6. HYPOTH 6: Agent A makes an inference explicit to demonstrate to agent B that agent A **believes** the inference rather than another conflicting inference (conflicting default inferences) [34].

Unfortunately the corpus analysis of naturally occurring data cannot test these hypotheses because each conversation occurs only once and thus conversations in which inferences are made explicit cannot be compared with the same conversation without those utterances. In addition, many of the hypotheses are related to human processing capabilities, which the corpus provides no access to. The number of utterances to complete a dialogue cannot be used as a measure of collaborative effort because this number bears no relation to the quality of the problem solution and only takes into account some of the processes involved.

What we need is a way to compare many conversations where the only thing that varies is whether an agent makes an inference explicit at a certain point in the conversation. We also require a way to measure the cost of communication, inference and retrieval for two agents involved in collaborative problem solving, to vary their communication strategy, and to measure how this variation affects performance on the task and collaborative effort. These factors motivate the development of Design-World, a computational dialogue simulation environment.

3 Design-World Domain and Task

The Design-World task consists of two agents who carry out a dialogue in order to come to an agreement on a furniture layout design for a two room house, similar to the task in the design corpus [39]. The agents' shared intention is to design the house, which requires two subparts of designing room-1 (the study) and designing room-2 (the living room). The agents' performance is evaluated based on how good the final design is, according to criteria to be discussed below. Figure 2 shows a potential final plan constructed as a result of a dialogue.

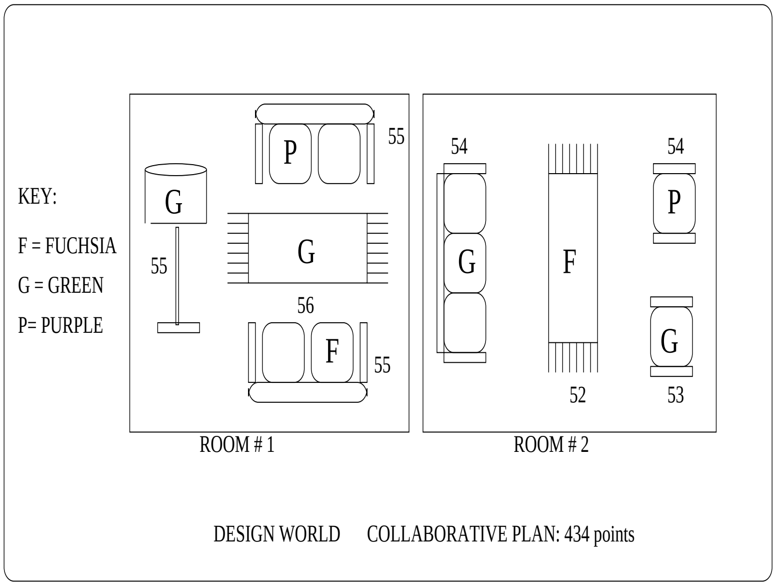

Figure 2: Potential Final State for Design-World Task: A Collaborative Plan Achieved by the Dialogue

At the beginning of a dialogue simulation, both agents know that a plan for designing room-1 or room-2 consists of four intentions to PUT a furniture item into the room. Agents start with private beliefs about the furniture items they have, and they know that any furniture item they have can be used in a potential PUT-ACT. Each furniture item has a color and a point value. Agents' belief states at the beginning of the dialogue with respect to what is mutually believed is similar to a card game. Both agents know which furniture items exist and how many points they are worth, but don't know which items the other agent has.

The colors and point values of the furniture items are provided in order to provide a basis for choosing among them when reasoning about the task. The colors of furniture items support task variations that involve more complex inferencing [31], and the point values give each PUT-ACT using a particular furniture item a utility that the agents use in deciding which items to use, as well as providing a way of objectively measuring the agents' performance on the task.

Agents' performance varies according to a set of cognitive and task parameters to be discussed below. One component of performance is the RAW SCORE that the agents get for their final Design-House plan. In the experiments below we compare communicative strategies for two versions of the task where RAW SCORE is calculated differently. RAW SCORE for the Standard task is simply the sum of the scores of all the furniture items for each valid step in the plan. The point values for invalid steps in the plan are simply subtracted from the score.

In general, the effect of invalid steps in a plan depends on how interdependent different

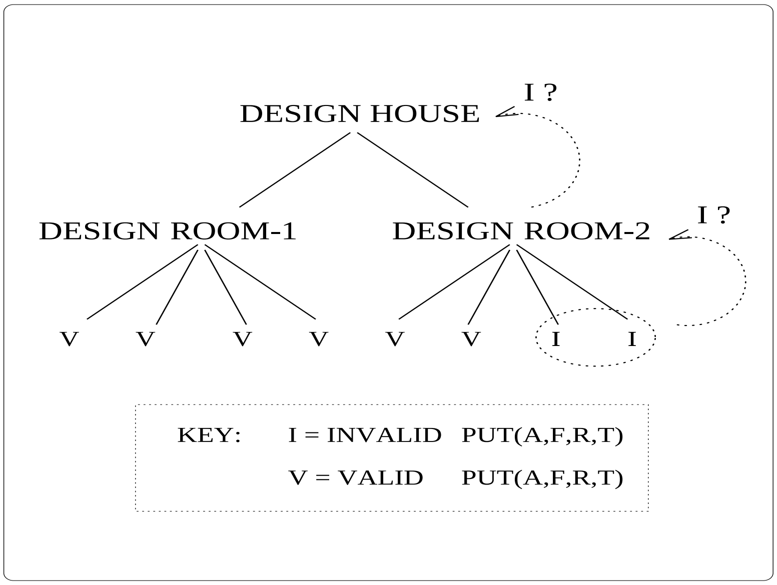

Figure 3: Evaluating Task Invalids: for some tasks invalid steps invalidate the whole plan.

subparts of the problem solution are.[3] Figure 3 shows the choices for the effect of invalid steps for the Design-World task. The score for invalid steps (mistakes) can just be subtracted out; this is how the Standard task is defined. Alternately, invalid steps can propagate up so that an invalid PUT-ACT means that the Design-Room plan is invalid. Finally, mistakes can completely propagate so that the Design-House plan is invalid if one step is invalid.

This final option defines the Zero-Invalids task as a fault-intolerant version of the task. In the Standard task, agents are not heavily penalized for making mistakes, but the Zero-Invalids Task is more cognitively demanding because any invalid intention invalidates the whole plan.

Because the Design-World task is simple, it is a good basis for developing a theory of collaborative communication. The components of the task are isomorphic with components of other collaborative tasks. The task is potentially executable by pairs of robots or by human robot teams, and can be set up to have a visual component. When both agents have access to information about the task, it gives rise to mixed-initiative dialogues [37].

[3] Contrast aircraft scheduling with furnishing a room.

4 Design-World Communicative Strategies

The primary purpose of Design-World is to allow us to compare a communicative strategies under different assumptions about agents' cognitive processes and the demands of the task. Here we discuss communicative strategies before turning to the discussion of agents' underlying cognitive processes.

Agents in Design-World communicate via 7 communicative acts: OPEN, CLOSE, PROPOSE, ACCEPT, REJECT, ASK and SAY. These acts are realized via the schemas below:

(Propose Speaker Hearer Option)
(Ask Speaker Hearer Belief)
(Say Speaker Hearer Belief)
(Accept Speaker Hearer Option)
(Reject Speaker Hearer Belief)
(Reject Speaker Hearer Option)
(Open Speaker Hearer Option)
(Close Speaker Hearer Intention)

The OPTIONS and INTENTIONS shown as the content of the communicative acts above are to PUT an item in a room or to start working on one of the rooms. An option only becomes an INTENTION if it is ACCEPTED by both agents, either explicitly or implicitly. Beliefs are either those that an agent starts with, beliefs communicated by the other agent, or inferences made during the conversation. Beliefs in ASK actions have variables that the addressee attempts to instantiate. The option in a REJECT schema is a counter-proposal and what is rejected is the current proposal. The belief in a rejection schema is a belief that the speaker believes is a reason to reject the proposal, such as a belief that the preconditions for the option in the proposal do not hold [37].

Excerpt 4 is part of a simulated dialogue: each utterance is shown both as a gloss in *italics*, and in the artificial language that the agents communicate with, as defined above.[4]

(4) 1: BILL: *First, let's put the green rug in the study.*
(propose agent-bill agent-kim option-43: put-act (agent-bill green rug room-1))

2: KIM: *Then, why don't we put the green lamp in the study.*
(propose agent-kim agent-bill option-61: put-act (agent-kim green lamp room-1))

3: BILL: *No, it's better to put the green couch in the study.*
(reject agent-bill agent-kim option-75: put-act (agent-bill green couch room-1))

[4] The generation of the glosses was not a focus of the work here, thus the generation procedures are adhoc.

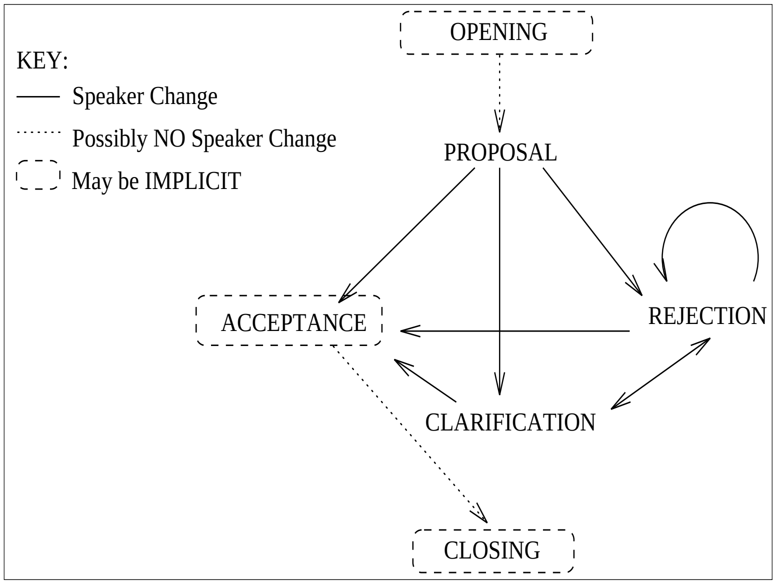

Figure 4: Discourse Actions for the Design-World Task

4: KIM: *Putting in the green-lamp is worth 55.*
(say kim bill bel-34: score (option-61: put-act (kim green-lamp room-1) 55))

5: *It's better to put the green-lamp in the study.*
(reject kim bill option-61: put-act (kim green-lamp room-1))

In dialogue 4, Bill makes a proposal in 1, and then Kim implicitly accepts this proposal with a new proposal in 2. In 4-3 Bill rejects Kim's proposal. Bill does this because option-75 is another option that Bill has identified which he believes is better than Kim's proposed option-61. Bill makes his reason for rejection explicit by realizing this option as the content of his rejection. Kim in 4-4 rejects Bill's counter-proposal and explicitly gives her reason for doing so, namely that the green lamp is worth 55, a fact which Bill has forgotten. The two agents eventually agree on an option to satisfy the current intention, after a sequence of proposals and rejections about various options, because they share a global utility function for actions in the domain as defined by the scores associated with each furniture item. The way that agents communicate rejection is patterned on observations about how rejection is communicated in human-human dialogue [40], which form the basis of a set of COLLABORATIVE PLANNING PRINCIPLES [37, 33].

Communicative acts are aggregated into discourse acts of OPENINGS, CLOSINGS, PROPOSALS, CLARIFICATIONS and ACCEPTANCES. Agents participate in a dialogue by following a schema of discourse acts, as shown in figure 4.[5] Each domain-plan step has

[5] Also see [5, 28, 30]. Schemas like this are probably not adequate to describe all discourse action transitions in every type of dialogue [18].

a corresponding proposal, and a plan step may have a corresponding explicit opening and closing. Each proposal can be followed by an acceptance, a clarification, or a rejection.

Discourse acts expand to zero or more communicative acts. The plan expansion that the agent uses for the discourse acts in figure 4 depend on agents' communicative strategies.

An expansion to zero communicative acts means that the discourse act was left implicit, as shown in figure 4. For example, if agents don't accept a proposal explicitly, acceptance can be inferred as a default. Similarly, if agents don't close a segment explicitly, the occurrence of an utterance related to a subsequent intention can support the inference of closure.

Agents may also realize a discourse act with one or more communicative acts. For example, a closing statement might consist of simply a CLOSE statement, or a CLOSE act preceded or followed by a SAY act, where the content of the SAY act is an inference or a summary of the result of the intentions that was just agreed to.

In the experiments presented below we compare two strategies: (1) All-Implicit; and (2) Close-Consequence. The All-Implicit strategy is a 'bare bones' strategy, exemplified by the dialogue in 4. Agents parameterized with the All-Implicit strategy do not include IRUs in any discourse act or produce discourse acts whose occurrence could be inferred. They do not make inferences explicit or remind other agents of facts that might be useful in inferencing, as we saw in dialogues 1 and 2.

Agents parameterized with the Close-Consequence strategy include IRUs at dialogue segment closings. In dialogue 5 agent CLC uses the Close-Consequence strategy (CLC is a mnemonic for CLose Consequence). CLC makes explicit Closing statements, such as 5-2, on the completion of the intention associated with a discourse segment. CLC also includes IRUs as in 5-3; CLC makes the inference explicit that since they have agreed on putting the green rug in the study, Bill no longer has the green rug (act-effect inference).

(5) 1: BILL: *First, why don't we put the green rug in the study.*
(propose agent-bill agent-clc option-30: put-act (agent-bill green rug room-1))

2: CLC: *So, we've agreed to put the green rug in the study.*
(close agent-clc agent-bill intended-30: put-act (agent-bill green rug room-1))

3: CLC: AGENT-BILL DOESN'T HAVE GREEN RUG.
(say agent-clc agent-bill bel-48: has n't (agent-bill green rug))

4: BILL: *Then, why don't we put the red-rug in the study.*
(propose bill clc option-50: put-act (bill red-rug room-1))

In dialogues 1 and 2, the informational relation between utterances asserted and the IRU was PREMISE:CONCLUSION. In Design-World, one possible realization of this informational relation is in act effect inferences, and that is the basis for the inferences that the Close-Consequence agents make explicit. Thus the Close-Consequence strategy parallels the naturally occurring examples in 1 and 2.

So far I have described the Design-World task and the communicative acts that are available to the agents when participating in a dialogue. In order for the agents to actually do the task, the agents' communicative behaviors must be linked with their underlying cognitive processes, which gives them a reason to communicate.

5 Cognitive Architecture of Design World Agents

5.1 IRMA Agent Architecture

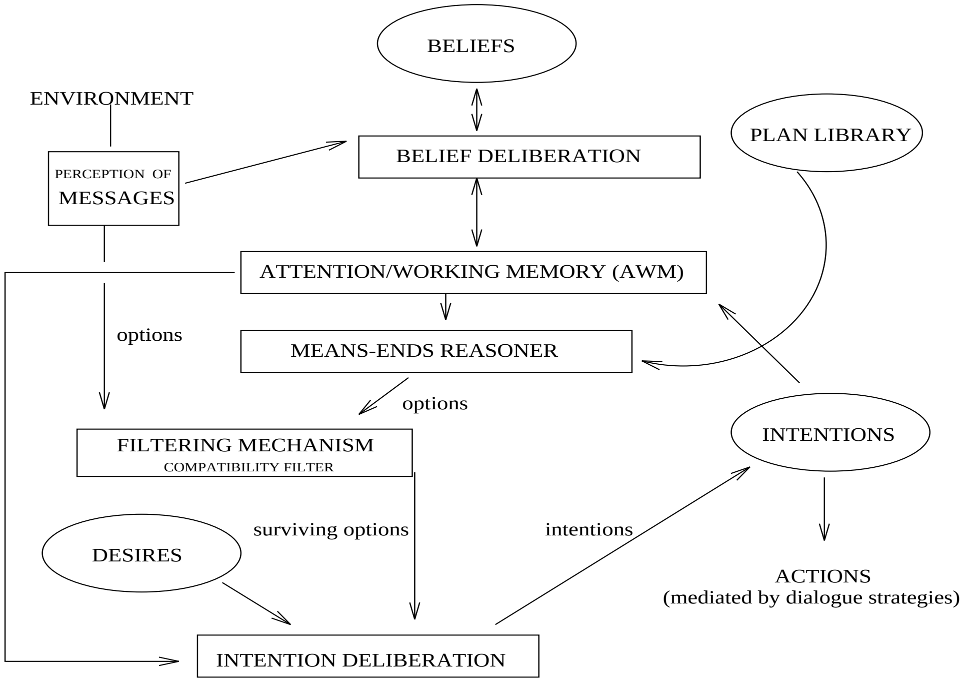

Figure 5: Design-World version of the IRMA Agent Architecture for Resource-Bounded Agents with Limited Attention (AWM)

The agent architecture is based on the IRMA architecture, also used in the TileWorld simulation environment [3, 25, 14], with the addition of a cognitive model of Attention/Working Memory (AWM) to be discussed below. See figure 5. As the figure shows, the output of each cycle of reasoning about a plan step is a set of communicative acts, mediated by the communicative strategies discussed above.

The content of each communicative act is a belief or a (potential) intention, e.g. the content of a PROPOSE utterance is an OPTION generated via means-end reasoning as shown in figure 5. Means-end reasoning is based on the agents' initial beliefs about which furniture items they have. As options are generated, agents deliberate and then select the best one as the basis of a proposal to another agent. As proposals are accepted, the agents make and store inferences that they no longer have the furniture items that they used in the design plan.

The simulated dialogues in 4 and 5 illustrate part of the cycle for achieving a DESIGN-HOUSE plan: (1) individual agents MEANS-END REASON about options in the domain; (2) individual agents DELIBERATE about which options are preferable; (3) then agents make PROPOSALS to other agents, based on the options identified in a reasoning cycle, about actions that contribute to the satisfaction of their intentions; (4) then these proposals are accepted or rejected by the other agent, or acceptance/rejection is postponed by a clarification.

Deliberating whether to accept or reject a proposal is based on beliefs about the proposed action's utility, the desire to maximize utility is the only desire that Design-World agents have [3, 8]. Utility theory gives a formal basis to an account of collaboration that emphasizes agents' autonomy, since agents use it as the basis for deciding what to accept or reject.[6] The use of utility as the basis for agents' decisions, and its use in formulating agents' autonomy with respect to the acceptance of each proposed intention is reflected in the definition of a COLLABORATIVE PLAN which describes the result of the dialogue:

COLLABORATIVE-PLAN A&B (Achieve A&B Goal) $\Longleftrightarrow$

1. MS A&B (Intend A $\vee$ B ($\alpha_1 \wedge \ldots \alpha_n$))
2. $\forall \alpha_i$ (MS A&B (Contribute α_i (Achieve A&B Goal))
3. MS A&B (Max-Utility ($\alpha_1 \wedge \ldots \alpha_n$) (Achieve A&B Goal))

The predicate MS used in the definition of a collaborative plan means MUTUALLY SUPPOSED, a type of defeasible mutual belief that reflects the fact that acceptance is sometimes only implicated [32, 22]. A COLLABORATIVE PLAN is a variation on a SharedPlan [12] that incorporates mutual supposition and utility theory [37, 33].

5.2 Attention/Working Memory (AWM)

The definition of COLLABORATIVE EFFORT proposed above depends on a model of agents' inferential processing and retrieval from memory. This implies a departure from simpler processing models that assume that agents are logically omniscient and can always access all of their beliefs. The cognitive architecture of Design-World

[6] This is a simplification, since in the real world, agents are often in the position of having multiple incomparable evaluation functions, e.g. of trading off aesthetic desires with desires for having money. The fact that there are multiple evaluation functions is exhibited in agents argumentation behavior in dialogue [34], however the incorporation of competing evaluation functions is left to future work.

agents includes a model of limited attention/working memory (AWM), first proposed by Landauer [17]. While the AWM model is extremely simple, Landauer shows that it can be parameterized to fit many empirical results on human memory and learning [2].

AWM consists of a three dimensional space in which propositions acquired from perceiving the world are stored in chronological sequence according to the location of a moving memory pointer. The sequence of memory loci used for storage constitutes a random walk through memory with each locus a fixed distance from the previous one. The fact that the propositions in memory are stored as a random walk produces variation in agents' behaviors on different runs of the simulation. If 8 propositions are stored, they may be stored closely together and be thus be likely to be accessible during the same search, but they may be stored as far as Hamming distance 8 apart. The frequency of exposure to propositions about the world is represented by storing items multiple times if they are encountered multiple times [16].

When an agent retrieves items from memory, search starts from the current pointer location and spreads out in a spherical fashion. Search is restricted to a particular search radius: radius is defined in Hamming distance. For example if the current memory pointer locus is (0 0 0), the loci distance 1 away would be (0 1 0) (0 -1 0) (0 0 1) (0 0 -1) (-1 0 0) (1 0 0). The actual locations are calculated modulo the memory size. The limit on the search radius defines the capacity of attention/working memory and hence defines which stored beliefs and intentions are SALIENT or ACCESSIBLE. Thus we distinguish between what is KNOWN and what is SALIENT [26, 2].

The radius of the search sphere in the AWM model is used as the parameter for a Design-World agents' resource-bound on attentional capacity. In the experiments below, memory is 16x16x16 and the radius parameter varies between 1 and 16. Agents with an AWM of 1 have access to 7 loci, and since propositions are stored sparsely, they only remember the last few propositions that they acquired from perception. Agents with an AWM of 16 can access everything they know.[7] Figure 6 shows how agent's performance improves as AWM increases, where performance is plotted on the y-axis and AWM settings on the x-axis.

The advantages of the AWM model is that it has been shown to reproduce, in simulation, many results on human memory and learning. Because search starts from the current pointer location, items that have been stored most recently are more likely to be retrieved, predicting recency effects [2]. Because items that are stored in multiple locations are more likely to be retrieved, the model predicts frequency effects [17]. Because items are stored in chronological sequence, the model produces natural associativity effects [1].

A further assumption of this architecture is that deliberation and means-end reasoning can only operate on beliefs that are accessible in AWM [19, 23]. See figure 5. One effect of this is that it is possible for agents to fail to make inferences because the

[7] The size of memory was determined as adequate for producing the desired level of variation in the current task across all the experimental variables.

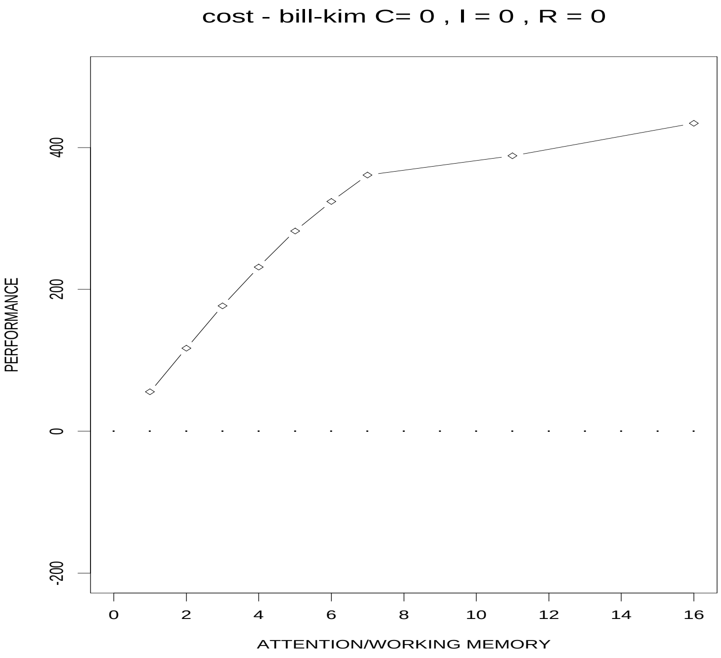

Figure 6: Performance increases for All-Implicit agents on the Design-World Standard task as Attention/Working Memory increases

premises are not accessible. Another effect is that agents can make mistakes in their planning process, because a belief that would allow them to produce an optimal plan is not accessible. It is also possible for agents to make mistakes and propose invalid steps in their plans because they forget that they have already used a furniture item.

A final experimental parameter supports parameterizing agents as to whether they always draw all the inferences that could be drawn from beliefs that are currently in working memory. This is done with a probabilistic inference model with a parameter that determines that each inference will be made with a particular likelihood, which provides a simple way of testing some of the effects of indeterminate inference. When this parameter is set to 0, the agent is called a NO-INFERENCE agent, when it is set to .5, the agent is a HALF-INFERENCE agent, and when it is 1, the agent is a ALL-INFERENCE agent. We will call ALL-INFERENCE agents logically omniscient agents.

5.3 Least Collaborative Effort

In the introduction, we defined PERFORMANCE as the difference between a measure of the quality of the problem solution and COLLABORATIVE EFFORT. In Design-World, the RAW-SCORE for the task is a measure of the quality of the problem solution. We can now refine our definition of LEAST COLLABORATIVE EFFORT within the cognitive architecture described above and shown in figure 5.

Collaborative effort involves costs for retrieval, inference and communication. The AWM model provides a measure of the number of retrievals from memory in terms of the number of locations searched to find a proposition. We can also keep track of the number of inferences that the agents make as they means-end reason and draw content-based inferences on their beliefs and the beliefs communicated from the other agent. The effort for communication is defined as the number of messages that the agents send to one another as they carry out the dialogue. So we have a way of collecting data for each of these different processes.

However we still are not in a position to produce a number that represents collaborative effort. The problem is that these various cognitive processes are not strictly comparable so we cannot simply add up the number of retrievals, inferences and messages. Even if we had a unit of cognitive effort, there is a danger that it would be specific to the particular implementation of Design-World.

The solution is to make the amount of effort required for each cognitive process a parameter of the model. We thus parametrize agents' retrieval, inference and communicative costs by (1) COMMCOST: cost of sending a message; (2) INFCOST: cost of inference; and (3) RETCOST: cost of retrieval from memory. Collaborative effort is then defined as:

$$\begin{array}{l}\text{COLLABORATIVE EFFORT} = \\ (\text{COMMCOST} \times \text{total messages for both agents}) \\ + (\text{INFCOST} \times \text{total inferences for both agents}) \\ + (\text{RETCOST} \times \text{total retrievals for both agents})\end{array}$$

In this paper the effects of varying the cost parameters is not discussed; the parameters are set so that the costs of communication and inference are equivalent and 100 retrieval steps is the same as an inference or a message.[8]

With the defintion above, PERFORMANCE in Design-World is:

$$\text{PERFORMANCE} = \text{Task Defined RAW SCORE} - \text{COLLABORATIVE EFFORT}$$

While we will not vary cost parameters here, one important point related to the method in general is that the cost parameters support modeling various cognitive architectures, e.g. varying the cost of retrieval models different assumptions about memory. For example, if retrieval is free then all items in working memory are instantly accessible, as they would be if they were stored in registers with fast parallel access. If AWM is set to 16, but retrieval isn't free, the model approximates slow spreading activation that is quite effortful, yet the agent still has the ability to access all of memory, given enough time. If AWM is set lower than 16 and retrieval isn't free, then we model slow spreading activation with a timeout when effort exceeds a certain amount, so that an

[8] See [35, 36] for results on how cost variations affect the results.

agent does not have the ability to access all of memory. Thus the AWM parameter supports a distinction between an agent's ability to access all the information stored in its memory, and the effort involved in doing so.

6 Experimental Results

In the experiments presented here we have two goals: (1) to demonstrate the simulation method, and (2) to test the hypotheses discussed above. In the experiments presented below we will compare the Close-Consequence strategy with the All-Implicit strategy. In the experimental environment as we have defined it, we can test hypotheses 1 through 4.

Hypothesis 1 is testable because Design-World agents can be parameterized to have limited AWM, with the result that they may not make particular inferences. Hypothesis 2 is testable because we can compare different strategies under different assumptions about cognitive costs for inference and communication. Hypothesis 3 is testable because all assertions make the asserted proposition accessible, thus if it is the premise for any new inferences, it is more accessible than it would be otherwise. Hypothesis 4 is testable because AWM produces both recency and frequency effects and because the Zero-Invalids task penalizes agents for forgetting.

To determine whether a strategy is COLLABORATIVE, we compare the performance of two strategies in situations where we vary the task requirements and agents' attentional capacity. Performance was defined above. Every simulation run varies the AWM radius from 1 to 16 to test whether a strategy only has an effect at particular AWM settings. Other parameters varied here are the agents' logical omniscience and whether the task is the Standard task or the Zero-Invalids task.

We simulate 100 dialogues at each parameter setting for each strategy. Differences in performance distributions are evaluated for significance over the 100 dialogues using the Kolmogorov-Smirnov (KS) two sample test [29]. A strategy A is BENEFICIAL as compared to a strategy B, for a set of fixed parameter settings, if the difference in distributions using the Kolmogorov-Smirnov two sample test is significant at $p < .05$, in the positive direction, for two or more AWM settings. A strategy is DETRIMENTAL if the differences go in the negative direction. Strategies that are BENEFICIAL are COLLABORATIVE strategies in the situation tested by the experiment.

A DIFFERENCE PLOT such as that in figure 8 will be used to summarize a comparison of Close-Consequence with the All-Implicit strategy. **Differences** in performance between the two strategies are plotted on the Y-axis against AWM parameter settings from 1 . . . 16 on the X-axis. Each point in the plot represents the difference in the means of 100 runs of each strategy at a particular AWM setting. These plots summarize the information from 18 performance distributions representing 1800 simulated dialogues. If the plot is above the dotted line for 2 or more AWM settings, then strategy 1 may be

BENEFICIAL, depending on whether the differences are significant.[9]

6.1 Close-Consequence beneficial for No-Inference agents

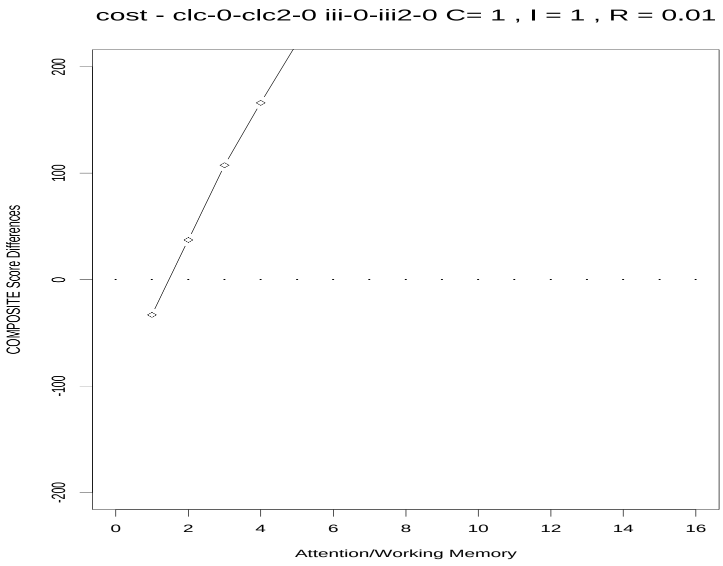

Figure 7: Close-Consequence is beneficial for No-Inference agents. Strategy 1 is the two NO-INFERENCE Close-Consequence agents and Strategy 2 is two NO-INFERENCE All-Implicit agents, Task Definition = Standard, commcost = 1, infcost = 1, retcost = .01

A version of hypothesis 1 is tested by using the logical omniscience parameter so that agents don't always make inferences. The experiment compared NO-INFERENCE agents who used the CLOSE-CONSEQUENCE strategy with NO-INFERENCE agents who didn't. Strategy 1 in the difference plot in figure 7 is a dialogue between two NO-INFERENCE agents employing the CLOSE-CONSEQUENCE strategy. Strategy 2 is two NO-INFERENCE agents using the All-Implicit strategy.

As we might have expected, figure 7 shows that the the Close-Consequence strategy of making inferences explicit is beneficial in cases where agents don't make these inferences (KS $>$ 0.5 for all AWM, p $<$ .01). The All-Implicit agents who don't make inferences explicit do very badly on the task because they have a high number of invalid steps in their plans.

In addition, the Close-Consequence strategy is beneficial for all settings of AWM above 1. This is because as agents' memory resources increase, they do better on the task, so

[9] Visual difference in means and distributional differences need not be correlated, and the KS test tests distributional differences. However KS significance values are given with each figure, and difference plots are much more concise than actual distributions.

at higher AWM settings the difference between agents who make a lot of mistakes and those who don't is even greater than at lower AWM settings.

6.2 Close-Consequence detrimental for Logically Omniscient Agents for Standard Task

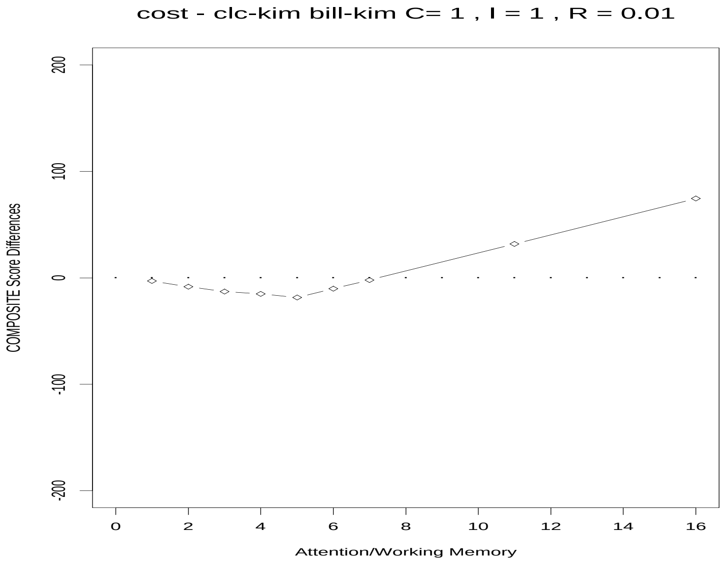

Figure 8: Close-Consequence can be detrimental for attention-limited agents. Strategy 1 is the combination of an All-Implicit agent with a Close-Consequence agent and Strategy 2 is two All-Implicit agents, Task = Standard, commcost = 1, infcost = 1, retcost = .01

Hypothesis 4 states that agents make inferences explicit just because it is important to remember them. We can test this by comparing Close-Consequence agents who are logically omniscient with All-Implicit agents who are logically omniscient in the Standard task. Because the AWM model means that agents can forget things, Close-Consequence might be collaborative in this situation.

However, the difference plot in figure 8 shows that Close-Consequence is DETRIMENTAL in the Standard task at AWM of 1 . . . 5 (KS $>$ 0.19, p $<$ .05). Remember agents can only access facts that are in their current working memory and each time an agent stores a fact, another fact may be displaced. Thus the additional utterances in the Close-Consequence strategy make agents forget other facts that would be more useful. Furthermore, these additional utterances cost effort to process, and this effort takes a large toll at low AWM when agents are not performing that well on the task.

6.3 Close-Consequence beneficial for Logically Omniscient Agents for Zero Invalids Task

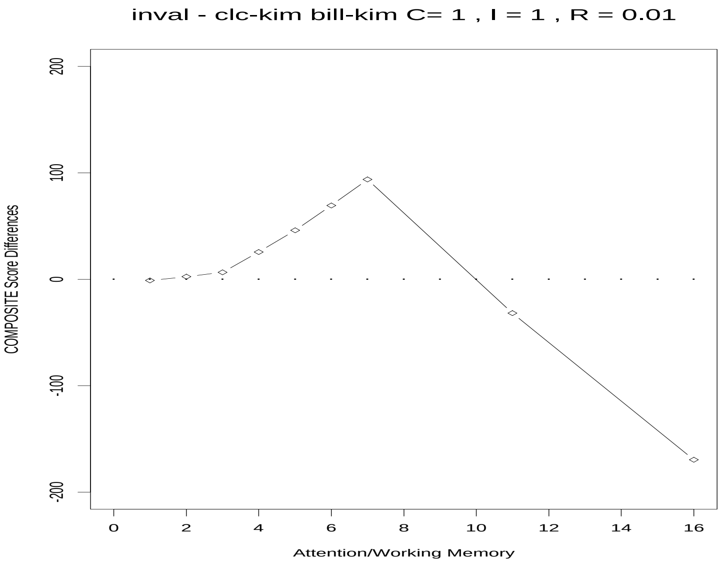

Figure 9: Close Consequence is beneficial for Zero-Invalids Task. Strategy 1 is the combination of an All-Implicit agent with a Close-Consequence agent and Strategy 2 is two All-Implicit agents, Task = Zero-Invalid, commcost = 1, infcost = 1, retcost = .01

Another experiment provides an additional test of hypothesis 4, by changing the task definition so that it is very important that agents remember what they have inferred. We test hypothesis 4 again by comparing logically omniscient agents using the Close-Consequence strategy with logically omniscient agents using the All-Implicit strategy in the Zero-Invalids task.

Remember that even when agents are logically omniscient they can forget that they have used a furniture item in the plan. In other words, the agents made the inference at the time that they agreed, but have since forgotten it. Because of this the Close-Consequence strategy is BENEFICIAL when the task is the fault-intolerant Zero-Invalids task. See figure 9 (KS for AWM of 3 to 6 $> .19$, p $< .05$).

Consequence IRUs can increase the robustness of the planning process by decreasing the frequency with which agents make mistakes. This provides support for the hypothesis that Consequence IRUs, like other types of repetition and paraphrase, can have a fundamental cognitive effect of helping the agent remember an important fact. In other words, the beneficial effect is a direct result of rehearsing the act-effect inferences, making it unlikely that attention-limited agents will forget that they have already used a furniture item.

As figure 9 shows, the Close-Consequence strategy is not beneficial At high AWM where agents are less likely to forget this information. this is because the All-Implicit logically omniscient agents with high AWM do not forget what they have inferred. When we calculate the performance difference between All-Implicit and Close-Consequence agents, the Close-Consequence agents don't do as well because they are expending effort in communicating IRUs that are not helping performance, since it is already at ceiling, and which are costing effort to process.

6.4 Discussion

The results presented above showed that the cognitive demands of the task, and their interaction with human cognitive limitations, determines whether a strategy is collaborative. Thus collaborative strategies cannot be defined in the abstract, but must be defined relative to particular assumptions about the capabilities of the architecture of the agents involved, as well as the relationship between the agent architecture and the demands of the task.

As we would have expected, figure 7 illustrated that the Close-Consequence strategy is collaborative when the inference made explicit would not have been made otherwise. In these experiments we used a probabilistic model of inference. A theory that made stronger claims about what determines process limits on inference would be useful for developing a more finely tuned strategy for making inferences explicit. Nevertheless, this result provides some insight into the situations in which Consequence IRUs are beneficial.

Figure 8 shows that the Close-Consequence strategy is actually detrimental for attention-limited logically omniscient agents as long as the task is not cognitively demanding. This is because making the inferences explicit displaces other facts from working memory and makes attention limited agents forget other facts that would have been more useful to them as they carried out the task.

Figure 9 showed that a cognitively demanding version of the task reverses this effect, and Close-Consequence is a beneficial strategy when agents have limited working memory. This is because rehearsing the act effect inferences makes it unlikely that the agents will forget that they have used a furniture item. Thus cognitively limited agents do better with a strategy that includes making inferences explicit such as Close-Consequence, when the task is sufficiently demanding.

These results parallel expected results for human agents and our intuitions about when we might make inferences explicit. However we did not test hypothesis 2 because the act effect inferences do not rely on having multiple premises simultaneously accessible. We did not test hypothesis 3 because no other inferences rely on the act effect inferences that are made explicit in the Close-Consequence strategy. We present the results of testing hypotheses 2 and 3 in other work [31], and leave hypotheses 5 and 6 for future work.

Finally the results presented here may be specific to the AWM model. AWM was chosen as an implementable model that reproduces many results on human memory and learning [17, 2]. If the goal is to test a different cognitive model of memory, or a non-cognitively based model, a different memory model can be substituted. We are currently carrying out experiments with other memory models which will allow us to investigate the extent to which our results depend on a model that produces both recency and frequency effects, and which allows agents to forget and to make mistakes in remembering, as humans do.

7 Implications for Human Computer Collaboration

This paper had two foci: (1) the proposal of a method for testing a theory of human computer collaboration; and (2) the presentation of results developed by the proposed method. I will discuss the implications of both of these below.

The simulation results show that systems for human computer collaboration must be designed to take cognitive factors and their interaction with the definition of the task into account [27]. Whenever a cognitive model can be formalized and implemented and hypotheses about collaborative strategies can be formulated, the method used here can be used to test the interaction of cognitive limitations, strategies, and aspects of the task or communication situation. The benefits of the method are both in testing a formalization of a theory and in the results, which can then be incorporated in the design of a system for human computer collaboration.

One advantage of the method is that it forces the theoretician to be precise about the models of agents' processing that a theory is based on, since these must be specified precisely enough to implement [14]. Similar methods have been used in exploring a theory of single agent resource bounded reasoning [25], in testing error recovery strategies for a dialogue system [5], in testing a theory of variable initiative [13], and testing negotiation strategies in the context of an automated librarian [10, 6].

Potential theoretical variations in the context of the work presented here would be to change the memory model, the dialogue model, or the definition of collaboration to only count human effort and not computer effort. All of these aspects of the simulation could be easily varied, and the effects would be of interest, but they would not affect the fundamental claims of this paper with respect to the the need for modeling cognitive processes and the benefits of computational simulation as a new method for theories of collaborative behavior.

8 Acknowledgements

Thanks to the participants in the workshop on Human-Computer Collaboration for discussion of this paper, especially Steve Whittaker, Susan Brennan, Chuck Rich, Candy Sidner, and Loren Terveen. Thanks also to the anonymous reviewers for many

helpful comments and to Aravind Joshi and Ellen Prince for their advice, insight and many useful discussions.

References

[1] J. R. Anderson and G. H. Bower. *Human Associative Memory*. V.H. Winston and Sons, 1973.

[2] Alan Baddeley. *Working Memory*. Oxford University Press, 1986.

[3] Michael Bratman, David Israel, and Martha Pollack. Plans and resource bounded practical reasoning. *Computational Intelligence*, 4:349–355, 1988.

[4] Susan E. Brennan. *Seeking and Providing Evidence for Mutual Understanding*. PhD thesis, Stanford University Psychology Dept., 1990. Unpublished Manuscript.

[5] Jean C. Carletta. *Risk Taking and Recovery in Task-Oriented Dialogue*. PhD thesis, Edinburgh University, 1992.

[6] Alison Cawsey, Julia Galliers, Steven Reece, and Karen Sparck Jones. Automating the librarian: A fundamental approach using belief revision. Technical Report 243, Cambridge Computer Laboratory, 1992.

[7] Herbert H. Clark and Edward F. Schaefer. Contributing to discourse. *Cognitive Science*, 13:259–294, 1989.

[8] Jon Doyle. Rationality and its roles in reasoning. *Computational Intelligence*, November 1992.

[9] Timothy W. Finin, Aravind K. Joshi, and Bonnie Lynn Webber. Natural language interactions with artificial experts. *Proceedings of the IEEE*, 74(7):921–938, 1986.

[10] Julia R. Galliers. Cooperative interaction as strategic belief revision. In M.S. Deen, editor, *Cooperating Knowledge Based Systems*, pages 148 – 163. Springer Verlag, 1991.

[11] H. P. Grice. *William James Lectures*. 1967.

[12] Barbara J. Grosz and Candace L. Sidner. Plans for discourse. In *Cohen, Morgan and Pollack, eds. Intentions in Communication, MIT Press*, 1990.

[13] Curry I. Guinn. A computational model of dialogue initiative in collaborative discourse. In *AAAI Technical Report FS-93-05*, 1993.

[14] Steve Hanks, Martha E. Pollack, and Paul R. Cohen. Benchmarks, testbeds, controlled experimentation and the design of agent architectures. *AI Magazine*, December 1993.

[15] Heritage and Watson. Reformulations as conversational objects. In George Psathas, editor, *Everyday language: Studies in Ethnomethodology*, pages 123–163. Irvington Publishers, New York, 1979.

[16] D. L. Hintzmann and R. A. Block. Repetition and memory: evidence for a multiple trace hypothesis. *Journal of Experimental Psychology*, 88:297–306, 1971.

[17] Thomas K. Landauer. Memory without organization: Properties of a model with random storage and undirected retrieval. *Cognitive Psychology*, pages 495–531, 1975.

[18] Stephen C. Levinson. Some pre-observations on the modelling of dialogue. *Discourse Processes*, 4:93–116, 1981.

[19] Gail McKoon and Roger Ratcliff. Inference during reading. *Psychological Review*, 99(3):440–466, 1992.

[20] Johanna D. Moore and Cécile L. Paris. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19(4), 1993.

[21] Margaret G. Moser and Johanna Moore. Corpus analysis of discourse cues. Technical report, University of Pittsburgh, LRDC, 1994.

[22] Gopalan Nadathur and Aravind K. Joshi. Mutual beliefs in conversational systems: Their role in referring expressions. In *Proc. International Joint Conference on Artificial Intelligence, Austin*, pages 603–605, 1983.

[23] Donald A. Norman and Daniel G. Bobrow. On data-limited and resource-limited processes. *Cognitive Psychology*, 7(1):44–6, 1975.

[24] Martha Pollack, Julia Hirschberg, and Bonnie Webber. User participation in the reasoning process of expert systems. In *AAAI82*, 1982.

[25] Martha E. Pollack and Marc Ringuette. Introducing the Tileworld: Experimentally Evaluating Agent Architectures. In *AAAI90*, pages 183–189, 1990.

[26] Ellen F. Prince. Toward a taxonomy of given-new information. In *Radical Pragmatics*, pages 223–255. Academic Press, 1981.

[27] Erika Rogers. Cognitive cooperation for visual interaction. In *AAAI Technical Report FS-93-05*, 1993.

[28] Candace Sidner. Using discourse to negotiate in collaborative activity: An artificial language. *AAAI Workshop on Cooperation among Heterogeneous Agents*, 1992.

[29] Sidney Siegel. *Nonparametric Statistics for the Behavioral Sciences*. McGraw Hill, 1956.

[30] Adelheit Stein and Ulrich Thiel. A conversational model of multimodal interaction. In *AAAI93*, 1993.

[31] Marilyn Walker. Planning to converse with an inference limited agent. In *AAAI Workshop on Planning for Interagent Communication*, 1994.

[32] Marilyn A. Walker. Redundancy in collaborative dialogue. In *Fourteenth International Conference on Computational Linguistics*, pages 345–351, 1992.

[33] Marilyn A. Walker. *Informational Redundancy and Resource Bounds in Dialogue*. PhD thesis, University of Pennsylvania, 1993.

[34] Marilyn A. Walker. Redundant affirmation, deliberation and discourse. In *Penn Review of Linguistics*, volume 17, 1993.

[35] Marilyn A. Walker. Discourse and deliberation:testing a collaborative strategy. In *Coling 94*, 1994.

[36] Marilyn A. Walker. Experimentally evaluating communicative strategies: The effect of the task. In *AAAI94*, 1994.

[37] Marilyn A. Walker and Steve Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proc. 28th Annual Meeting of the ACL*, pages 70–79, 1990.

[38] Steve Whittaker, Susan Brennan, and Herb Clark. Coordinating activity: an analysis of interaction in computer supported cooperative work. In *CHI 91*, page 1, 1991.

[39] Steve Whittaker, Erik Geelhoed, and Elizabeth Robinson. Shared workspaces: How do they work and when are they useful? *IJMMS*, 39:813–842, 1993.

[40] Steve Whittaker and Phil Stenton. Cues and control in expert client dialogues. In *Proc. 26th Annual Meeting of the ACL, Association of Computational Linguistics*, pages 123–130, 1988.