Meanings and Messages

Iain D. Craig
Department of Computer Science
University of Warwick
Coventry CV4 7AL
UK EC

January 24, 1995

Abstract

The problem of meaning in DAI systems is examined in some depth. Agents in DAI systems communicate by exchanging messages, as well as by sensing and acting upon their environment, although message exchange is the focus of the current argument. The problem for such systems is that of determining the meaning of a message and acting upon it in a way that displays understanding. The current analysis centres on the grounds on which meanings can be ascribed by agents to each other. After considering the role of context and expectation, private states are examined: the conclusion that is drawn is that internal states underdetermine meanings and that other factors must be brought into play when deciding upon the meaning of messages and the understaning of agents.

1 Introduction

For a long time, I have been dissatisfied with my CASSANDRA architecture [4]. In a recent paper [5], I proposed a number of extensions to the published architecture. The main proposals were:

- The provision of an organisational context within which agents operate.
- The provision and use of reflective capabilities within each agent (this included the provision of a declarative representation to augment the purely procedural one used in the original architecture).
- The use of more structured communications between agents (the model that was proposed was speech acts [1, 11])[1].

In [5], I took it as read that the inclusion of an organisational structure was a 'good thing', and did not argue for it. This position seems to be very much in line with

[1] This proposal dates back to [4].

some recent trends in DAI work (see, for example, [6] and [9]). In the short time that has elapsed since writing [5], I have come to view the organisational aspects as being more important than a mere 'good idea'. This change of view is a result of work on reflection and representation. Indeed, the subject of this paper results from this work.

The aim of this paper is to suggest that there is more to meaning than meets the eye. Much of what I have to say may seem obvious, even trivial, to some, but it seems a natural consequence of my work on meaning and knowledge representation, and on reflection in problem-solving agents, and also appears to be in line with the work of a number of other researchers (e.g., Barwise and Perry [2], Smith [14, 12, 13], Suchmann [15]). Indeed, the context of distributed and autonomous agents has always seemed a natural one within which to examine the problems of self-modelling and reflection. The arguments and examples that I will present below have been on my mind for some considerable time, but they have not, hitherto, seemed to add up to more than a collection of observations.

More specifically, I want to tackle the question of how messages can have meanings. In a DAI system, agents communicate by sending messages. The flow of messages conveys information of different sorts between agents. By sending and receiving messages, agents in this kind of system obtain information about the state of the world and about the state of other agents in the system.

The question can be paraphrased thus:

> *If A sends $\mathcal{M}$ to B, and $\mathcal{M}$ means ψ to A, how can one say that B means ψ by it and not σ (where ψ and σ are incompatible or even contradictory)?*

In other words, how can we (as observers), or A, know that B understands that $\mathcal{M}$ means what A means by it, and act on it in accordance with what A requires or intends? The problem is important because there is no guarantee that B will act a reasonable or sensible way in response to $\mathcal{M}$ without some assumption of shared meaning or interpretation. For coherent or coordinated behaviours amongst agents, such a requirement seems inescapable. As will be seen, this question has implications for the concept of the representation of knowledge.

The idea of questioning meanings in messages may seem a little absurd, for unless a message has a meaning, it fails to communicate anything. In a computational setting, a message can have any meaning that *we* assign it, for it is, after all, only a sequence of bits (or, more abstractly, is only a sequence of symbols). As will, I hope, be seen, this view misses the point by a long way. Indeed, the very idea that we, the designers of systems, can arbitrarily assign meanings to messages begs a number of important questions, one of which being how we as designers can construct intelligent systems composed of many autonomous agents[2].

[2] In an analogous fashion, Nilsson's remarks in a recent paper [10] begs a number of questions about the knowledge and role of designers in the construction of conventional intelligent systems.

The starting point for the discussion that follows is this: messages must serve some purpose. For a long time, I have believed that the *very fact that* there is a communication between two agents is, in itself, significant and that inferences can be made on the basis of that fact. The account that I want to give rests upon a complex of ideas, including meaning-as-use, a functional account of mind, the idea that reference is, at base, causal, and that understanding and inference are intimately connected with actions.

2 Assumptions and Background

Imagine two agents A and B who exchange messages in order to communicate. Both A and B are autonomous: they act independently of each other. Furthermore, assume for the time being that A and B can *only* obtain information by sending and receiving messages: these other messages can come from and go to agents distinct from A and B. The other agents (which are also assumed to be autonomous) might be able to sense the environment in which they operate, and they might also be able to alter it in various ways, as might A and B: these possibilities are ignored for the time being, but are worth bearing in mind as background setting. The environment within which A and B find themselves can be assumed to be composed of both A and B and the other agents in the system or community; if it is in any way real, it also contains non-agentive objects (physical objects, for example). The environment containing A, B and all the other agents can be considered to be the actual world in which we all live, or some part of it. The reason for making the environment the world is that I do not want to assume any artificial constraints: I want it, that is, to be as general, rich and seemingly problematic as the world in which we live.

Both A and B are charged with performing certain tasks: some tasks will require co-operation or co-ordination between A and B. The nature of the tasks that they are charged with performing is not important: what is important is the fact that they both *do* things. The interest, in this paper, is not in the details of A and B's tasks, but in the other things that they do in order to perform these tasks. As has been said, one thing that they must both do is *communicate* with each other. It can be assumed that they communicate by passing messages (whether the messages are on pieces of paper, or electronic mail, or are symbolic structures is immaterial). The question that must be answered is how A and B mean things by their messages. To begin with, the fact that these two agents only communicate via message-passing is merely a simplifying assumption: later (particularly in sections four and five), I will allow the possibility that each can sense its environment and also act upon it in ways which the other (or both) can detect.

Clearly, if A and B communicate by sending messages, they must communicate in order to pass information. What is assumed is that the messages sent between A and B mean something to both: that is, it is assumed that if A sends B a message, A must mean something by the message, and B must also interpret the contents of the message in such a way that the message can be said to be meaningful. It seems to make little sense for A to send B a message that A considers meaningless while, at the same time, expecting B to do something with that message. In a similar fashion, it makes little sense for B to receive a message from A and believe that A means nothing by it. The question, in other terms, is "how can both A and B mean something by sending a message?" A related question is the following: if A sends B a message $\mathcal{M}$, and A means ϕ by $\mathcal{M}$, whereas B means ψ, who is to tell who is right?[3] The second question requires a theory of meaning for messages just as much as the first. It also raises the question of how to go about determining whether two agents mean the same thing by a message.

It seems reasonable to focus, then, on how agents can mean things in and by messages and on how they can diverge in their ascriptions of meaning. Otherwise stated, given the above setting, how is it possible to ascribe meanings to messages? To answer this, it is necessary, first of all, to consider what constitutes a message and to account for why it should be exchanged. Next, it is necessary to examine the context within which messages are exchanged: context takes into account a number of different factors, not just those directly related to the tasks which A and B are trying to accomplish.

3 Agents and Messages

Agents exchange messages in order to communicate: for present purposes, messages will be taken to be distinct entities which are exchanged in an act of communication. That is, I am taking 'message' to be interpreted in a technical sense, one that is akin to its sense in "message-passing architecture" (I do not intend 'message' to be interpreted in, for example, the sense of "the medium is the message"). For present purposes, I will usually assume that only two agents are of immediate interest.

The aim of communication can be to inform (for example, uttering indicative sentences), to instruct, request, enquire or suggest: these are clearly only some of the uses of communication. The following can be separated out from an act of communication:

- The *content* of the message that is exchanged.
- The *mode of presentation* of the message (see below).
- The *fact* that the message has been sent.

These three aspects can be explained as follows. The content of a message is taken to be that which the message is about (in some appropriate sense): it is the information that is to be communicated. The mode of presentation is the way in which the

[3] This is, as far as I understand it, the problem posed by Gasser [7]. Unfortunately, I do not, at present, have a copy of it.

message is to be taken: it can be explained in terms of the speech act that is associated with the message. Messages can ask, promise, inform, instruct, order, and so on. As will be seen, what I call the *mode of presentation* is important to the way in which the message is interpreted (in a way analogous to that identified in Speech Act theory). The final aspect, initially at least, seems to require no explanation, although it will become more important below.

This analysis of a message exchange will help in trying to determine how meanings could be ascribed. Given the above analysis, a message can be assumed to be of the form:

$$\mathcal{M} = \langle \mu, \phi \rangle$$

where μ is the mode of presentation and ϕ the content. The problem is accounting for the content of the message. An initial account is immediately clear.

A message is exchanged between two agents with some mode of presentation μ if the sending agent (A) wishes to communicate ϕ to the receiving one (B). In the case of messages whose mode is that of informing, it is clear that A will send B a message $\mathcal{M}$ just in case A wants B to know that ϕ or that A believes that B needs to know ϕ.

Upon receipt of $\mathcal{M}$, B knows that A wishes it to know that ϕ; it also knows that A knows or believes that ϕ. What B infers upon receipt of $\mathcal{M}$ is a direct result of the fact that A has sent B that message (assuming that A did not send B the message by mistake, but this possibility will be ignored—it will be assumed that all messages are always sent to the correct or intended recipient). A similar account can be given for messages whose mode of presentation is different from informing.

For example, if A asks B whether ϕ, B can infer that A does not know whether ϕ, and can also infer that A believes it to have that information. (If the message were something like "Do you have a ψ?", similar things can be said.) What this shows is that A must hold beliefs about B and that B holds them about A. It also shows that both A and B will alter their beliefs about each other as a result of exchanging messages. In the case of replies, this is obvious. In the case of, for example, A informing B that ϕ, after sending the message, A is entitled to believe that B now knows or believes that ϕ (assuming that communications are reliable: in some cases, A might want to wait for a response from B before becoming committed to the belief—again, the simpler version will be adopted).

This immediate account seems fine as far as it goes. It fails, however, to account for the content, ϕ, of the message $\mathcal{M}$. What the account does do is to talk about the beliefs that each agent has as a result of exchanging $\mathcal{M}$. It shows that the agents not only exchange the content ϕ, but also are able to make inferences about each other's beliefs as a result of communicating. What is needed is to account for the way in which the content is taken by each agent: i.e., the meaning of $\mathcal{M}$ must be given in such a way that both A and B would agree on what $\mathcal{M}$ means.

The immediate account is as follows. For A to send a message $\mathcal{M}$ to B, and for A to mean ϕ by $\mathcal{M}$, and for B to understand ϕ by it, the following conditions must be met:

1. A and B must share a common language.

2. A and B must have background knowledge in common (i.e., they must posses knowledge that is common to both).

3. A and B must make the same inferences from ϕ, hence form the same beliefs as a result of knowing that ϕ.

4. The actions of A and B as a result of ϕ, all other things being equal, must lead to the same or to similar results.

Although containing the basics of an answer, as well as informing the analysis given above, I believe that this account fails to give a true account. The reason for this is, I believe, that it takes too narrow a view of the agents, and that it fails to take into account the situation in which the agents find themselves.

By taking the situation or the context into account, we end up with a more complex account of meaning. Indeed, we end up with an account in which it might seem that meanings cannot be the same for two agents. The above account only considers knowledge and inferences seriously: actions are accorded a relatively minor role in the theory. This is because it is the inferential aspects of understanding that are taken to be central: understanding is seem in terms of the inferential consequences of ϕ together with common knowledge—in other words, it is the body of common knowledge that licenses similar conclusions to be drawn by both agents.

The context within which the message is exchanged is ignored for the reason that the body of common knowledge is seen as the means, together with inference, by which the content is explicated. The concept of action is minor because it need not affect common knowledge in any serious way: in any case, action is relegated to the role of being an additional way of determining that A and B understand the same thing. Furthermore, the conventional account concentrates on the *propositional content* of messages in attempting to give an account of meaning and interpretation.

4 Messages and Contexts

I want to argue that the conventional account of content and meaning fails because it ignores the context in which the agents are located. The above account rests upon a corpus of background knowledge which is used to fix the interpretation of a message. If the content of a message contradicts something that is known by an agent, what should be said of that message? One response is to reject it. Another would be to determine why it does not fit. Given the individual beliefs of agents, the interpretation of a message may vary, even contradict. If A has beliefs $\mathcal{B}_A$ and B has beliefs $\mathcal{B}_B$, and A sends ϕ to B, B may assign ϕ an interpretation very different from A, even though ϕ is consistent with the shared knowledge: i.e., background knowledge may suggest a particular interpretation, but current beliefs another.

In this section, I want to concentrate rather more on the beliefs that an agent currently holds. The reason for this is that the beliefs that are held by any two distinct agents may differ in radically different ways, even though there are propositions which both would, in the normal course of events, assert. Even if the background knowledge possessed by agents is held constant, there remains the possibility that their beliefs will differ. This is because agents will experience different things and will be engaged in different tasks. In addition, agents will be in different contexts. Part of what must be done is to relate the internal states of agents with the state of the external world. As was said above, the external world is assumed to consist of other agents as well as things like physical objects and processes. This is not to say that background or shared knowledge is not important: all I want to do is suggest that the current beliefs play a more important role, even though such dynamism causes problems.

The account briefly presented in the last section centered around the role of shared or background knowledge in interpretation. The discussion began with a consideration of A sending a message, $\mathcal{M}$, to B. It can be inferred that A wants B to do something: that something can be an externally visible action or it can be altering its internal state in some way. The basic assumption is that messages make agents do things, and that doing things (either internally or externally) is the point—the purpose, if you will—of the agent's being there in the first place. In DAI, it makes little sense for a system to contain only *solipsistic* agents: these agents merely reflect on the world they observe, and do not act in any way. In such a system, how would an external observer, or even one of the agents, know that anything had been achieved—how, in more conventional terms, would one know that a problem had been solved? It seems to me, at any rate, that one would not: since the agents in such a system merely reflect, they do not act, and, hence, no solutions or conclusions can be extracted; equally, the environment is not modified in any way. At least one consequence of this is that agents need to *act* if only to produce solutions.

What this is meant to indicate is that the context in which an agent is situated influences its behaviour: communication is, after all, an action, and, in a DAI system, it is the action by which one agent exchanges information with another. In other words, communication is one way in which an agent interacts with its environment. For the agents defined in section two, communication is the *only* way in which they can learn about what is happening and about what they should do in response to environmental changes. When one agent sends a message to another, it must do so in order to achieve something. What it intends to achieve will depend, in part, upon its context (i.e., upon its environment—environment construed in the wide sense). It would appear sensible to examine potential reasons for agent A to send agent B a message $\mathcal{M}$ of mode μ and with content ϕ.

Because of the role of action outlined above, it seems useful to begin by considering actions and intents. To send a message, A must be doing something that is related to what B is doing: simply sending a message at random serves no purpose. If, for example, I open the telephone book at random, pick out a number, dial it and, when answered, say "Please pass me the large saucepan" or "Do you know where the *Radio Times* is?", I will be doing something very odd and rather pointless, and for fairly obvious reasons—if I were even to elicit a response, it would probably not be one that I would record in a paper. If, on the other hand, I went through the process of randomly selecting a number, dialled, and then said "Do you know that headline inflation is down to 5.8%", I will still be acting in an odd way, but I might elicit a resonable response (the response might contain a question about who I am and why I am calling, but it might lead to a discussion about the state of the economy).

In the first case, the person at the end of the line is not situated in the same context as I am, and probably does not care whether I can find the large saucepan or the *Radio Times*[4]. In the second case, the other person is involved in the process of coping with the UK economy in 1991. It would not be reasonable to expect someone randomly selected from the population of a large city to know where my large saucepan is, and it would be unreasonable to expect them to know. What is expected is, thus, important in sending a message.

If, on the other hand, I were to ask my wife where the *Radio Times* is, not only do I expect her to provide me with some sort of answer (which includes the possibility that she does not know), but *she* will then expect things of me. One thing that she will expect is that I will behave in a way that can be predicted from my now knowing where to look up today's Radio 3 programmes, provided I believe her reply. That is, my wife would expect me to go to the place where she claims the *Radio Times* to be and to locate it. If it is not there, I can say that it is not there and ask her again where she thinks it is. She would probably come and look at the place where she initially claimed it to be, or suggest another place[5]. Expectations can be formed by both parties to a communication. The way that my wife shows she understands my request to be told the location of the *Radio Times* is to tell me where she thinks it is or to hold it up to show me. The way that I show her that I have understood her reply is to go and look or take it from her.

In the case of looking for the *Radio Times*, I want to achieve something (finding my copy), and, by communicating, I want to achieve the state of knowing where it is. The communication is, of course, motivated by my belief that my wife knows where to find our copy. In this example, both parties have an interest in the activity:

[4] A less bizarre example is one that often happens. Imagine you are on the phone while there is someone else nearby. It is possible to engage in two conversations simultaneously, or to respond to a question from the person in the same room. Sometimes, the two conversations get mixed up and the person at the other end of the phone can become very confused about what is going on.

[5] She could accuse me of being blind or just curse me, but (i) I discount this possibility, and (ii) she does not usually behave in these ways—another expectation!

I want to find out what's on the radio and my wife wants to help me. If we both want to listen to a certain piece of music, and the hunt for the *Radio Times* is aimed at determining when we should listen, we both have a common interest in that determination in a clear sense.

Similarly, if it is my turn to cook and I cannot find the large saucepan, my aim will be to do the cooking and my wife's will be of helping me achieve that for it means that she is able to do something that she wants to do (and not cook). It is, of course, possible, that I will get a helpful reply to my question *even if* my wife has no particular interest in my doing some cooking: she may simply wish to co-operate with me. The point though, if it needs spelling out, is that beliefs form a basis for communication.

In the case of finding the large saucepan, I believe (amongst other things[6]) that:

- I need the large saucepan to do some cooking.
- My wife knows where it is.
- My wife will tell me where it is if I ask her.

On the other hand, my wife has the following beliefs (amongst others):

- That I need the large saucepan to do some cooking (this may be inferred or I may have told her—the distinction does not matter for present purposes).
- That I cannot locate the large saucepan.
- That I believe that she knows where it is.
- That by telling me where she thinks it is, I will find it.

It should be noted that some of the things we both believe deal with knowledge that the other person has. This knowledge of the other agent is important, for helps in the formation of the request and of the reply.

Another example will bring this out a little more clearly. Imagine that we have just moved house and I want to cook the first meal (say, because we have a new hi-tech cooker and I want to try it out). Imagine that I have unpacked the cooking utensils and have arranged them in the kitchen, but was distracted when I put the saucepans out and put the large one in a cupboard. My wife does not know where I have put the saucepans, but I believe that she does. My wife knows that I want to try out the new cooker by cooking something. In response to my question "Where is the large saucepan?", my wife can justifiably reply "I don't know". My belief about her knowing its location is changed, and I can show that I understand her reply by hunting for the saucepan. She will believe that I think she knows where it is, but she will also believe that I was mistaken. She will still understand my original request and will display this understanding by her denial.

[6] These other things might be quite irrelevant to cooking, note.

What this is intended to show is that, even when one party does not know something, the beliefs that both agents have are still important in the formulation of the messages they exchange. It also shows that the knowledge possessed by one or both agents enters into the exchange. If one agent is incorrect in its ascription of knowledge, the behaviour of the other can be used to indicate this (taking the utterance "I don't know" as a behavioural item, which seems fair enough). If I randomly telephone someone and ask them where the large saucepan is, I will not have the slightest idea as to whether they know that I have such an article, or where it is. There is much more to be said about knowledge and belief, but enough, I believe, has been said to give a general feel for the account that I am proposing.

Note, also, that, in the case of the large saucepan, there are important contextual factors. By 'large saucepan', I mean the largest saucepan that we actually have at the time I utter that phrase, and not some other—say, the largest saucepan owned by our neighbours, or the largest one in the city where we live. If I went out and bought another, larger, saucepan, *that* would become the referent of 'large saucepan'. This would be so, unless we were to dub it 'largest saucepan': that, given my idiolect, is an unlikely option, for I would normally utter "largest saucepan" when explaining or clarifying the referent, or intended referent, of the phrase 'large saucepan' as in "I mean the *largest saucepan* that we have". Similarly, when I say '*Radio Times*', I am taken to be referring to the edition which contains programme details for the day on which I utter the name, not the one for the following week (which I usually buy on Mondays), nor the one for the previous week (which I usually keep so that I can try to finish the crossword). Of course, I might say "Where is the *Radio Times*?" and then make the referent clearer by saying something like "No, the one with next Tuesday in". Even in the last case, there is still a dependency, for the successful reference will depend upon the fact that I have bought next week's edition and that the hearer believes that I have.

What this all amounts to is the fact that, not only must the referent be known to speaker and hearer, but they must be in a context which will allow the referent to be determined. If the context is changed, so too will be the referent.

For example, if my wife and I go to my parents' house in order to cook a birthday dinner for my mother and I say "Where's the large saucepan?", I will not be referring to the saucepan that satisfies that description in our home, but will be referring to the largest saucepan that my mother owns. By changing the location, the referent of 'large saucepan' will change. The case of buying a new and bigger saucepan shows that time intervenes in the process of determining the referent. Time and location seem obvious factors in reference (compare the current referent of 'the King of France' with the referent in May, 1770), so none of this is surprising (or even new). What I am suggesting, though, is that time, location, knowledge and belief are important in determining referents and in determining meaning[7]. What

[7] Of course, I am assuming that a theory of meaning must, to some extent, involve a theory of reference: I can see no reason why this should not be so.

matters for DAI, in particular, is that each agent must be aware of its context.

Now, the exchanges that have been considered so far have been of a particular kind. That is, in the examples, both my wife and I are assumed to have good grounds for uttering what we do and for ascribing knowledge and beliefs to the other (although, as was seen in the last example, such ascriptions might be mistaken or wrong). It seems fair enough to claim that in the cooking examples, my wife will have good grounds for viewing my request for the saucepan as being well-founded and genuine. These grounds condition her response. Now consider the case of someone, say X, who is sitting in a pub and who has a sudden, intense chest pain. The person sitting next to X, say Y, who is a carpenter[8] and who is known or believed by X to be a carpenter, says "You have angina". X, with the presence of mind only to be found in examples, asks Y what he does: Y replies that he is a carpenter. It would be quite reasonable for X to discount Y's diagnosis and to believe that the pain was heart-burn. Such a conclusion is warranted by the fact that carpenters are not usually qualified physicians, so their diagnosis of chest pains may not be relied upon.

Again, overt behaviour also plays a part in understanding the message. In other words, X can claim of Y that he does not understand what he was talking about: the message is, therefore, meaningless in an appropriate sense of the word.

The conclusion I want to obtain is the obvious one: when agent A sends a message to B, B's view of the content of the message will be conditioned on how it views A's ability to communicate what is in the message. That is to say, B must hold the belief that A is able or competent to communicate the content of the message. Even if the carpenter reached into his pocket and handed X a glyceryl trinitrate tablet, one would still not be convinced that the pain was caused by angina. If Y were X's G.P., matters might be a little different, especially if Y examined X before making the diagnosis, or if Y stated that he had long suspected X had this condition, or if Y said "Try sucking this" and gave X a nitro tablet (if X's pain subsides, it might be reasonable for Y to state that he believes that X has angina). The behaviour displayed by Y would lend credence to the claim that he knows what he is doing and, hence, has good grounds (or, at least, better grounds) for producing that diagnosis. It would not be reasonable for anyone, physicians included, to make a diagnosis on the basis only of chest pains, so an immediate diagnosis was too hastily reached for it to be treated as correct (it might be a good guess or the best working hypothesis, but that is a different matter), a point which indicates that, unless Y has previous information, or engages in some questioning or an examination, it is reasonable to conclude that the diagnosis is not final[9].

[8] I am not attempting to denigrate carpenters. All I want to do is to point out the grounds on which we would confidently believe in someone's competence. Painter or airline pilot are perfectly good alternatives for anyone who objects to my use of carpenter.

[9] Of course, if X and Y were doing something like putting up a fence, and if X had accidentally driven a nail through his hand, Y could quite reasonably and accurately determine that X had hammered a nail into his hand and needed to be taken to hospital for the nail to be extracted and

The essential point is that there must be *grounds* (ideally, very good ones so that acceptance is rational) for accepting that Y is able to say whatever he or she says. We expect G.P.s to be in a position to diagnose illnesses or to give accounts of the cause of pain; we do not expect carpenters to be as well-placed. What we expect of someone, given our knowledge of, or beliefs about, them thus provides additional information with which to interpret what is said by them. The context in which a utterance is made determines whether credence is placed in what is said. We are all fairly capable of determining whether someone has nailed his hand to a fence, but we are not all as capable of diagnosing angina or providing remedies for the country's current economic problems. If, on the basis of good evidence, it is concluded that Y is not qualified to state ϕ with any authority, X may choose to ignore what Y says in that connection (and might even claim that Y's utterances are "meaningless").

I think it fair to infer that expectation is part of the process of interpreting a message—that is where the argument leads. What is expected of the person who makes an utterance determines, at least in part, what we take the content to be. Furthermore, the determination of content gives rise to new expectations. In addition, the current context will also determine a variety of conditions that one would expect to be satisfied by anyone who utters meaning ϕ by it. Of course, one's beliefs about someone or something may be false. The carpenter in the last example might turn out to have been a physician at some time (after making X swallow the nitro tablet, Y might say "I used to be a surgeon at Guy's" and then provide convincing evidence to support that claim).

The relationship between expectation and content needs to be clarified, for it might seem that the explanation just offered is paradoxical for the reason that an utterance depends upon, and is, at the same time, instrumental in forming expectations. The argument has been that expectations of one kind help in determining content; expectations of another kind are involved in deciding whether someone has understood the content of a message. The expectations of the second kind are expectations about future behaviour—physical or linguistic (if one can make this distinction with any clarity). Expectations of the first kind are brought to bear when one hears an utterance: what one expects of and believes about the speaker enable one to determine whether the speaker has grounds for engaging in that communicative act. What I have been arguing for is that it is against the background of previously formed expectations and beliefs that one judges the current utterance. There is no claim, it should be stressed, that the expectations must be *consciously* entertained: that remains an option, but I am not claiming that it is either necessary or sufficient for determinations of this kind. The seemingly paradoxical nature of expectation arises, quite simply, because the two kinds are confused. Of course, it remains the case that, because of some utterance, the hearer may alter his or her expectations, but this is as a result of determining the content of an utterance: it

an anti-tetanus injection administered.

is uncontroversial to claim that utterances, once understood, can alter beliefs.

What, now, of the case in which A utters $\mathcal{M}$ meaning ϕ at time t_1, and utters $\mathcal{M}_1$ meaning $\neg\phi$ at time t_2, where t_2 is later than t_1? If $\mathcal{M}$ and $\mathcal{M}_1$ are both asserted, apparently A has contradicted himself by first stating ϕ and then stating $\neg\phi$. If everything is held constant, then this seems inescapable (for some other modes of presentation, this conclusion is less certain, although contradiction reappears in the case of imperatives—A says "Do F" and later says "Don't do F"). On an account of meaning which is based on model theory (as is the conventional one), A's beliefs and knowledge can have no model, yet there are occasions on which we all contradict ourselves in a variety of ways. If A asserts that "Grass is green" on Monday, yet asserts "Grass is not green" on Tuesday, B is entitled to belief that A knows nothing about "green" or about "grass", or that A is not in a legally or socially acceptable state of mind. If, on Sunday lunchtime, A says "Mow the lawn", and then at three o'clock says "Don't mow the lawn, help me repair the fence", B may think nothing of it. A might have changed his mind, or may consider repairing the fence as being more important at that time. However, if, at noon on Sunday, A says "That sweater is green" and at one o'clock says "That sweater is brown", B may still not be ready to draw a contradiction, but, instead, believe either that A was wrong at one o'clock, or that the lighting conditions have changed[10]. If, on Easter Sunday, A asserts that "The lawn is green", but on June 15th asserts that "The lawn is yellow", B may still refuse to claim a contradiction, even though the lawn referred to on June 15th is the same one as on Easter Sunday, for the weather in the intervening period may have been hot with little rainfall and the lawn may have become discoloured as a result. Drawing an absolute contradiction appears to be something of a special case when context is taken into consideration[11].

5 Messages and Privacy

So far, discussion has concentrated on the *public* aspects of communication: that is, on the behaviour that can be expected of agents and on the publicly available content of a message. In this section, I want to go inside the agents and try to determine what sorts of things go on and what sorts of things are represented. That is, I want to concentrate on those aspects of communication that can be taken to be *private* to an agent. The assumption is that when I (or anyone else, for that matter) receives a

[10] The carpet in my sitting room at home has the property of seeming to change colour in different lighting conditions: sometimes it looks green to me, sometimes brown. At different times, I make different assertions as to its colour, even though the *actual* colour does not change.

[11] Subject-matter is also important: the class of sentence that seems most obviously to be prone to the accusation of contradiction is that dealing with *a priori* properties and objects—numbers, for example. Things which change very slowly such as stars, mountains, or the locations of countries are similar in that one can be accused of self-contradiction when making two or more assertions. However, with purely 'factual' subjects, one always has the defence that one was ignorant of the true facts.

message in some form, something "goes on" in my head which is not observable by anyone else. What goes on in the head may condition responses—actions—at some later date, or it may remain private for ever. That is, no external observer may ever determine what happened when the message was received. What I do not want to consider are the kinds of epistemically private entities that inform Wittgenstein's [16] Private Language argument: I am rather more interested in private entities that *could* become public if need be (say, by uttering a sentence or by performing a certain action).

Now, the point of this paper is to determine how agents can mean things by exchanging messages. The account that I have proposed is based upon the beliefs that an agent holds when it receives a message, and also upon the idea that beliefs are part of how actions are performed. Part of the thrust of the argument in the last section was that one agent can determine whether another has understood a message by examining the second agent's behaviour after the message has been received. Messages seem, therefore, to play a *causal* role in determining future behaviour. The account rests upon observable behaviour because I do not want to make agents perform psychological or physiological experiments on each other in order to determine responses: people use behaviour as a guide—often, they think a good one—to making determinations.

Before moving on, I want to argue against one or two objections to the idea that two different agents can mean the same thing by a message.

The first objection is simply put. The way in which meaning is construed is in terms of causal relations (and not purely logical ones). Because of the explicit effect of context on the account of meaning, and because the causal influences and previous experiences of any two agents will be different (in the case of the two agents above who only experience the external environment via message-passing this reduces to the fact that they will previously have received different messages), the beliefs that they have will be different. Because beliefs are used as a way of determining content or meaning, it immediately follows that any two agents will differ in their ascriptions of meaning or content. Therefore, the objection goes, two agents cannot agree about the meaning of a message because, necessarily, they will not have the same beliefs.

This argument is wrong for two reasons. The first is the implicit assumption that sameness is identity. Of course, two people will hold different beliefs if the criterion is identity. Trivially, only I can hold the beliefs which I actually hold; more reasonably, it is highly unlikely that I hold *identical* beliefs to anyone else. It seems highly dubious to claim that mono-zygotic twins share *identical* beliefs: they are different people and will differ in subtly different ways—their experiences, even if they have never been parted, will differ because they are *distinct* people. Twins, though, can hold similar or equivalent beliefs. Because each twin can respond in the same or in similar ways to some utterance (for example, "Please pass me the *Radio Times*"), the requirement that beliefs be *identical* seems too strong. Of course, it is necessary to say in respect to what the beliefs are similar, and that turns out to be more difficult than one might expect. A first account requires that similarity be based upon the similarity of action: performing what observers would count as an action-token of the same type where the action-token is counted as an appropriate one for the circumstances. Imagine the situation in which my wife and I are at her parents. My mother-in-law at different times of the day asks us each to put the kettle on. Both my wife and I know where the kettle is and how to fill it and switch it on. We both correctly boil water in the kettle on two different occasions. We do this, however, when we hold beliefs that are distinct but similar.

Futhermore, consider the case of understanding 'red'. It does not matter that I might have a mental "picture" of the colour of a stop light and someone else has a picture of the colour of a well-tended lawn: i.e., it does not matter that *I* have a picture of something that is "red", but *you* have a picture of what I would call *another*, and totally different, colour. Colour-blindness apart[12], as long at the internal classification is systematic in its ascriptions, the form (or whatever) of the internal state does not matter. This is because, even with a different internal state, overt behaviour will be such that someone can, in all reasonableness, ascribe correct usage to another. The point is that, for the person who systematically mis-classifies 'red' as 'green', all behaviour to do with 'red' will appear normal (i.e., conforming to what everyone would agree as the "correct" behaviour). A further, though obvious, point is that, even if someone mis-classifies like this in a systematic fashion, there is no way for anyone to determine it.

Admittedly, this case is only *logically* possible. Whether it *actually* occurs is another point, but any theory must take it into account. However, the mere possibility that is opened up entails that the internal states of one agent may, in fact, be incomparable with those of another. Thus, the claim that identical states lead to identical behaviour is further weakened.

The second reason is the assumption that because C_1 causes E, nothing else can cause it. It seems perfectly reasonable to say that for any effect E, it is logically possible for there to be many causes which are not equivalent, and which are not identical. This seems to be particularly true for psychological entities such as beliefs. I can make water boil in at least two ways: I can heat it (in a kettle, for example), or I can evacuate the vessel which contains it. Both heating and pressure reduction lead to boiling. Heating is not the same as evacuation: the causes are not, at least at this level of description, identical. What these two methods have in common is that, if you will, they put more space between the water molecules (a liquid becomes a gas when its molecules become less densely packed; another explanation is that the phase change occurs when the mean free path is lengthened). It remains the case that the two methods are different, even though the *explanations* which underpin them may have a lot in common. When a liquid is heated, its molecules acquire additional kinetic energy; when a liquid is contained in a vessel that is being

[12] But note that colour-blindness does not necessarily imply the kind of mis-classification that I have in mind, for many colour-blind people mis-classify in different ways than this.

evacuated, the energy required for molecules to escape from the surface of the liquid is reduced, so they escape to fill the region above. Even with these explanations, the two methods are still not equivalent: heating involves putting in *more* energy; evaculating involves converting energy that is already in the system. Certainly, energy lies at the bottom of the processes, but the fact that there is a concept in common does not make the processes equivalent or even identical[13].

In an analogous fashion, I can come to believe things by different causal routes. I believe that Rajiv Gandhi was murdered last Tuesday. I formed that belief by watching BBC television news, a source which I usually trust. I could have formed it by reading a newspaper, or by being told by someone in person (which is the way I came to believe that the *Challenger* had exploded). The details that will be conveyed in each of these circumstances will be different.

The point I made about *similar* beliefs above relates to a second objection. Consider the actions of an agent, B, subsequent to receiving a message. On what grounds does A, the sender of the message (the agent which uttered the sentence), believe that B is acting in accordance with the content of the message? That is, how does A identify B's behaviour as being a result of receiving that message and not a result of some random factor or of a memory? (Given the above account, this clearly relates to the problem of different internal representations of colours.)

One answer (given by Craig, [3]) depends upon an intuitive notion of similarity. For any two agents, A and B, A believes that B is similar to itself: i.e., that A would behave in a similar (although perhaps not identical) fashion if it had received that message. In other words, A believes that its own behaviour would have been similar in important ways if it (and not B) had been the recipient of the message. This clearly requires A to hold beliefs about itself. When B does not behave in ways that A would expect, A would assert that B had not understood the message, not that B differed from A in significant ways. If I am sitting down to lunch, and I say "Please pass the butter", the butter dish will be handed to me. If the butter dish actually contains margarine, the other person is entitled to say "It's margarine", to which I can reply "Well, pass me the margarine, then" (I am also entitled to think that the hearer is being pedantic, depending on the tone of voice). Although I have mis-classified, and, presumably, hold an incorrect belief as to the contents of the butter dish, the hearer does not *fail* to understand, but can reply in a variety of ways (depending upon disposition, mood, etc.). What *I* am expecting is to have the butter dish passed to me so that I can have some of whatever it contains. Although the reply "It's margarine" is strictly irrelevant to my request (which is merely to *pass me* the dish), it, too, is not meaningless. It is, though, quite reasonable, to refer to the contents of a butter dish as 'butter', even though, on further inspection, it is revealed to contain margarine; what is more, I *assume* that other people will behave or form beliefs in a way similar to me. I assume that the person who hears the noun phrase "butter dish" in that particular context will understand by it the

[13] I wouldn't use the evacuation method to boil water to make tea, though.

same as I do.

Because I make the *assumption* that someone else is the same as me in all important respects (ignoring, for example, fine details of biography), that their physiological and psychological states are roughly the same as mine, and that they use language in roughly the same ways as I do[14], I make the assumption that they will behave in ways roughly similar to me[15].

An objection to this is that one is making an assumption, and possibly a big one. I do not want to argue that such an assumption is reasonable, or even that it is natural (which I believe it to be), but want, instead, to point out that there is another account (due to Heal [8]). The alternative claims that, rather than assuming similarity, the same effect can be achieved by simulation. If A says $\mathcal{M}$ to B, for A to believe that B understands $\mathcal{M}$, A simulates (Heal's term is *replicates*) the possible behaviours of B on receipt of $\mathcal{M}$. In other words, A reflectively infers possible behaviours. Heal argues [8] that this is a different, though not incompatible account. Certainly, inference is often performed when the speaker is not sure how the audience will react to a particular utterance, and reflection upon an audience's actions can take place in order to reassure the speaker that his or her utterance was understood. However, I believe that the amount of knowledge of the audience that the speaker must have in order to engage in replication must be large: I doubt that a replicationist would argue that the response would have to be inferred in all its detail, however. All the same, the inference must start somewhere, and there must be some set of assumptions which serve to give it foundation. If I am walking around a strange town, say one in an English-speaking country other than the UK[16], and I go up to a stranger to ask directions, I have only the barest of information upon which to judge the stranger's responses to my request. The barest of information must include the fact that the stranger is a human being and that the stranger speaks English. If this were not the case, I might ask directions from a life-like statue in a park or a dummy in a shop window. In other words, it appears that the replicationalist strategy reduces to the more obviously functionalist one proposed by Craig, at least in certain circumstances.

In order to give an account of the private aspects of communication, it seems useful to summarise the points made so far.

1. Different people will be involved in different causal chains.

2. There can be different causes for the same or similar effects.

3. Two different people my exhibit the same overt (or external) behaviour, but have different internal states which cause this behaviour.

[14] The hedge is necessary because confusions of language use are commonplace.

[15] Compare this with Wittgenstein's remarks about different forms of life in [16]

[16] It does not really matter that it is an English-speaking country as long as it is one in which any language I speak will be understood by its inhabitants.

4. External or overt behaviour is taken as the yard-stick in determining understanding.

5. The assumption of uniformity: because someone looks similar to me, I will assume that they behave in the same way as me.

Furthermore, if the point needs to be made again, only my internal states are introspectively available to me; even here, not all of my states are available to introspection. It is, of course, not possible to introspect on some sensation in my arm in the same way that I can introspect on other items (such as the meaning of an utterance or on the origin of some of my beliefs)—all that I can do is to know that I *have* that sensation, but I am the only person who can know that fact with certainty.

The argument concerning different causal chains and that different causes can be found for the same or similar effects was intended to show that it is not *necessary* for identical causes of beliefs. The intention was to show that two agents can come to hold similar beliefs even if they have participated in different courses of events. The conclusion I wanted to draw was that behaviour can be reasonably said to be the same even without the assumption of causal identity. The possibility that two agents will behave in similar (or even identical) ways when they posses totally different internal states then follows. The irrelevance of the difference in internal state results from the requirement that the different states are evoked in a *systematic* fashion (and this need not be a merely representational difference, although some might argue it can be reduced to one—perhaps this says more about our understanding of representation than anything else). The immediate outcome of all of this is that there is no guarantee that when A says $\mathcal{M}$ to B, B will thereby enter an internal state that is identical to the one A would enter. Thus, B's observed behaviour does not *necessarily* indicate understanding. In fact, the concept of meaning, at least if approached in this way, seems to fall.

If meaning is taken as that which remains invariant across contexts or situations, as Barwise and Perry take it to be [2], there remain no necessary grounds for saying that B (or A for that matter) can recognise the invariant. The reason for this is that the invariant that A detects may not be detected as such by B. Does this make sense? My answer is that I do not think that it does. Just because A and B participate in different causal chains, and because identical states are not caused in them by a particular event-token, it does not follow that neither is able to detect invariants in the causal field. What *is* a consequence is that the invariants may be *qualitatively* different, but this is a very different matter from claiming that what is an invariant for A cannot a fortiori be an invariant for B—we, the external observers of B, like A, cannot say *what* the invoked state is (presumably, B can), but the claim that there is no such invariant appears absurd.

At present, I consider that the following is inescapable: we cannot know exactly what states another person or agent has. All we have is external observation upon which to base inferences. Furthermore, it seems reasonable to *assume* that things which are sufficiently similar to us will behave in ways that are similar (this, as Craig argues in [3], is not equivalent to the *Argument from Analogy* since it deals with an assumption and not an extremely weak chain of inference). Do internal states matter? That is, are they more than a fiction? It seems unwarranted, on purely introspective grounds, to claim that they are. As has been argued, the states that are purely internal to an agent seem, in a sense, to under-determine the actions which the agent performs. Of course, *any* state will not do in determining behaviour: that was the thrust of the first part of the argument in this section. Although internal states need not (indeed, cannot) be identical, it does not follow that they are irrelevant to the determination of behaviour. The fact that very very different internal states can be posited to account for similar outward behaviour appears to suggest that internal states do not give a neat picture for determining behaviour.

A congenitally blind person may use, for example, colour words with complete fluency, and may give the impression that he or she understands colours (say, the person is giving a radio talk, and the audience does not know that the speaker has been blind from birth). Until it is known that the speaker has never seen colours, someone in the audience may ascribe correct usage and full understanding to the speaker—when the fact of the speaker's blindness becomes known, matters will, in all probability, change considerably. One might then say that the speaker does not "really" understand colour language, or that the speaker has learned to use the words without understanding them fully (as in the case of the carpenter in the last section). It must be admitted, though, that the grounds for ascribing full and complete understaning become much less firm. This does not render the speaker's words unintelligible or meaningless (in a similar way to the words of the diagnostic carpenter). The problem here is that there seems to be *no* way in which someone who has been blind from birth can ever come to a full understanding of, say, 'red'—understanding must, one would like to say, be partial[17]. In this case, one might want to argue that colour words are being used in purely linguistic ways: that the meanings employed are entirely linguistic and do not depend upon having the requisite experiences. It remains the case that the internal state of the blind speaker will be different from the internal states of his or her audience, but that does not entail that what the speaker says is meaningless in any literal sense. By the speaker's correct uses of colour words, the audience must be prepared to grant at least *some* understanding, enough, in fact, to convey some sort of meaning.

The conclusion I think it fair to draw is that meaning is not entirely "in the head" if it is anywhere. That is, the internal states of the speaker do not entirely determine meanings; the internal states of the audience can be viewed in a similar light. As I argued in the last section, context plays a significant role in determining

[17] On a purely truth-conditional account, one would be forced to say that the blind person's utterances were literally meaningless because some of the conjuncts of the truth-condition could not be satisfied—perhaps analysis depends too much on a mentalistic account of truth-conditions.

significance. The general context in which a speaker is situated also plays a part. Carpenters, in our society, are not often trailed medical practitioners[18], so we do not, in the normal run of things, place great store by their diagnostic pronouncements. This suggests, amongst other things, that conventions play a part in our ascriptions of meaning (this is implicit in the argument of the last section). This should not, though, be considered as an argument for rule-following and meaning, because, if it need be said, conventions do not have to be explicitly stated as rules, even if an analysis of them is in terms of rules. To understand an utterance, I do not consciously need to apply a rule: I may behave *as if* I were applying one, but that is a different matter, and a subject for another time.

6 Agents, Meanings and Messages

The conclusion of the argument so far is that a great variety of factors influence ascriptions of meaning and understanding. Some of these factors are contextual, some conventional. The argument of the last section showed, I believe, that what is in the head of an agent does not uniquely determine meaning. Because of this, the final arbiter of meaning cannot be said to reside uniquely with an agent, be it human or otherwise.

In this concluding section, I want very briefly to turn the arguments of the last two sections back onto the problem I began with: how can messages exchanged between agents in a DAI system be said to have meanings? In the context of traditional, sequential, centralised AI (what I often refer to as "solipsistic" AI because the agents—programs—exist in a universe in which only they count and in which their interactions with an external world are, to say the least, marginal), the issue does not immediately arise. This is because the observer/designer/experimenter is always at hand to provide the agent with meanings. In a DAI system, on the other hand, the observer cannot be onmiscient: the experimenter/designer/observer cannot inspect all internal states at the same time, and it is assumed that DAI agents operate concurrently, possibly at widely separated spatial locations (very much in the way in which people do). In other words, the external human agent cannot be on hand to give each agent in the DAI system the meanings it requires as and when it requires them.

A model-theoretic account of meaning for a DAI system would, or so it seems at first sight, require that two agents possess the same model in order to understand the meaning of a message. This entails that they must be in identical (or, at least, provably equivalent) states. By the arguments of the last section, this appears to fly in the face of the evidence: it is far too strong a requirement, and it also impacts upon the causal chains in which any two agents participate. On a model-theoretic account, the final arbiter of meaning is the model: this, as has been argued above, cannot be, for it appears that there is no *final* arbiter in such an unambiguous and definite

[18] They might have been more common in post-Cultural Revolution China, but that is not here.

a sense. The current proposal is very much messier than one which proponents of model theory would like to provide. It is messier because it posits information from a large number of sources as assisting in the meaning-determination process, and it is also messier in the sense that no one single agent in a DAI system can be said uniquely to determine *the* meaning of a particular utterance. In order for agents to be able to ascribe meaning, they must, amongst many other things, have access to their context, to the context of the other agent, to expectations and to conventions about use; they need to know about their social roles and what that implies for them. It should not come as a surprise to find that what agents need to know, or to have access to, or to reason about is somewhat large and complicated, for DAI is, at least in part, about systems that act in a much more social way than do the agents in solipsistic AI systems.

References

[1] Austin, J.L., *How To Do Things with Words*, Oxford University Press, 1962.

[2] Barwise, J. and Perry, J., *Situations and Attitudes*, Bradford Books, MIT Press, 1983.

[3] Craig, E., Privacy and rule-following in Butterfield, J., *Language, Mind and Logic*, pp. 169-186, Cambridge University Press, 1986.

[4] Craig, I.D., *The CASSANDRA Architecture*, Ellis Horwood, Chichester, 1988.

[5] Craig, I.D., *Extending CASSANDRA*, Research Report No. 183, Department of Computer Science, University of Warwick, March, 1991.

[6] Fox, M.S., An Organizational View of Distributed Systems, in *Readings in Distributed Artificial Intelligence*, Bond, A.H. and Gasser, L. (eds.), pp. 140-150, Morgan Kaufmann, 1988.

[7] Gasser, L., Social Conceptions of Knowledge and Action: DAI Foundations and Open Systems Semantics, *Artificial Intelligence*, Vol. 47, pp. 107-138, 1991.

[8] Heal, J., Replication and functionalism, in Butterfield, J., *Language, Mind and Logic*, pp. 135-150, Cambridge University Press, 1986.

[9] Kornfield, W.A. and Hewitt, C., The Scientific Community Metaphor, in *Readings in Distributed Artificial Intelligence*, Bond, A.H. and Gasser, L. (eds.), pp. 311-320, Morgan Kaufmann, 1988.

[10] Nilsson, N.J., Logic and Artificial Intelligence, *Artificial Intelligence*, Vol. 47, pp. 31-56, 1991.

[11] Searle, J.R., *Speech Acts*, Cambridge University Press, 1969.

[12] Smith, B.C., *The Correspondence Continuum*, Report CSLI-87-71, CSLI, Stanford University, 1987.

[13] Smith, B.C., *The Semantics of Clocks*, Report CSLI-87-75, CSLI, Stanford University, 1987.

[14] Smith, B.C., The Owl and the Electric Encyclopedia, *Artificial Intelligence*, Vol. 47, pp. 251-288, 1991.

[15] Suchman, L.A., *Plans and Situated Actions*, Cambridge University Press, 1987.

[16] Wittgenstein, L., *Philosophical Investigations*, trans. G. E. M. Anscombe, Blackwell, Oxford, 1958.