Agents That Model Themselves

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

March 1994

Abstract

In this paper, we adopt a symbolic interactionist perspective on Multi-Agent Systems. We investigate the uses and genesis of agent models, both Self and acquaintance models and argue that the symbolic interactionist view is superior to the *ad hoc*, static and somewhat low-level approach adopted at present. We consider agent models—both Self and acquaintance models—as devices for making predictions about and the formation of explanations of agents' actions. We then argue that Self models are prior to and instrumental in, because of perspective and because of role taking, the construction of acquaintance models. We also argue for our position that social structure is dynamic and that socially organised agents must be dynamic in their nature.

1 Introduction

In this paper we attempt to show how the symbolic interactionist conception of the Self can be used in Multi-Agent Systems as the basis for complex behaviour based in part upon models of other agents. We also aim to show that a model of the Self that is conceived in symbolic interactionist terms can be used in the generation of agents' acquaintance models. The term *acquaintance* refers here to the other agents known to a given agent. When an agent has acquaintances, it is able to interact with them, and it has knowledge of and about them. (The acquaintance relation is usually symmetric: if agent A is an acquaintance of agent B, agent B is an acquaintance of agent A.).

We will argue that an agent's model of the Self is prior to its models of its acquaintances and is used in the generation of acquaintance models when the right cognitive processes and structures are activated. We also argue that models of various kinds can be used in prediction and explanation. These uses of models are constrained: models are not used for the entire span of an interaction as is the case in some current Multi- Agent Systems, a property which makes the possession of dynamic models plausible. The Self is further developed through interaction.

Our conception of the Self is that it is essentially dynamic, evolving as a consequence of meaningful interaction with others and with the environment. The Self is founded on privileged information that is private to the individual. In a sense, this information is prior to all other because it composes the individual. This privileged information is composed of an agent's knowledge, beliefs, dispositions, motives, goals and the contents of other mental states and processes. These forms of information are privileged in the sense that they relate to an agent's internal mental (and perhaps physiological) processes: they cannot be *directly* inspected by anyone other than the agent. Other agents must infer them or rely on being told by the agent. More central to our purposes here, this privileged information is deployed by an agent in forming the perspectives upon which are based its observations and interactions.

Without internal information about itself, perspectives cannot be constructed, and, without perspectives, views of others cannot be formed. In the sense that perspectives are the basis upon which models of others are formed, it can be seen that models of the Self are prior to models of others. This is because we always view the world from a particular perspective (which may change from situation to situation and from time to time): perspectives depend upon how we are feeling, how we are disposed, what we know, what we believe and upon our past history—perspectives depend, that is, on privileged, internally derived information about ourselves that is prior to information about others because it is used in forming beliefs about others.

The Self is dynamic and constantly evolves: evolution occurs as a result of an agent's interaction with the external world, but also as a result of interaction with itself and its changing internal processes, dispositions and states. Privileged information changes with time—is itself dynamic—and contributes to the development in time of the Self.

We must be clear about the kind of Multi-Agent System of particular interest to us. Most Multi-Agent Systems are composed of a stable collection of agents: in these systems, it is not typically the case that agents come into and go out of existence, their tasks are highly circumscribed and determined *a priori*. We are interested in systems which can exhibit such dynamic organisation. The collections of agents that interest us are such that all agents do not necessarily possess exactly the same knowledge: they are heterogeneous when viewed in terms of what they know (although there may be certain kinds of common knowledge that is possessed by all agents—knowledge of social conduct and rules for effective and socially acceptable communication, for example).

Because Multi-Agent Systems are naturally distributed, we may not assume that all agents have the same perspectives with respect to the environment, with respect to the other agents in the system, or with respect to the tasks they may perform.

The fact that agents necessarily have different perspectives is an important point as far as this paper is concerned. Two agents have different perspectives as a result of:

- Their location in the environment.
- The other agents with which they communicate (or more generally: the information available to them from various sources).
- The individual histories of the agents (we want agents to develop over time—is an essential consequence of a symbolic interactionist analysis).
- The specific knowledge possessed by agents as a result of their identities and personal history.
- The tasks, goals and other intentional structures of an agent and their ability to execute actions.
- The agent's motives and drives.

An agent's perspectives are based upon a variety of factors some of which depend upon physical properties, some based upon cognitive factors, and others that depend upon the agent's history (history viewed in terms not only of the tasks the agent has attempted in the past, but also the episodes in which it participates and the events it witnesses, as well as the factors that affect its motives).

The fact that no two agents have exactly the same perspectives has far- reaching consequences for Multi-Agent Systems and for Self models. For example, two agents may tend to interpret the events they witness in different ways: if it is necessary for them to act in response to the events they observe, agents may not necessarily agree on what the best course of action: some form of negotiation may be necessary. Difference of perspective is that no two agents will necessarily behave in the same way after witnessing an event or participating in an episode: if the perspectives are the same, future development and behaviour will be the same because agents' interpretations of the event or episode will be the same—the agents will derive identical information from an event or episode; a simple argument shows that perspectives must necessarily differ.

We expect agents to change with time: indeed, this is *required* by symbolic interactionist theory. Consequently, an agent's experiences and the outcomes of its previous actions determine its future behaviour: experience shapes the formation of the various, context-dependent Selves an agent exhibits. The interactions in which an agent has participated will also shape future behaviour. This contrasts with many Multi-Agent Systems which contain agents that are static over time (our CASSANDRA [3] is an example of an architecture for building such static systems).

Under this static interpretation, agents are seen as problem solvers that cooperate in the solution of larger problems; communication is seen as a necessary evil forced upon agents by their distributed and (semi-) autonomous natures. If agents are problem solvers with added communication (an interpretation very similar to our own early view of CASSANDRA agents—see [3], Chapter Four), their most important components deal with the problem-solving process. In the conventional view of problem-solving agents (one which is almost a tradition in AI), an agent is equipped with all knowledge necessary to solve a given repertoire of problems: agent development can proceed in a completely off-line fashion (in the sense that the agent is performing no actions related to interaction or to problem solving[1]—we have an second use for the adjective 'off-line' which will become clear in the penultimate section). At runtime, agents do not and need not change in any way.

Our view of agents is that they are highly dynamic. They change as a result of interaction with other agents and with the environment. Agents also change as a result of internal cognitive processes (for example, forming beliefs as a result of reflecting on external or internal events—we will consider these processes explicitly below). We view agents as firstly agents belonging to and forming a social structure, secondly as actors within social situations and finally as problem solvers: problem solving is important but not everything. Consequently, considerable activity will be devoted to action within a social structure and to the formation and maintenance of social order. Even while engaged in problem-solving activity, agents interact with others and these interactions cause agents to change—we will return to this important point below.

A consequence of the dynamic nature of the Self is that an agent's acquaintance models must change with time. On a common-sense level, we all accept that other people will be *roughly* the same the next time we meet with them as they were last time. There may be slight variations: larger variations may only detected after longer of separation (if one has not met the other person for some considerable time, differences between the way they were and the way they are now may be much clearer).

We are nonetheless able to adapt dynamically to any changes in the behaviour of an acquaintance (we infer their current attitudes and affective state on the basis of such perceived changes). Once an acquaintance has behaved in a manner we might consider atypical or merely different, we might remember the change in demeanour for future reference and might adjust our behaviour on subsequent occasions to take account of it. We are able also to make inferences about the conduct of other people and use these inferences in adjusting our own behaviour. For example, if one has hitherto had cordial relations with someone and then that other person becomes aggressive and unpleasant, one will remember the change in behaviour and will in future act in a different fashion towards that other person.

As a result of the change in behaviour and imputed knowledge of the other person, we might attempt to reason about potential causes of the change and use the results of reasoning as an explanation for it. Such an explanation can be used to

[1] What we really mean here is that the agents are 'developed' or modified by their programmers who change (edit) their database or mechanisms by hand.

guide our future behaviour both with respect to the person whose behaviour we have found to be different and also towards others who may deal with that person (for example, warning them of the altered disposition). This example indicates that our models of other people are dynamic and subject to change given sufficient grounds: we should expect the acquaintance models of an agent in a Multi-Agent System to be equally dynamic and context-sensitive.

The reader will notice that the arguments below deal with the concepts of action and behaviour, and propose structures that are entirely cognitive—such as mental models and belief systems—as theoretical constructs that are 'internal' to agents. On the one hand, we have overt, observable entities, and, on the other, private entities whose existence we can only infer. This is, however, entirely in keeping with symbolic interactionism as Mead envisaged it. Mead was influenced by behaviourism, both negatively and positively. The positive influence was that "[h]umans should be understood in terms of their behaviour, not who they are," [2], p. 28). The negative influence was that the narrow conception of behaviour held by Watson (and later Skinner) was, Mead believed, inadequate to explain complex, particularly social, behaviour: Mead insisted that people's actions be understood in terms of definition, interpretation and meaning—clearly, these are cognitive and intentional terms. Symbolic interactionism is a theoretical perspective that encourages the analysis afforded by cognitive theories in the explanation of social and interpersonal behaviour. This aspect of symbolic interaction, we believe, makes it an ideal framework for theories about Multi-Agent Systems.

1.1 Organisation

This paper is organised as follows. In the next section, we present the central concept of symbolic interactionism—the interaction—in brief outline and also present some properties of the Self from this perspective. Next, in Section 3, we consider first the concept of the mental model and some of its properties, then we describe the concept of the acquaintance model and mention some of its uses in Multi-Agent Systems. While considering acquaintance models, we will mention issues that we explore in more detail later. In this paper, we are concerned with the concept of the Self and its role in autonomous agents: we address issues raised by the symbolic interactionist concept of Self in Section 3.3. Role taking is a central concept in symbolic interactionism, one which we examine in Section 4: we will describe its use as a device for predicting and explaining behaviour. In Section 5, we argue that models of the Self are prior to models of acquaintances in the sense that Self models can be used in the generation, modification and application of acquaintance models. It can be argued that the kinds of model that we propose in this paper are far too complex to be of real use in practical settings: they appear to be too dynamic and subject to frequent change, properties which imply high computational cost. We argue (Section 6) that the conceptual framework within which such claims about cost and complexity are couched are mistaken, and argue that people are able to manipulate dynamic structures with ease and at low cognitive (and computational) cost, provided the structures are processed at cognitively and socially appropriate times. Finally, we summarise our arguments and mention some issues not raised in this paper (Section 7).

1.2 Acknowledgements

We would like to acknowledge a snow storm which enabled the timely completion of this paper. As is usually the case, this work benefited in no way whatsoever from support or encouragement from the Department of Computer Science, University of Warwick, or any of its academic staff, nor from the University of Warwick as usual.

2 Symbolic Interactionism

We adopt symbolic interactionism [19, 14, 2] as the general framework within which we will discuss the structure and use of agents' models of themselves and of others. Symbolic interactionism is a theory which concentrates on the *interactions* between individuals. It describes individuals as constantly evolving symbol-using agents. Individuals are *not* stable and structured unitary personalities. Instead, each agent has many selves that are context- dependent: the selves are themselves constantly evolving. They evolve as a result of dynamic selection and interpretation (in terms of *symbolic* structures) of stimuli that are derived from interaction with other individuals and with the environment.

The symbolic interactionist view of the Self contrasts strongly with the common-sense view that frequently informs the Self models of agents in most current Multi-Agent Systems. Common sense tells us that there is a unique thing which is ourself: this thing persists through time, is associated with the memories that we call our own, and appears to form the basis of our claims of identity. Recent work in cognitive psychology, for example [26], has strongly denied the common-sense view of the self: very powerful arguments against the unitary self have been developed. The Self is now seen as an unstable construct that develops over time. Symbolic interactionism, as has been mentioned, proposes a similar view of the Self, but places the origins of the self in the interactions in which the agent participates and which the agent interprets. Seen in these terms, the Self becomes a collection of context- sensitive processes or perspectives which respond to and interpret the various situations in which an agent is located.

3 Mental Models, Acquaintance and Self Models

3.1 Mental and Agent Models

It is necessary, particularly within a symbolic interactionist framework, to examine the role of agent models: models that an agent has of itself and of its acquaintances.

The reasons for engaging in such modelling are often taken for granted in the Multi-Agent Systems literature. For example, the descriptions of the acquaintance and Self models in ARCHON [27], there is little justification of their use: from the text, we are expected to provide the justification for ourselves. Initially, we will concentrate on the role of acquaintance models; then we will consider the self model. Discussion of the self model will entail consideration of the role of the Self as a symbolic interactionist construct.

Both acquaintance and Self models can considered a species of mental model [16, 21, 17]. Norman [21] says that they are:

- They are parsimonious.
- They are incomplete.
- They lack firm boundaries.
- They can be used in prediction and explanation.

Mental models are parsimonious because they are simpler than the object or process they represent. This simplicity comes at the price of incompleteness. Because mental models do not contain all the information a person has about an object, individual, task or process, they are incomplete and cannot be used in all circumstances. It may be the case, though, that someone has multiple, overlapping mental models that relate to the same object: each model will contain different knowledge or information, but there will be overlaps—thus, models do not have firm boundaries. Finally, mental models have predictive and explanatory power: it will be seen these issues are of particular importance.

3.2 Acquaintance Models

Acquaintance models are necessarily simpler than the acquaintances they describe. They will also be incomplete. This clearly derives from simple analysis and it can be derived from a symbolic interactionist view of our knowledge of others. In general, we may know a lot about another person, but we do not use all of that knowledge in forming a model of them. We select what appears to be the most relevant information (which may be based upon situational factors as well as prejudices and so on): one's *perspective*[2] is important in forming a model of another. From a symbolic interactionism viewpoint, acquaintance models will be incomplete because the self is viewed as a dynamic construct which is always being constructed from novel interactions. It is therefore impossible that anyone can have *complete* knowledge of another.

The use of mental models in prediction and explanation is of considerable interest in the context of Multi-Agent Systems. Acquaintance models have a variety of uses in Multi-Agent Systems.

[2] Charon [2] emphasises the perspective as a concept important to symbolic interactionism.

In some contemporary systems, models are used to support reasoning about acts of communication: interaction is planned on the basis of the model of the agent who will be the other party to the communication. Before this can be done, the other party (or parties in some systems) must be identified: this is done by consulting available acquaintance models to determine which acquaintances are the most appropriate participants. The content and structure of interactions are also planned on the basis of the contents of the acquaintance model. Acquaintance models are consulted during interaction as well as during the interaction planning stage: this, as will become clear, is not very plausible.

We have a different view of the use of acquaintance models. In common with some contemporary systems (e.g., MACE [10, 11]), we see them as containing information about other agents. Such information might include:

- The areas of expertise of the acquaintance.
- The tasks which the acquaintance is likely to work on.
- The goals, motives and other intentions the acquaintance is likely to have. This might also include information about the likelihood of engaging in co-operative activities with the acquaintance and information about the acquaintance's reliability in various contexts.
- Information about the acquaintance's progress (or assumed progress) on certain tasks (say, tasks which it is performing in co- operation with or at the request of the agent which holds its model).
- Expectations about the acquaintance's behaviour in certain contexts (this might include expectations about things promised by the acquaintance).

This information can certainly be used in planning interactions and in *role taking* (see Section Four below). Acquaintance models can also be used in prediction and explanation as well as general reasoning about the acquaintance: we will briefly explore these now.

Acquaintance models contain information about the assumed expertise, attitudes, dispositions, etc. of other agents. They are *theories* of other agents. During interaction, if an acquaintance communicates something or behaves in a way that is not readily comprehensible or that is unusual, the acquaintance model can be used to explain the anomaly. If the explanation is adequate (and this may require subsequent investigation), the acquaintance model can be updated to reflect the newly acquired information. An explanation might relate to general or background information; it might relate to previous information about the acquaintance, or it might be derived by analogy with the behaviour of another acquaintance. In the last case, an agent's knowledge of more than one agent becomes important: we must, in any case, always remember that each agent is a member of a community of interacting agents.

The process of explanation need not be restricted to the case of anomalous or unexpected behaviour on the part of one acquaintance: it can be applied (although it will be more complex) in cases in which more than one acquaintance's behaviour needs explanation. For example, an agent might be informed about the joint actions of two or more acquaintances, but may not know the details or the purpose of the joint activity. The agent may engage in the process of constructing an explanation in order to determine what the other agents are doing; such an explanation may have some relevance to the agent's own current behaviour or to future actions it intends to take. Explanation of this kind may involve assessment as to the likely threat posed, or benefit incurred, by the acquaintances' actions, or the importance of their actions to the agent constructing the explanation. Explanations of this kind may also (will probably) involve reasoning about the motives behind and about the aims of the acquaintances' joint actions—such reasoning may also be applied to the actions of a single acquaintance, it is clear. The results of this process can be added to the acquaintance models in question and lead to an enrichment of the models[3].

It is, of course, the case that such explanations may prove incorrect: it must be remembered that agents operate in the context of information that is generally incomplete or inexact. In such circumstances, the results of the explanation process may need to be revised in the light of future developments. This is a very important point: in the kind of Multi-Agent System we are considering, acquaintance models are always in a state of revision and reconstruction: they can never be perfect. Explanations are a kind of empirical theory of limited extent: as such, they are subject to revision in the light of better information. (Even when an explanation is incorrect, it may still be useful for an agent to remember it because it might, at some later time, form the basis of a correct explanation of some other behaviour or interaction. We will return to this point below.)

There is a use of explanation that we have not considered. It is the use of explanation to determine the current state, actions or dispositions of an acquaintance based upon the interactions between the agent and the acquaintance. When interaction occurs, it may be necessary for both parties to explain the content of the messages which are exchanged. This is because interaction will assume background information that is common to both parties. Furthermore, interactions may contain implicit information and inferences may be made about the structure and content of the interaction: inference may reveal information that is not present in the original communication. We have argued elsewhere [4, 5] that the very fact that one agent has engaged in communication with an acquaintance can be a significant event for the receiver of the message. Given the content and the kind of interaction (the speech act performed in the communication), inferences can be made about

[3]We view acquaintance models as having a network structure based upon a generalisation hierarchy that is rooted in representations of central concepts, event types, etc. Explanations and other information are assumed to be stored in various parts of the hierarchy in such a way that they can be accessed in different contexts. The model we have is similar, although not identical, to a multiple inheritance lattice.

the other person. For example, if the communication is based upon a factual question, the receiver can infer that the sender is ignorant of something. Information derived from inferences about the content and form of interactions can be used in the enrichment of the acquaintance model.

Mental models can also be used to make predictions about their content. Acquaintance models can be used in exactly the same way. However, within a symbolic interactionist framework, matters are not quite as simple as using an acquaintance model to make predictions about what the acquaintance will do or how it will react in certain circumstances. It is necessary for the agent making the predictions to view things from the acquaintance's *perspective*: to a great extent, acquaintance models are about perspectives. How acquaintance models are used in prediction must wait until we consider the process of *role taking*.

3.3 Self Models and The Self

In some Multi-Agent Systems, agents are provided with models of themselves. These models contain descriptions of the kinds of task the agent can work on, the areas of its expertise, and the goals it has. In some cases, for example the ARCHON system [27], the Self model contains information about the low-level primitive operations the agent can perform. The Self model is used by the agent to reason about its own capacities and capabilities as well as about the tasks that the agent can perform: in other words, Self models support a form of reflective or introspective processing.

In many respects, models of this kind give rise to behaviour similar to that exhibited by one use of our ELEKTRA system [7, 8]. ELEKTRA is a production rule interpreter that permits rich forms of reflective behaviour, and allows rules to access system primitives. Such access conveys upon rules the ability to reason about the primitives which implement the system, as well as using them to perform reflection (by engaging in reflection about control, ELEKTRA can be used to describe and implement its own forward-chaining interpreter, see [8] for more details). One central feature of control reflection in ELEKTRA is that it is at a relatively low level, and is based upon the implementation of the interpreter.

The ARCHON agent self model [27] appears to be at a level similar to that of ELEKTRA. The ARCHON concept of Self model is static. ARCHON agents appear not to change with time (for one thing, they are intended to be completely domain-independent—problem- and domain- specific problem-solving components can be slotted into the agent structure without change to the agent) and their behaviour does not appear to change greatly as a result of interaction.

The kind of agent we have in mind contrasts strongly with this. For us, agents change as a result of interaction with other agents, with the environment, and with time. Agents learn more about themselves, other agents, about the environment as they interact; agents learn more about their acquaintances through interaction and the use of acquaintance models and by the process of role taking (which tells them about themselves). We also see the Self model as a construct an agent uses to make sense of the world. Because an agent will view the world from different perspectives, its Self will change from situation to situation according in various ways.

The Self in symbolic interactionism is constructed as a result of interaction with others: in particular, significant others and reference groups. The latter form is a source of a perspective which shapes the development of the Self and is clearly related to the concept of an agent type or to the conception of some agents as being in some way similar (either in structural, social or teleological terms). The reference group might be a political or social group, members of a club, former students of a particular Oxford college, and so on. Each reference group is associated with a particular perspective on the world: that is, with a particular way of interpreting the world of interaction.

The Self is constructed (at least in part) as a result of the content and form of interaction with these others. Although the symbolic interactionist concept of the self is dynamic in the sense that every interaction causes the Self to alter, it is not necessarily as dynamic as one might think. Once the self has been constructed, it remains fundamentally the same even though small changes will occur (such changes might be viewed more as adjustments); it is possible for a personality to undergo a spectacular change, of course (for example, political or religious conversion).

The important point to remember at this point is that *the self is a social construct and is developed as a result of social interaction.*

For one of our purposes, the Self is the construct that allows introspection to take place. In symbolic interactionist theory, the agent interacts—using symbols—with its Self. This allows the agent to determine which courses of action it will follow. Blumer [1] states that:

> "The possession of the self provides the human being with a mechanism of self-interaction with which to meet the world—a mechanism that is used in forming and guiding his conduct." (p. 535)

Mead [19] remarks that:

> "the essence of the self ... is cognitive: it lies in the internalized conversation of gestures which constitutes thinking, or in terms of which thinking or reflection proceeds." (p. 173)

The Self is a mechanism by which an agent can understand what it communicates and what is communicated to it. Mead [19] explains this thus:

> "What is essential to communication is that the symbol should arouse in one's self what it arouses in the other individual." (p. 149)

This conception of communication with the self is similar to the concepts of reflection and introspection that are implicit in the self models in Multi-Agent Systems and elsewhere. The Self model in this context is used to determine what the agent should do next: for example, the Self model might be employed to determine which step of a plan should be executed next. This situation may occur because the agent has reached an impasse in its activity and needs to deliberate on what next to do. (It is interesting, here, to note that this view suggests that the creation of a plan- –'plan' interpreted in the classical fashion—is an act of deliberation, an interesting point that is different from the usual one encountered in the planning literature—see below.)

If one adopts a different terminology, for example with 'reflection' being understood as having externally derived information as its object and with 'introspection' having internally-generated information as its object, the conception of the Self as defined by symbolic interactionism provides a unifying concept with which to understand the two processes. Reflection (in the above sense) is the *act* of understanding an external interaction. Introspection (in the above sense) is the act of understanding an internal interaction. The symbolic interactionist concept of the self, then, provides a framework for engaging in reflective behaviour in a way familiar from existing research into Multi-Agent Systems and programs that perform meta-level reasoning (such as our ELEKTRA).

We do believe, however, that the kinds of information and knowledge which are manipulated during reflection (in the conventional, all- encompassing sense) in current Multi-Agent and reflective systems are at far too low a level and that the models are too complex to permit easy and efficient use. The level at which conventional Self models are couched relate to the implementation (this is the case, at least for some kinds of process, in our ELEKTRA rule interpreter [7, 8]). Such low-level models are psychologically implausible, however. We are not privy to the fundamental psychological and physiological workings of our minds, brains and bodies. Instead, we have a relatively high-level view of ourselves in terms of our beliefs, desires, goals, memories of others and of episodes, etc. But access to the primitives that implement an agent is close to the conception of a Self model in many Multi-Agent Systems. Equally, although we may infer some pathology (mental or physical) in others and we can make inferences (provided we have adequate knowledge) about others in terms of their appearance and behaviour (including verbal reports), we generally think of other people in the high-level terms mentioned above. For example, we may think of someone as bombastic, self-centred, self-important, willing only to promote selfish aims, ignorant or power-crazed, but it is less likely that we will think of someone in terms of how they are encoding memory traces, and so on. In the case of our views of ourselves, we certainly have privileged access to our states (for example pain) and we may prefer not to make some thoughts or beliefs public, but we do not have low-level access to many of the processes (e.g., procedures that access the mental lexicon). This property of the human mind has been called 'cognitive impenetrability' [22].

We can reason about ourselves and about others in terms of high-level concepts. Our reasoning about others and about ourselves appears often to be relatively speedy but can be wildly incorrect (for example, we may refuse to accept certain inferences or facts that pertain to ourselves, to our best friend or to our partner). If the models that we have of ourselves and others were massively detailed (and as mental models, they appear not to be), they would be difficult to use in reasoning. As mental models, they appear to be relatively simple; the fact that reasoning with them is rapid seems to imply that they contain information in a readily manipulated form and that the amount of information is relatively small.

A further criticism of the kind of self model found in conventional systems is that they do not change with time. The self, in this case, is merely a snapshot. Indeed, one can argue that the self model is not really a model of an agent's self: it is a term that has been misappropriated and which really means a device to allow reasoning about the static structure of a static agent. As we have seen, the symbolic interactionist conception of the Self is dynamic. The models that we propose of the Self must be dynamic because our agents operate in dynamic environments. Furthermore, since agents interact with each other over time, dynamic models are of utmost importance because they must represent the most up-to-date view of acquaintances and of the Self. The Self changes as a result of interaction with acquaintances and with its own agent. Interaction is itself a dynamic process. Consequently, agent models of all kinds must be dynamic because they are required to represent the latest information so that agents can act and interact in ways appropriate to the current situation (which changes with time and as a result of interaction between agents and with the environment).

If models are static in the sense of containing unchanging information, agents are unable to represent the current situation in terms of changes in themselves and changes to the environment and other agents. One role for the Self is to direct behaviour. Unless the Self is able to adapt to changes in its agent's position within the social groups in which it participates and to changes within the agent itself (as a result of interaction in all its forms), it cannot direct behaviour. This is because the Self relates to circumstances which no longer obtain. Thus, unchanging Self models cannot be used to direct the agent's actions in a dynamic context.

(It might be possible to argue that under some version of behaviourism, static Self models might be suitable for directing actions, but it might turn out that simplistic, static models only fully perform their function in static—or relatively static—environments. In particular, they may apply only to low-level behaviours and only in agent organisations in which the agents' roles do not change: such organisations are typical of the current generation Multi-Agent System.)

We believe that part of what it means to have a 'Self' is to change with time and with situation. In other words, we believe that part of the concept of 'Self' is just those properties of adaptation and development.

We believe that models of the Self are composed of high-level structures such as goals, intentions, beliefs about others and about the Self, as well as descriptions of actions. Such models will change as a result of interaction with others and as a result of learning and training. The beliefs that an agent holds about itself at any time will guide its actions and will assist in the formation of its goals (equally, intentions will assist in belief formation). The beliefs held by an agent will alter with time and as a result of interaction: beliefs are formed as a result of observation and of understanding what others do and say. Explanation of its own and other agents' behaviour also contributes to the development of the Self, as do prediction (and the consequent production of expectations, [5]) and role taking. We will discuss both prediction and explanation in some detail below.

It may seem surprising that explanation of one's own behaviour is included in the list of processes that contribute to the development and change of Self models. There are two reasons for such explanation: they can be created to determine the reasons for the success of an action, and they can be performed in order to determine the reasons for the failure of an action[4]. In both cases, the context in which the action was performed is important, as are the original reasons for performing the action. The more important of these two cases would appear to be that in which an action fails. Schank [24] argues for the importance of failure in learning: we learn more from failure than success. One reason for this is that when an action fails, it shows that either our view of the situation is flawed in some way (some predicted event or state of affairs may be at fault) or that we do not understand the action or the situation in the way we thought when deciding to perform the failed action. By determining the causes or reasons for failure, we learn about the world and the factors which constrain actions.

When we decide to perform an action, we have a goal in mind, as well as a view of what the world will be like when the action is to be performed; we also have an idea of some of the effects of the action[5], typically the effects which make the performance of action desirable as far as we are concerned (the effects which satisfy our goals). When an action fails to produce the desired or expected result, we need to explain that fact so that we can try to avoid making the same mistake again. The explanation process may involve reasoning about causality, other agents, and so on. Schank [25] proposes a theory of explanation and learning which is relevant in this connection. The result of the explanation process will be new information about the world and about the actions one can perform in given situations—a new theory, in fact.

An explanation may be faulty because it does not generalise or that it is not the 'true' reason for the action's failure (for instance because the information available to the agent during the explanation process is incomplete or incorrect). 'Incorrect' explanation is a problem for all empirical and heuristic methods: it is not necessarily a great problem to an agent. Even though an agent may not have the correct or true reasons for determining that an action should not be performed in a given situation

[4] A third reason would be to explain why one has changed one's mind about the viability of an action just as one was about to perform it.

[5] We may not be able to say all the effects because of the frame problem. It is possible that the frame problem can be ignored.

(or class of situation), the fact that the agent has determined *some* reasons for avoiding the action in future is enough: the agent will avoid such actions when similar situations occur later. Subsequently, the agent may discover a better, more accurate or deeper reason for avoiding the offending action, but it is the behavioural modification of the agent that is the important point.

Explanation of failed actions serves to alter behaviour by inhibition. Explanation of successfully performed action serves as a reinforcer. Successful performance does nothing to increase the information or knowledge possessed by an agent. Failure causes the agent to change in both its behaviour and its knowledge.

It is interesting to note that an action that hitherto has been successfully performed, but which suddenly fails. In this case, knowledge of previous success is important to the explanation process because the factors which led to past success will be important in determining what has now changed. Agents cannot ignore success in favour of failure.

The Self model is important in producing explanations of failed actions. This is for a variety of reasons. Firstly, as we have seen, the Self is the object which directs behaviour. Second, selfhood means that one is able to *see oneself in the situation of interest* (the Self is present in all situations): the Self "makes judgements and subsequent plans of action toward the many other objects that appear in each situation" [15], p. 43). Charon rephrases this point as follows:

> "Selfhood allows us to examine situations and how they affect us and to determine the kind of action we might take by imaginatively testing proposed action first on its effects on the self, that object we seem to know best. When I engage in conversation I engage in a self-interaction that attempts to assess the other's image of me and how I am acting in relation to the other." ([2], p. 77)

The Self acts on the basis of knowledge or information that determine the kind of action to be taken. This implies that the Self has information that can also be used in explaining failure and success: part of this is derived from the process of putting oneself in the position of another agent (assuming the other agent's perspective or role taking—we will discuss this in more detail below).

Third, we believe that part of what it is to have a Self is having a memory of past events. To common sense, memory seems to be one of the defining factors of personal identity: without memory of past events, actions, feelings, etc., there is no concept of the continuity of the Self through time. Memory records the events one has participated in and the actions one has performed, together with the reasons for and outcomes of those actions: these can be retrieved (with greater or lesser accuracy) from memory and used for explanation. Memories of past failures and previous explanations of failure can be employed in current explanations. The use of past events and actions in explanation links previous to current experience; it emphasises the continuity of experience that we identify with the our selfhood.

Previous experience can guide current action and also assist the process of explanation by relating current action to previous events, actions and explanations. In the opposite direction (present to past), a new explanation may cast light on a previous episode that could not be previously explained or which was incompletely or incorrectly accounted for. The fact that the Self changes with time and experience implies that in the future it may have more information to apply than in the past. By re-examining previous events, a new perspective or new information can be applied in the explanation of previous episodes: this may lead to a different explanation or a different attitude to (perspective on) previous events—it leads to the formation of a new theory about how one acted in the past. This new theory can be adopted in the present to determine behaviour both in the present and in the future. This theory revision process leads, therefore, to an alteration to the Self: the Self is seen from a different perspective to be dynamic. We will return to the construction of explanation below when we consider the symbolic interactionist concept of role taking.

The Self can also be used as a basis for prediction. In the quotation from Charon ([2], p. 77) there is the explicit notion of trying out actions in imagination. The symbolic interactionist concept of the Self is of interaction between the agent and its conception of itself: the Self is seen as an object with which interaction occurs. When considering a course of action, the agent imagines what will happen when the action is actually performed. Part of this imaginative simulation is the prediction of the effects of the action on the environment and on other agents. Prediction of events allows the agent to determine whether a proposed action will achieve the desired effects when it is performed. This leads to an expectation which can be supported or refuted by actual circumstance.

The predictive use of Self models relates to the process of role taking to which we now turn. As will be seen, role taking is important in the development of the Self as well as in interaction with others.

4 Role Taking

The agents of a Multi-Agent System need to co-ordinate their actions, co- operate and compete in performing various tasks. Agents must make rational or reasonable decisions about what to do; they must make such decisions based on their knowledge of their environment and of the other agents with which they cooperate or compete. Co-operation and competition are dynamic—we consider that agents may sometimes wish to cooperate, sometimes compete and sometimes remain neutral, the choice of response being determined by factors which include the nature of the task or action, and the nature of the situation in which the action is to be taken.

Our view of Multi-Agent Systems is, as we have said, a dynamic one in which inter-agent (social) structures can form and disintegrate over time; agents can come into and go out of existence, and can change their allegiances to suit their own needs.

Any single agent may belong to more than one social group and will have to respect the codes of conduct of each of these groups at different (sometimes overlapping) times. This view of Multi-Agent Systems contrasts strongly with many systems in the literature: conventional systems have the character of having been designed, top down, to perform some particular set of tasks. We have in mind a much looser and more flexible collection of agents which can interact in ways that produce social groupings.

Even if we were to adopt the less flexible view that seems most common in current Multi-Agent Systems research, it would still be necessary for our agents to act in various ways and to various ends. In conventional systems, this is viewed in terms of making use of communication, perhaps also making use of skeletal plans that describe what each agent should do (Durfee's [9] Partial Global Planning is an example of this approach). We question the plausibility of repeated communication of plans between agents in a society; although intentions may be communicated from time to time, frequent communication of this kind, with its concomitant need for costly evaluation, does not appear to figure often in interaction between people. In conventional systems, each agent's role is relatively constrained and not subject to development: our view obliges us to adopt a more flexible approach. The symbolic interactionist approach provides theoretical constructs that directly relate to cognitive processes.

The processes of interest here relate to interactions between an agent and its acquaintances (we can ignore the problem of newly encountered others for the time being). Interaction between agents comes at a cost and cannot represent *all* possible information about an agent's goals and other intentions. Agents must interact and behave in ways that make sense to other agents. This way, agents can act in ways that are both intelligible and suitable—interaction is meaningful.

We have already seen that an agent's Self directs its actions. The Self is involved in mental simulation of proposed courses of action. There is also the concept of *role taking* and the concept of the *definition of the situation*. We will examine them in turn and see how they relate to Multi-Agent Systems. While reading what follows, it is essential to remember that all interactions (more generally, all actions) take place in situations and relative to the actor (agent's) perspective in that situation.

The concept of role taking is simple to explain. It is the process by which an agent mentally occupies the role of another agent in order to look at the Self and the situation from the viewpoint of the other. Hewitt [14] states that:

> "Role-taking always involves cognitively grasping the perspective of the other whose role is taken, and sometimes it also involves identification ... the child who responds to parental discipline—as in the case where the child apologizes and resolves to try to do better in the future after having been called to account for some transgression—engages in role-taking in which affect and evaluation are significant and identification is clearly evident. The child not only makes a factual prediction that

> the parent will be pleased with an apology and a resolution to do better in the future, but may also identify with the parent's perspective." (p. 99-100)

Role taking is a conscious activity in which we place ourselves in the position of the other involved in the situation. By putting ourselves in the position of the other, we adopt their perspective on the situation and see how they view us. In Hewitt's example, the child takes the role of its parent and sees first that an apology will please the parent. The child determines that a sincere promise to avoid repetition of the transgression accords with the parent's view of the situation. These decisions made by the child occur after it has recognised that it has done wrong: it can only make such a recognition after considering the situation from the parent's viewpoint and seeing that, from that viewpoint, the action it took was wrong.

By taking the role of the other, agents are able to determine various pieces of information about the situation, in particular how it is viewed by the other—the other's perspective—and how the other will react to various actions taken by the agent. In this sense, role taking serves to frame 'what if?' questions about the situation. The answers to these questions allow the agent to decide upon a course of action that the other will consider suitable and appropriate by the other. When seen in these terms, role taking is useful for many aspects of life.

Taking the role of the other enables an agent, *inter alia*:

- To engage in co-operative behaviour.
- To engage in symbolic communication.
- To control or direct others.
- To learn perspectives.
- To work through all social situations.

and it allows us to see the present from our past and from our future perspectives. Charon [2] discusses these and other issues in detail (pp. 109-114). Those aspects listed above, however, are most relevant present because they directly relate to 'what if?' questions of a kind that immediately and obviously relates to Multi-Agent Systems.

One can view role taking as being part of a dialectic between the Self and the other. The Self proposes a course of action and then the agent takes the role of the other to determine whether it is appropriate or suitable. The agent can alternate between the Self and role taking until a satisfactory outcome is reached. It is possible that no satisfactory outcome is reached: no course of action may be found that will meet the approved of the other. In such a case, the agent can determine some of the more likely consequences of the course of action (that is, how the other will react to the its actions). This enables the agent to protect itself against possibly harmful effects or to plan additional actions that are expected to mitigate against the unwanted effects of the inappropriate course of action. As a result of the dialectical exchange between the Self and the other, a strategy for acting within a situation can be derived. This strategy should have a better chance of success than would an arbitrarily chosen course of action because the strategy takes into account the position of the other.

As a result of this process, an agent may have decided to take some course of action the situation. It will also have learned more about itself. By taking the role of the other, an agent is able to determine how the other views the action that the Self proposes to take.

It is important to remember that actions (of all kinds) taken by agents are *situated*: they are relative to a situation. Interactions between agents take place within situations and are, indeed, conditioned by the situation. The *definition of the situation* is an important concept because it constrains the behaviour of the parties to an interaction. Hewitt [14] describes the definition of the situation as:

> "an overall grasp of the nature of a particular setting, the activities that have taken place there and are seen as likely to occur again, the objects to be sought or taken into account, and the others who are present. More formally, a definition of the situation is an organization of perception in which people assemble objects, meanings, and others, and act toward them in a coherent, organized way." (pp. 90-91)

Situations are constrained as to what they contain and what can happen in them. Each situation is either constructed by those agents which act in it or else is a construct that is known to the agents. Situations are not uncontrolled and chaotic, but are regular, meaningful objects that have purposes and that can be understood by agents. Understanding is important because situations have meaning according to symbolic interactionism and the meaning is constrained by the components of a particular situation.

The definition of the situation can considered as a representation of the situation that also constrains behaviour in it. It resembles Schank's [23] concept of the script, but is considerably more flexible. One reason why there should be such flexibility is that an agent's definition of a particular situation may be altered as a result of interaction with various other agents. Interaction (including role taking) need not always be with a concrete agent: it can also be an act of imagination involving generalised others and with reference groups—these are both abstractions that agents generate as a result of experiencing interaction with real agents. Through experience, situation definitions can be generalised and adapted (specialised as well as analogised) so that they can be used in concrete situations: agents learn about situations and their objects. This contributes to their knowledge about situations and about other agents. In summary, situations are flexible, varied, and resemble each other in various ways; they are also dependent upon the others with whom we interact as well as upon our own definitions.

The definition of the situation constrains how agents behave and how they interact with each other. An agent's knowledge of a situation *together with* role taking determine how the agent will interact in that situation. In order to engage in role taking, an agent must have knowledge of the situation and the perspective of the other agent. The perspective of the other agent can be equated with the acquaintance model we discussed above. Reasoning using the acquaintance model (as a mental model) and knowledge of the situation (the definition—another mental model) supports the process of role taking. We view role taking as a form of 'what-if?' reasoning that is constrained by various factors including the previous behaviour of the other (as represented by the model of the other) and the definition of the situation—another mental model. It should be noted that new acquaintances can also be catered for by this approach, particularly when reference group and generalised other information is available. The results of these inferential processes can be used to enrich (or maybe, even simplify) acquaintance and situation models; the Self model will also be affected by these processes because the point of role taking is to produce information for the Self.

Role taking is a process by which an agent can determine how to act in situations by placing itself in the position of the other agent so that it can view matters through the other agent's eyes. In order to do this, the agent needs information about the other agent (the acquaintance or general type of acquaintance) as well as information about the kind of situation in which the it is to act. One product of role taking is information about the Self: this is information about how the Self is viewed by the other and which is used in the development of the agent's concept of its Self. Another product is an outline of the course of action chosen by the agent, together with predictions of its possible and potential consequences.

5 The Priority of The Self

We now come to one of the important points of this paper, the concept which we will call *The Priority of the Self.* This principle states that the Self model is the basis for all acquaintance models; indeed, the Self is the basis for action and social behaviour, and also for the interpretation and understanding of symbols (actions and interactions) and situations. The Self model is used to generate acquaintance models. Acquaintance models are generated, we will argue, as a result of the combination of role taking, interaction, generalisation and analogy. The processes that support the construction of acquaintance models are all relative to situations: they take account of, and are situated in, the situations in which interactions occur.

The most important point is that process of role taking not only aids the formation of the Self and of perspectives, it also uses information about the Self in order to make predictions about the behaviour of others. Of course, we are unable to do this with complete accuracy (there will always be things which another will do that we would not do, and there will be things that another will not do that we would do). We do not have complete information about others and, instead, must rely upon observation and inductive generalisation (both of situations and of person types). We may have information about the other and we have information about the Self.

The act of imagination which places the Self in the position of the other must rely upon information that is provided by the Self. This information is in terms of the perspectives formed by the Self but which are imagined in altered form as being those of the other. This implies that the Self will bias view the world seen from the other's viewpoint unless there is balancing information. The reason that there is this bias is that information about the other may be lacking.

The Self is produced through interaction. Others are the original source of the information upon which the Self is based. Even as the Self is formed, information from external sources is used in its construction, so the perspectives of others are built into (so to speak) one's Self. Yet, at the same time, the Self's own perspective develops. These perspectives derived from others can form the basis of models of those who have contributed to the development of the Self. As a result of interaction with *significant others* (parents, guardians, siblings), children's selves are formed (initially as a result of imitation). Other agents (who have distinct and clear identities and relationships to the agent) have attitudes towards the agent and report them to the agent. This provides the agent with information about the attributes and the perspectives of the others. From this information comes the primitive acquaintance model. As a result of the development of the Self, agents also form their own views of others and these views—these perspectives—also serve to develop models of acquaintances. The Self's perspectives can be the subject of introspection: this provides another way in which the Self can develop and change its views of others; role taking will also contribute to the process.

In the face of (a relative) lack of information about the other, the Self has to rely on its own perspective and on what it knows. Clearly, what the Self knows about most is the perspective of its agent and its own history of interaction with others and with the world: the Self creates meaning out of situations and remembers these meanings—this forms the basis for the Self's understanding of the world and other agents. It would appear that the early development of the concept of the Other (other agents, acquaintances) will depend upon the concept of the Self for it is the Self that interprets interactions with others.

In terms of mental models, the Self model is therefore used to interpret the actions of others. These interpretations (together with other information) is remembered by the interpreting agent and forms the basis for acquaintance models as well as definitions of the various situations in which interaction occurs.

With experience of—after interaction with—the other, information about the other (attitudes, information about perspectives) can be brought to bear in taking the role of the other. Agents learn from experience, interactions with others, changes to the Self, and so on, in interpreting situations. The Self is better able to take the role of the other when more information is available to form the basis for its prediction of the possible behaviours and the perspectives of the other. Increased self knowledge results from interaction with the Self (introspection), role taking and from interaction with others. Increased self knowledge comes as a result of interaction with external agents for the most part: interaction brings with it information that is derived from the meanings of the actions taken by others (especially interaction with the agent). Increasing self knowledge comes hand in hand with increased knowledge about situations and others. As a consequence, the Self is exposed to information which it can interpret as part of the relevant acquaintance models.

Self knowledge also results from role taking and directly introspective action. Purely private, internal interaction serves to deepen and broaden the Self. Role taking shows the Self how it is seen by others and depends upon the models that have been formed of the others; introspection requires a dialogue with the Self.

Role taking and the interpretation that accompanies interaction inform the Self, but they also assist in the formation of acquaintance models. For example, role taking helps an agent to make predictions about situations and the behaviour of other agents. The predictions thus formed can be verified or falsified as a result of future interaction. The results of evaluating predictions feed the development of the Self (by means of an explanation process that will affect the Self's actions in similar situations that occur in the future) as well as the models of the acquaintances involved in the situation. The behaviour of the acquaintances in the situation can be remembered and interpreted in order to alter the associated acquaintance models—the acquaintance models will, it must be stressed, be updated in a way that reflects the agent's understanding of the situation—the updates will reflect the meaning of the situation and the perspective of the agent.

Matters are made more difficult to explain by the fact that the Self and the various acquaintance models are dynamic structures that change as a result of interaction in time, and by the fact that the development of both are closely interrelated and interdependent. As the Self develops, knowledge about others develops as a direct consequence. Knowledge of others develops as a result of generalisation, first in the creation of generalised others and then of reference groups, each of which has its own associated perspective. It can be argued that role taking, both in determining action in a particular situation and also action with respect to a generalised other or reference group, is used to produce possible revisions of the agent's model of the other—role taking also contributes to the development of the Self.

We noted above that one aspect of role taking was that of allowing the agent to review its past and to reason about its future actions. In reviewing the past, it can revise beliefs about others and about situations: this leads to a reevaluation of both and also leads to alteration of the associated models. Role taking, in this form, depends upon knowledge of the Self; in particular, it requires new information, knowledge and newer perspectives in order to perform the review. It is clear that this use of role taking exploits experience in the revision of beliefs.

On the other hand, as acquaintance models develop, they provide a richer and potentially more accurate basis upon which to take the role of the other—they provide more insight into the potential behaviours of others. When role taking, the agent can make use of more information about the perspectives and potential interactions of the other. This will lead to predictions that have greater plausibility. As a result of interaction, agents will become more competent in predicting the behaviour of other agents and will be better able to form more accurate plans for engaging in the interaction. Equally, more developed acquaintance models provide a better basis for explaining the behaviour of the other (this is another use for role taking, and one we have not explored).

As an agent's models of its acquaintances become more developed, they exert influences on the Self (we have already seen the use of role taking in developing the Self). Thus, we have a circle, but one which starts with the Self[6].

6 The Cost of Dynamic Models

The models we have discussed are all dynamic. Initially, the Self model will be highly dynamic, perhaps being completely reorganised after every interaction in the early stages (the rate of change probably slows with time). As a consequence of the dynamism of the Self model, the acquaintance models are dynamic in nature. However, acquaintance models will not necessarily reach the degree of stability of the Self model: this is because acquaintance models will change as a result of situations and role taking.

It can be argued that dynamic models of the kind we propose are computationally too expensive to consider with any seriousness. We will argue that this is not the case and that dynamic models are necessary if we expect agents to respond appropriately in widely varied and varying situations.

Firstly, we note that people are able to reason about other people—family members, friends, colleagues, media personalities, politicians and so on— and are able to do so with apparent ease. This reasoning is based upon our mental models of others. If we are interested in making our artificial agents more like people than the simple and relatively inflexible structures we have at present, we need to endow them with the ability to build, modify and manipulate models of themselves and of other agents. In addition to this, one important aspect of the argument presented in previous sections is that symbolic interactionism analyses human behaviour in terms of these mental models and it is because of the existence of these models that people construct and find meaning in situations and interact socially with other people. In

[6] We could, and probably should, have mentioned definitions of the various situations in which agents find themselves (even though this might seem daunting because of their variety and because situations look suspiciously like contexts) because the definition of the situation depends, at least in part, upon meanings and participants. We do believe that definitions of the situation become more defined with time, but prefer not to discuss this at present.

other words, symbolic interactionism analyses things in terms of structures that can be regarded as dynamic mental models.

Second, these mental models are not used all the time during interaction. This is a problem with the models proposed in the Multi- Agent Systems literature to date: too much happens during exchanges. Agents are always engaged in plan monitoring and replanning as a result of failed expectations.

In addition to these behavioural modification problems, agents containing dynamic models in current Multi-Agent Systems are always adjusting their models of themselves and of other agents, activities that are computationally expensive. Updates to an agent's dynamic models are needed to keep the models up-to-date: this is essential for the production of appropriate, co-ordinated behaviour that helps satisfy the current global goal.

One possible reason for the need for such monitoring and replanning is that current agents do not have information about the situations in which the act. As we have noted, the definition of a situation represents a grasp on the part of the agent of the objects, activities, meanings, and so on, that are expected in any given context—a kind of model of the situation. For any situation, its definition serves to limit the participants' possible actions and interpretations. Given a situation, its definition (which is shared knowledge) constrains what the participants can do and what resources they can employ in the situation. Viewed in another way, the definition of the situation outlines a kind of plan or script that defines the roles of the agents acting in the situation. However, the range of behaviours is less restricted than a plan: the definition of a situation is more like a flexible strategy for behaving in a given context (it is normative rather than prescriptive). Some variation in behaviour is clearly permitted in the situation: any behaviour that respects the constraints of the definition will count as acting in accord with it.

This view contrasts strongly with the notion of planned behaviour encountered in much current work on Autonomous and Multi-Agent Systems. In these systems, agents sometimes react to environmental events (including interactions), but more commonly they create and then execute fine-grained plans which determine their actions—classical plans in one form or another. The symbolic interactionist account of behaviour in a situation is, by contrast, considerably more flexible: the definition of a situation generates an equivalence class of behaviours, all of which are acceptable in the situation. We could argue (we defer the details) that as long as the agents in a situation perform certain key actions, they will be deemed to be acting in accordance with the situation's definition (this also simplifies the problem of classifying and learning situation definitions).

Not only do current agents not possess definitions of the situation, they are unable to engage in role taking as part of the process of determining reasonable behaviour. Role taking involves the use and modification of mental models in the production of action.

In addition to the problems we have seen with respect to behavioural modification, dynamic models are expensive, as we have noted above, because the models used in current systems contain information that is at too low a level. This has the consequence that there is a great deal of information that will be irrelevant to any particular situation and must be eliminated before appropriate reasoning can occur. Some models may not even contain abstractions of a kind that most suits analysis. We have already suggested that models should contain relatively high-level information: now we also suggest that it should be organised as a collection of abstractions.

Some approaches to belief revision (which is clearly akin to the problem of dynamic models) require every consequence of a revision to be propagated through the belief system before use can restart. This is clearly implausible (in fact, as plausible as the idea that we find *all* logical consequences of an inference). People do not try to alter all of their beliefs and are perfectly content to maintain contradictory beliefs. The solution to this problem is only to update directly affected beliefs (and at the appropriate level of abstraction) and to leave the rest until needed (i.e., perform only local updates to the belief network).

A further problem with the use of models in current Multi-Agent Systems is that they are used *during* interaction. Part of the justification for this is that behaviour is determined by planning activity: plans must be constantly updated as a result of changes to the environment. Models must be updated as a result of interaction because they partially determine how plans are to be modified when they fail or when they are found to be inappropriate. Interaction is also viewed as a kind of planned activity, both in terms of production and in terms of understanding. As an interaction proceeds, the models upon which it is based will be updated for the reasons just given. The various models an agent possesses serve as the 'world models' the planner consults and manipulates: as such they must be kept up-to-date so that the planner can determine what to do next, and so that the failure diagnosis mechanism can determine where and how a plan has failed. Given this framework, it comes as no surprise that dynamic models are considered expensive: they appear to inherit all the problems with world models from classical planning and appear to incur all their overheads.

Our view of the models held by each agent and their roles in the operation of Multi-Agent Systems is very different from the current view. Above, we have stressed the use of models (of all forms) in *prediction* and in *explanation*. These are different from constituting a world model for classical planning. Indeed, it is to be expected that they will not be performed during interaction: their major uses will be before and after interaction, respectively. When an agent is able to made predictions and engage in explanation, other tasks need to be performed during interaction. For example, processes checking that predictions have been satisfied (and noting failures when they occur as well as triggering improvised behaviour that will be appropriate within the current situation), and noting that significant constraints of the situation have been satisfied (again, noting cases of failure and inducing remediating, or coping, behaviour when failure occurs).

Given our view of the use of models, they may still be consulted during interaction, but it will be more usual for them to be used while interaction is not taking place. An example from human activity is mentally acting out an interview before it takes place and then reviewing the interview after the event. During review, the various models may be updated as a result of explanation and as a result of what is remembered about the interaction and its content: the expensive processes that modify models can be executed at a time when processing loads are lighter and when there is less demand for resources, in particular when there are lower demands on models[7]. However, review is not the same as performance, and mentally acting out an interaction before it takes place is not the same as the interaction proper. Models are used during mental simulation (prediction generation) and are used for explanation.

We believe that these properties of our interpretation lead to more appropriate and plausible demands upon agents' cognitive systems. The expensive operations of updating and restructuring models implied by interaction are performed when interaction within the target situation is not taking place. This appears, from introspection, to be more in line with the way in which people deal with their beliefs about others.

We are not claiming that interaction proceeds without access to models. One clear place for access is when an expectation fails and an agent has to deviate from its expected behaviour (for example, one might be expecting a hostile response from the other but, instead, the other behaves in an avuncular fashion). It may be enough to improvise behaviour or it might be necessary either to engage in role taking or to examine the definition of the situation. (Role taking will probably be avoided because it is too costly.) New predictions about how the situation will unfold will be needed so that the agent knows what to expect and so that the agent will take appropriate actions: predictions can come from the acquaintance model or from the definition of the situation (probably both). The formation of predictions is not as costly an operation as updating a mental model and propagating the consequences of the change: it depends upon retrieving structures from models and operating on what has been retrieved. Most of the cost of expectation formation occurs away from the models: it is only later when the expectations are analysed that costly modification processes become involved.

We believe that the above account provides an effective argument against the claim that agent models must be static because dynamic models are too expensive. The argument that dynamic models are impossible does not, in any case, square with the facts of psychology.

[7] In an implementation of a model-using agent, there will be contention for the various models if they are simultaneously accessed and updated during interaction. The demands of planned interaction imply that the latest information about the acquaintance or about the Self will be needed. Contention for the various models may become a severe problem for a totally deliberative agent of the 'classical' type.

7 Conclusions

We have considered in detail the role of various mental models that agents in a Multi-Agent System possess. We believe that cognition is mainly about the formation, use and revision of mental models: this is a stronger formulation of the mental models theory as presented by Johnson-Laird [16]. Mental models, as McGinn [20] argues, provide a mechanism for the consistent assignment of meaning to mental structures. This, as we have argued, is entirely consistent with a symbolic interactionist interpretation of agents.

Throughout the paper, we have stressed the dynamic nature of agents and the environment in which they live; we have also stressed the fact that agents change as a result of interaction with their environment, with other agents *and* with themselves. Symbolic interactionism is a theory of *situated* action, and one that relates internal (mental and symbolic) activity to overt and symbolic behaviour that is subject to direct observation and interpretation. A symbolic interactionist account of behaviour naturally implies structures that are very similar to mental models and processes that operate upon them. The models that symbolic interaction proposes are highly dynamic in nature, a property which suits our conception of Multi-Agent Systems.

So far, we have only considered human and artificial agents s separate 'forms of life.' We have proposed a framework and a set of mechanisms that apply to both kinds of agent when considered separately: we have, in effect, proposed an account of action that is both public and private—that is, a theory which deals with the production of public behaviour given private processes. What we have not mentioned is the possibility that both human and artificial agents may coexist and cooperate within an overall system. Such a system (which we refer to as a *Symbiotic System*) would be a society of agents in which each agent is able to perform some tasks better or more easily than others, and in which social behaviour occurs between agents of different kinds. Such a system or society of agents appears natural when considering Multi-Agent Systems from a symbolic interactionist perspective; however, as with all kinds of artificial agents, it is necessary for people to understand artificial agents in their own (human) terms. This places behavioural constraints of the kinds that we have discussed above on the artificial agents.

We believe strongly that work such as that represented by this paper, as well as the modelling of human social behaviour will increase our understanding of social process. Above, we approached the problems of agent modelling from a human perspective. This is for a number reasons: in particular, we want to suggest that our understanding of social theories can be deepened by constructing faithful simulations using Multi-Agent Systems (i.e., by the construction of computer systems in a way similar to that which has benefited cognitive psychology). Symbolic interaction is a framework within which to conduct such experiments (we could enrich it by incorporating some form of attribution theory—for example, Heider's original [12]—

as one of our inferential processes). We suggest that the approach is not limited to symbolic interactionism, but can be applied to other theoretical positions as well. We have concentrated on symbolic interactionism because of its naturalness and because it appears most obviously suited to Multi-Agent Systems. However, our interest in cognitive science leads us to make the suggestions in this paragraph, and we hope to expand the argument at a later date.

From the Multi-Agent Systems, as well as a general scientific, perspective, it is all very well to propose theories and mechanisms: the acid test comes in the implementation—can the proposals made above be implemented and, once implemented, do they work in the way suggested? We are aware that a computational model needs to be constructed and tested. We are also aware that the account given above misses many details: for example, how do agents make inferences about new agents, about situations and the behaviour of other agents? Taking the last case, how would an agent know what to infer during role taking? These issues need to be addressed because only then can the proposals be fully evaluated: it might be useful, for example, to allow agents to use implementations of Attribution Theory [12] or Personal Construct Psychology [18] as internal submechanisms which guide some of the processes that we have described. *Implementation* is the key concept: social structures are, by their nature, dynamic. Dynamic processes are hard to comprehend when they are presented in an entirely static fashion. Dynamic systems have subtle and far-reaching effects and interactions that cannot be determined from static views of them. Multi-Agent Systems (implementations) are a way of animating theories of dynamic, social systems.

References

[1] Blumer, H., Sociological Implications of the Thought of George Herbert Mead, *American Journal of Sociology*, Vol. 71, pp. 535-44, 1966.

[2] Charon, Joel M., *Symbolic Interactionism*, Fourth Edition, Prentice Hall, Engelwood Cliffs, NJ, 1992.

[3] Craig, I.D., *The CASSANDRA Architecture*, Ellis Horwood, Chichester, 1989.

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

[5] Craig, I. D., *Meanings and Messages*, Research Report No. 187, Department of Computer Science, University of Warwick, 1991.

[6] Craig, I. D., *Replacing CASSANDRA*, Research Report No. 232, Department of Computer Science, University of Warwick, 1992.

[7] Craig, I. D., A Reflective Production System, *Kybernetes*, to appear, 1994.

[8] Craig, I. D., Rule Interpreters in ELEKTRA, *Kybernetes*, to appear, 1994.

[9] Durfee, E. H., *A Unified Approach to Dynamic Coordination: Planning Actions and Interactions in a Distributed Problem-Solving Network*, Technical Report 87-84, Department of Computer and Information Science, University of Massachussetts at Amherst, Amherst, MA, 1987.

[10] Gasser, L., Braganza, C. and Herman, N., MACE: A Flexible Testbed for Distributed AI Research, in *Distributed Artificial Intelligence*, ed. Huhns, M. N., Pitman, London, 1987, pp. 119-152.

[11] Gasser, L., Braganza, C. and Herman, N., Implementing Distributed AI Systems Using MACE, in *Readings in Distributed Artificial Intelligence*, ed. Bond, A. H. and Gasser, L., Morgan Kaufman, Los Altos, 1988, pp. 445-450.

[12] Heider, F., *The Psychology of Interpersonal Relations*, Wiley, New York, 1958.

[13] Hewitt, John P., *Self and Society, Fifth Edition*, Allyn and Bacon, Needham Heights, MA, 1991.

[14] Hewitt, John, P., *Self and Society*, Fifth Edition, Allyn and Bacon, Boston MA, 1991.

[15] Hickman, C. A. and Kuhn, M. H., *Individuals, Groups and Economic Behavior*, Dryden Press, New York, 1956.

[16] Johnson-Laird, P. N., *Mental Models*, Cambridge University Press, 1983.

[17] Johnson-Laird, P. N. and Byrne, R. M. J., *Deduction*, Erlbaum, Hove, 1991.

[18] Kelly, G. A., *The Psychology of Personal Constructs*, Norton, New York, 1955.

[19] Mead, George Herbert, *Mind, Self and Society*, University of Chicago Press, 1934.

[20] McGinn, C., *Mental Content*, Basil Blackwell, Oxford, 1989.

[21] Norman, D. A., Some observations on mental models. In Gentner, D. and Stevens, A. L. (eds.), *Mental Models*, Erlbaum, Hillsdale, NJ, 1983.

[22] Pylyshyn, Z., *Computation and Cognition*, MIT Press, A Bradford Book, Cambridge, MA, 1984.

[23] Schank R. C. and Abelson, R., *Scripts, Plans, Goals and Understanding*, Erlbaum, Hillsdale, NJ, 1977.

[24] Schank, R. C., *Dynamic Memory*, CUP, 1982.

[25] Schank, R. C., *Explanation Patterns*, Erlbaum, Hillsdale, NJ, 1986.

[26] Varela, Francisco J., Thompson, Evan and Rosch, Eleanor, *The Embodied Mind*, MIT Press, Cambridge, MA, 1991.

[27] Wittig, Thies (ed.), *Archon: An Architecture for Multi-Agent Systems*, Ellis Horwood, Hemel Hempstead, 1992.