Foundations of an Agent Theory of Mind Model for Conversation Initiation in Virtual Environments

Christopher Peters
*LINC Laboratory
University of Paris 8
`c.peters@iut.univ-paris8.fr`

Abstract

We investigate a theoretical model of conversation initialisation that utilises a theory of mind model from evolutionary psychology for agents in a virtual environment. Agents attend to the level of attention that other agents pay to them and found their decision to engage in interaction on this interpretation as well as internal goals. For example, sometimes one person may want to engage in discourse while the other would prefer just to nod or say ‘hello’ and move on. The theory of mind module is primarily based on an agents perception of the others gaze behaviours, which we deem here to be a significant cue to an interest to interact. We hope that such a model will provide a link between currently disparate scenarios involving agents moving freely in virtual environments and those involving ECAs during close-up interactions.

1 Introduction

Many scenarios involving conversing characters presume the interaction has already started, with all parties standing within speaking distance, and the roles of speaker and listener assigned. Less studied has been the question of how such interactions begin in the first place; a flexible system cannot presume that agents will always schedule meeting arrangements with each other, but should also consider chance encounters e.g. seeing friends while walking down the street. Also, even if meeting engagements are made, it is possible that our agents will not all be very good timekeepers!

In many animations involving agents, groups either congregate based on relatively high-level rules suitable for large group interactions (see for example, Villamil et al. (2003)), or else, as is the case with most ECAs, it is presumed that conversation has already been joined and those involved are known.

Here, we describe a model where agents are provided with basic attributes encoding their social relations with other agents as well as their goals to engage in conversation. Agents cannot access other agents conversational goals directly and therefore they do not know if the other agent wants to engage in conversation with them. Rather, agents are endowed with synthetic senses and perception, and must formulate their own theory on whether the other agent wants to converse, based on their perceived *level of interest* in conversing. Level of interest is determined primarily through gaze and direction of intention, but our model also facilitates the inclusion of gesture and facial expression. We give gaze and direction of attention special importance; it is known that the ability to ascertain social signals directed towards the self, such as gaze signals, are important for establishing communicative intent (Kampe et al., 2003). After all, if somebody makes a waving gesture, it may not be significant to an agent if it is not directed at that agent: gaze direction is an important way of indicating to whom gestures and facial expression are directed.

The mechanism for doing this reasoning, and the primary focus of this paper, is a theory of mind module, or *ToMM*. The theory of mind module, which we base on an important model from evolutionary psychology (see Baron-Cohen (1994)), gives special importance to the direction of another’s attention, and to the eyes in particular. Indeed, it has been argued that the ability to read the behaviour of others in terms of their mental states would be advantageous for the survival and reproduction of an organism and that this may have strong links to the interpretation of another’s gaze (Baron-Cohen, 1994). The social importance of gaze is perhaps underlined by recent findings that privileged processing in brain areas such as the amygdala takes place when eye gaze is directed as opposed to averted (Wicker et al., 2003). The theory of mind module therefore contains special eye and head direction submodules that detect when the eyes and head of another agent is oriented towards a subject agent, S1.

Our theory of mind module is therefore specialised towards storing an agents theories that are of importance in the context of social initiation, and is not intended to be anything like a full theory of mind of another agents intentions or beliefs. As such, the range of variables in our ToMM are kept minimal and simple, although the determination of their values is complicated and their use allows agents to acquire extra reasoning and behaviours that an agent without such a ToMM would not possess.

2 Background

A number of researchers have emphasised the importance of a theory of mind for social functioning. Our model is primarily based on that of Baron-Cohens, who has postulated a number of modules in infants that may give rise to Theory of Mind (Baron-Cohen, 1994). We chose this model, since the modular information processing approach that it adopts is easily adaptable and very useful for constructing a computational model for computer agents. It also represents a nice higher level layer that can be added to previous work on synthetic sensing and memory for agents (Peters and O' Sullivan, 2002). We will now consider theory of mind in more detail.

2.1 Theory of Mind

Theory of mind research considers the mechanisms and interplays that are involved in using perceived information to create theories regarding the intentions of others. One influential model of theory of mind comes from evolutionary psychology and has been proposed by Baron-Cohen (Baron-Cohen 1994). It suggests that the ability to read the behaviour of others in terms of their mental states is advantageous for the survival and reproduction of an organism and that this may have strong links to the interpretation of another's gaze. We look at this model in more detail, as it forms a framework for our research.

Baron-Cohen suggests that the brain contains a series of specialised modules that enable humans to attribute mental states to others (see Figure 1). These modules are thought to be present and functioning in most humans by four years of age. The modules of interest here are enumerated as follows:

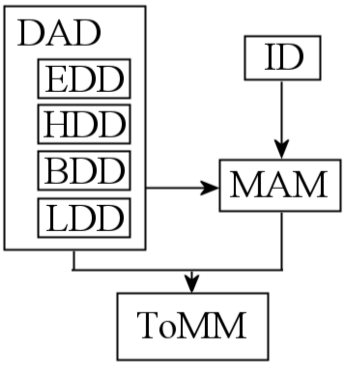

Figure 1: Simplified schematic of our version of Theory of Mind based on description by Baron-Cohen and elaborated by Perrett and Emery. The constituent modules detect volitional behaviour and the direction of attention of an entity and attribute theories of mind to it.

- Eye-direction Detector (EDD) The EDD is a social cognition module exclusively based on vision. It functions by detecting the presence of eyes or eye-like stimuli in the environment and computing the direction of gaze (e.g. directed or averted).

- Intentionality detector (ID) The ID module attributes the possibility of an object having goals and desires based on self propulsion, i.e. notions of animacy and intention. One should not, for example, attribute volitional behaviour to a brick, even if it is moving in the environment.

- Theory of Mind Mechanism (ToMM) This module stores the attribution of mental states to the other agent and is based on the results of interactions between the other modules. It contains working theories that may not necessarily be correct, but are nonetheless vital for forming an internal representation of the possible motives behind the actions of other living entities.

Perret and Emery (Perrett and Emery 1994) build on this work to propose further module classifications:

- Direction of attention detector (DAD) This is a more general form of the EDD above, that combines information from separate detectors that analyse not only gaze, but also body and direction of locomotion.

- Mutual attention mechanism (MAM) This is a special case of shared attention where the relationship is dyadic, involving mutual gaze and eye contact. In this situation, the goal of the participants attention is each other.

These models, provided by Baron-Cohen and Perrett and Emery, have been inspirational to us for creating a direction of attention and theory of mind model applicable to autonomous human-like agents in virtual environments. Before we look at this model in Section 4, we will first see how related work in robotics is already using such ideas to successfully enhance the social capabilities of robots in that domain.

3 Related Work

In the field of social robotics, Scassellati (2000) is constructing a humanoid robot as a test bed for the evaluation of models of human social development. The robot, Cog, has been endowed with social abilities using models of social development in both normal and autistic children. Scassellati proposes a merger of two models of theory of mind, including Baron-Cohen's model. The model first considers the movement of environmental stimuli in terms of the physical laws in order to distinguish between animate and inanimate objects. Animate stimuli are then further processed by Baron-Cohens model. Unlike robotics systems, our approach is easier to implement since we are dealing with a virtual environment and virtual sensors: using the synthetic vision module, difficult and time-consuming issues such as segmentation and recognition are avoided.

Horvitz and Paek (1999) present a computational model of conversation called the Bayesian Receptionist. The system uses Bayesian user models to infer the communicative goals of speakers based, not only on natural language processing of their utterances, but also on visual findings, such as spatial configuration and attire. Importantly, this work stresses the critical role of uncertainty in conversation.

In the area of agent and avatar simulation, the importance of conversation initialisation has been outlined (Cassell et al., 2001), although computational models do not appear to be widespread. One system that does consider the opening of engagements is the BodyChat system (Vilhjálmsson and Cassell, 1998). People are represented in online virtual worlds through avatars that behave automatically using socially significant movements, including salutations and back-channelling, based on text entered by a user. This work continues previous research outlining the role of conversation initialisation in generating plausible social behaviours (Vilhjálmsson, 1997).

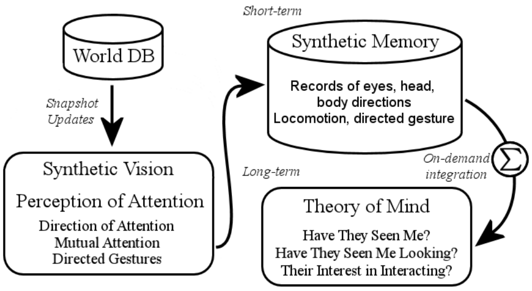

Figure 2: An overview of our model.

4 Our Theory of Mind Model

The core components of our model are the synthetic vision, direction of attention detector, synthetic memory and theory of mind modules (see Figure 2). The high-level operation of the model is summarised as follows:

The vision system takes frequent snapshots of the environment in order to provide visibility information. The ID module filters out any visible objects from the vision system that are not agents in order to provide a fast approximation to the ID module mentioned in Section 2.1. We have expanded on the model by Baron-Cohen to take into account enhancements suggested by Perrett and Emery (Perrett and Emery, 1994), in particular, the use of the more general Direction of Attention Detection module, that encapsulates eye and head direction, body orientation and locomotion direction. These extra features are of significance to us, because at greater distances or in occluded situations, the eyes and head may only be partially visible (or not distinguishable), and other body parts may have to be relieved upon for perception of the direction of another's attention. Thus, the direction of attention of visible agents is measured at each update of the vision system by the DAD module (Section 4.1). The DAD stores perceived attention entries recording this information over a time period in the memory system. The memory system therefore acts as a short-term storage for an observed agents attention behaviours. MAM is implemented by simply checking the information from DAD for mutual gaze.

The consideration of all of the entries in memory (Section 4.3.2) for a single agent provides a profile of the attention they have been paying; when viewed as a whole, this provides a more global indication of their overall interest. For example, consider an agent who gave a small wave upon passing by, but didn't intend to stop to converse. If we only considered the attention level at the time of the gesture, it would be relatively high and, interpreted in isolation, could indicate a willingness to interact. However, studying the full profile might indicate that this was just a peak in attention following by a drop that could be interpreted as an uninterested, but perhaps mannerly, agent.

Our ToM module (Section 4.4) sits on top of DAD and MAM and is connected to memory. On demand, it may integrate and interpret the attention profile present in memory into a coherent *interest level*. This allows the formation of a theory about the observed agents intention to engage in interaction. The connection to DAD and MAM also allows agents to store theories about another agents awareness of them and if they think another agent is aware that the current agent is aware of them. As such, our ToMM contains information on whether our subject agent, S1, (a) thinks another agent has seen it, (b) thinks the other agent has seen it (S1) look at that agent (S2) and (c) its theory of the goal of S2 in engaging in conversation. All of these are of use in conversation initiation. First of all, we presume that interaction cannot take place unless our subject, S1, has an awareness that another agent, S2, is present. But for conversation to take place, mutual awareness is required, as well as each agents knowledge that the *other* agent is *aware of it*. In this way, (a) (b) and (c) can be used to establish if another agent has an intention to interact and can also be used for enhancing the automation of more interesting conversation initiation behaviours (Section 5).

4.1 Direction of Attention Perception

Here we provide an quick overview of our DAD and ID modules for detecting the direction of attention. We do not cover these areas in detail, since they are not the focus of this paper: rather, we view them as "black boxes" that reliably provide information for use in the ToMM and illustrate that they are technically feasible for implementation through the use of synthetic vision and memory: interested readers are referred to Peters and O' Sullivan (2002) for more details on the concepts involved.

The synthetic vision module provides sensing of the virtual environment in a manner that is a crude approximation of human vision. A rendering is taken from the point of view of the agent and visible object lists are extracted and stored in a short-term storage area. The ID module may be implemented as a simple filter that only allows those objects that constitute agents through to memory for further processing. In essence, this contends that all agents are perceived to have the characteristic of animacy and agency and can therefore also be perceived as being capable of having goals and intentions. The DAD module consists of four submodules: an eye direction detector, head direction detector, body direction detector and body locomotion detector. Using information from the database, these modules detect if there are eyes and agents out there, and if so, if they are directing their attention towards the subject, S1. The DAD module also provides heuristics for selecting the contribution of the eyes, head, body and locomotion detectors to the overall judgement of attention direction: these weightings change depending on distance and occlusion information, which is provided by the synthetic vision module.

In order to obtain some temporal notion of the direction of anothers attention in order to link it to their overall *level of interest*, attention behaviours are stored in a memory mechanism. When considering another agent, the coupling of their current attention direction information from the DAD with previous information from the memory module provides this level of interest estimate which is used to formulate the theory of whether or not they wish to interact.

4.2 Agent Attributes

We have defined two key attributes for our agents that shape their goals and how they will interact: *relationship* and *conversational stance*. The *relationship* attribute indicates the state of the social relationship between S1 and S2. It can have one of the values *good*, *bad*, *neutral* or *stranger*. The relationship variable is a simple way for modelling the type of encounter that is taking place and determining the type of behaviours that will be animated during that encounter. Most models concerning conversation initiation presume that the speakers recognise each other and the encounters are always friendly. However, many encounters in everyday life may concern strangers (a tourist approaching you to ask directions) or even be of a confrontational nature (an angry neighbour approaching you to complain about your child smashing their window with a ball).

Conversational stance is defined as an agents goal or willingness to engage in conversation. It is presumed that the agent can be in one of three stances: it wants to interact (*interact*), it doesn't want to interact (*avoid*) or it is passive and has no particular preference (*don't care*). In the final case, an agent S1 with a passive stance will base their decision to interact on the perception of whether S2 intends to interact. An agent with stance set to *avoid* does not want to engage in conversation: such agents will not attempt to attract the attention of other agents in the environment, even if they are on friendly terms.

Both of these attributes allow high-level, albeit somewhat limited control of social encounters. For example, an agent that has its stance set to avoid would still provide a salutation behaviour to an agent it has a good relationship with, whereas it could be free to simply ignore the other agent if the relationship was bad.

4.3 Interpreting Another's Attention

Since our model is concerned with conversation initiation, the main interpretation that an agent tries to make about another agents attention behaviours is the willingness of the other agent to engage in conversation. That is, our model links the concept of attention and interest to the perception of the desire to engage in conversation; agents who do not show an interest in our subject, S1, are presumed not to want to engage in conversation. Agents that show a high interest in the subject will be perceived as candidates for engaging in further communicative acts or conversation. We propose the use of synthetic memory and belief networks to aid the calculation of the likelihood that another wants to engage in conversation based on the others current direction of attention, short term history of attention and any directed gestures that were made.

4.3.1 Directed Gestures

Among other cues such as verbal communication and facial expressions, directed gestures may have the effect of amplifying the perception of the interest of another. We use the DAD module to differentiate between normal gestures and what we call *directed gestures*. We regard directed gestures to be those gestures that one perceives to be directed towards them due to the coinciding fixation of the gaze of the other on the perceiver.

Our model currently takes account of whether a directed gesture was made towards S1 in a binary fashion. When agent information is being queried from the database by the DAD module, agents are also scanned for gestures that they are making. Gestures are only processed if the DAD or MAM considers that they are being directed to the agent in question based on the attention direction. Our model then accounts for the effect of such gestures on perceived attention, presuming they have been categorised as indicating a willingness to interact (or not interact) and a magnitude.

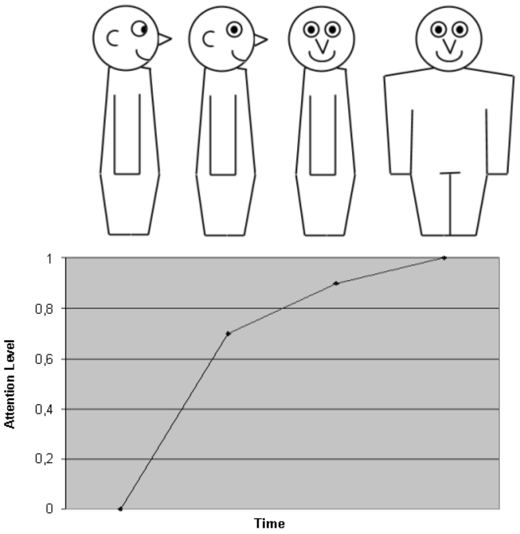

Figure 3: An example attention profile for an agent that gradually increases its attention towards S1.

4.3.2 Memory

The memory system contains records of the direction of attention of agents in the environment for each perceptual update, including their attention level at the time, flags indicating parts of the body that were directed and timing information. The memory system also stores records of locomotion from the LDD and directed gestures. Of key importance here is the ability to concatenate multiple separate memory entries, each with a separate attention level, into a single coherent indicator of the agents attentive actions over a certain period of time. We do this by constructing and analysing an *attention profile* from memory. An attention profile is a curve that is created to intersect attention levels over a specified time period. Analysis of the magnitude and slope of the profile encapsulates the information that an agent needs to later theorise, in the ToM module, about the intention of the other. A curve that is increasing over time indicates increasing attention over time by another agent, for example, if the agent was initially looking away from the subject and then looked towards them (see Figure 3). Peaks in the curve may be interpreted as 'social inattention' or salutation behaviours without the intention to escalate the interaction (see Figure 4). We regard an overall increasing or maintenance of an attention curve profile to be indicative of a likelihood that the other agent is willing to get involved in conversation. Entries regarding locomotion towards the agent are used to maintain the level of attention in cases where the profile drops while directed locomo-

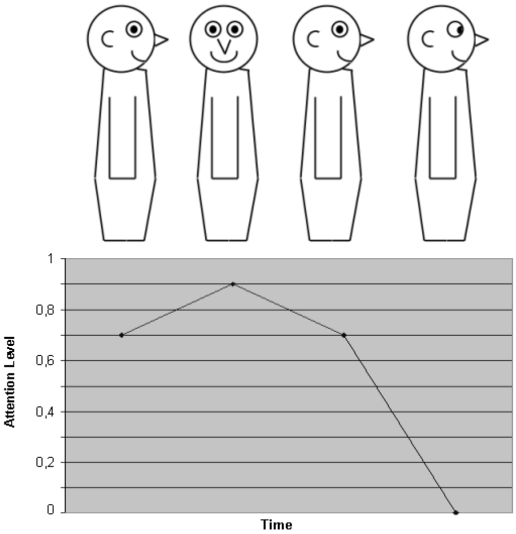

Figure 4: Example of a peak in an attention profile that may be interpreted as a sign of a social attention or greeting behaviour without the intention to become involved in conversation.

tion behaviours are occurring.

4.4 Theory of Mind Module (ToMM)

The Theory of Mind module for each agent stores a number of simple, but important, descriptors relating to the theories of the agents awareness of each other and the theory of whether the other agent intends to converse.

An agent will only commit to conversation when it perceives that there is a high chance the other agent wants to interact based on their converse theory. This attempts to allow simulation of human conversation initiation protocols, where smaller signals or probes for conversation initiation are first sent to save the sender from the potentially embarrassing social situation of starting a conversation with somebody who does not wish to talk.

4.4.1 Theory Representation

The theory of mind module stores a number of high-level variables that represent theories based on the perception of directed attention behaviours of others:

	Abbreviation	**Theory**
1.	HTSM	Have they seen me
2.	HTSML	Have they seen me looking
3.	IL	Interest level

1. *HTSM* Have they seen me: Does S1 think the other agent is aware of it. This theory is based on the consideration of eye gaze directions from memory and the DAD and MAM.

2. *HTSML* Have they seen me looking: Does S1 think the other agent is aware that S1 is aware of it. This theory is also based on the consideration of eye gaze directions from memory, particularly the MAM.

3. *IL* Interest level: How much interest has the other agent being paying to S1. This is based on the *attention profile* that is queried from memory (described in Section 4.3.2), as well as current attention direction information.

Even though these variables appear simple, they are very high level and their calculation is not trivial.

5 Conversation Initialisation and the ToMM

Here, we show how the previously described theory of mind module could be applied to to a more concrete example of agent conversation initiation. Agent behaviour is guided by finite state machines that run on each agent. As shown in Section 4.4, our ToMM contains information on whether our subject agent, S1, (a) thinks another agent has seen it, (b) thinks the other agent has seen it, S1, look at that agent, S2 and (c) its theory of the goal of S2 in engaging in conversation.

1. (a) If S2 has not seen S1, S1's theory shouldn't be that S2 doesn't want to interact, but that S2 is simply not aware of S1. If this is the situation, S1 can try to grab S2's attention in some way if it wants to interact, or can ignore S2 without invoking any social repercussions if it doesn't want to interact. This theory is referred to as the HTSM flag, which stands for Have They Seen Me. It encapsulates S1's perception of whether S2 is aware of it.

2. (b) In this case, we are storing S1's perception of S2's awareness of whether S1 has seen S2. That is, S2 may have seen S1, but may not be aware that S1 has seen S2. This is important for conversation: you must be aware of the other person, know that they are aware of you, but additionally, know that they are aware of you being aware to them. This theory is referred to as the HTMSL flag, which stands for Have They

Seen Me Looking. It is also useful for deception when S1 has seen S2 but does not want to interact: even if S2 subsequently sees S1, from S1's perspective, if S2 does not know that S1 has it, then it can attempt to ignore S2 and continue on its way without incurring any social repercussions. With this type of behaviour, the agent 'pretends' not to see another so it does not have to interact, even though such an interaction may only have been to signal that it could not engage in more lengthy interaction.

3. (c) This is the highest level theory stored in our system, and is S1's perception of whether S2 wants to engage in conversation. Essentially, it is S1's guess at the stance attribute of S2. This theory is called *converse*. Converse is based on the level of interest from the DAD module and memory, and gestures and facial expressions signalled to the agent.

In this way, the actual state changes that are made in the FSM are based not only on the agents goals and current state in the FSM, but also on their relationship and their theory of mind of the other based on information from the ToMM: that is, their respective perception of the others conversational stance and as well as the other theory variables in the ToMM.

5.1 Description of States

At any one time, an agent can be in one of the following five states:

1. Monitor Environment (ME) While in this state, the agent is attending to the environment in a general manner, watching out for agents that it is familiar with, or who may want to interact.
2. Grab Attention (GA) In this state, the agent attempts to grab the attention of another agent.
3. Passive Monitoring (PM) This state represents discreet monitoring of another agent without trying to attract their attention.
4. Gauge Reaction (GR) While in this state, an agent is actively sending signals and interpreting received signals to decide whether it should commit to conversation, or abort and return to monitoring the environment.
5. Starting Conversation (SC) In this instance, SC is presumed to be the terminating state of the state machine. In a full implementation, it would represent a transition to a node for handling in-conversation behaviours.

5.1.1 State Changes

The general operation of the FSM occurs as follows. It is presumed that the initial state of agent when in the environment is ME: that is, the agent is actively monitoring the environment, paying attention to other agents and related features i.e. gaze, gesture, facial expression. The agent stays in this state (edge 1 in Figure 5) while no social contacts are visible.

When an agent is in state ME and sees a social contact, S2, it can change to one of three different states:

It can switch to state GR in order to gauge the reaction of the S2 - it will look at S2 in order to try to ascertain its intention, resulting in a close to conversation (edge 3) or a decision not to engage in conversation and return to attending the environment (edge 4).

It can switch to state GA in order to attempt to grab S2's attention (edge 5). The agent will continue to do this while HTSML is false (edge 6) until it succeeds in grabbing S2s attention and gauges its reaction (edge 7), or else gives up and returns to attending to the environment (edge 8).

The agent can switch to state PM in order to passively monitor the other agent (edge 9). While in the passive monitoring state, an agent will continue to monitor S2 while it is in front of S1 and HTSML is false (edge 10). If HTML becomes true, then S2 has seen S1, and S1 switches to state GR in order to gauge its reaction (edge 11). If S2 passes out of range, then S1 switches back to monitoring the environment (edge 12).

It should be noted that in this paper, the start conversation state (SC) is the final state that the system may go into: once this state has been reached by agents, they will remain in it. Agents in this state are now analogous to those in systems consisting of close proximity conversation interaction and such systems could now be used to take control of the simulation for actual discourse.

State changes are not based solely on the current state and the HTSML flag - the agent attributes, *relationship* and *conversational stance*, also have a large impact, not only on what states the FSM transitions to, but also on the types of signals that the agent sends, in the form of facial expression, gesture and so on. That is, agents that are meeting to have an argument may have different initiation signals to those meeting to have a friendly discussion e.g. shaking of fist vs. wave (Kendon, 1990).

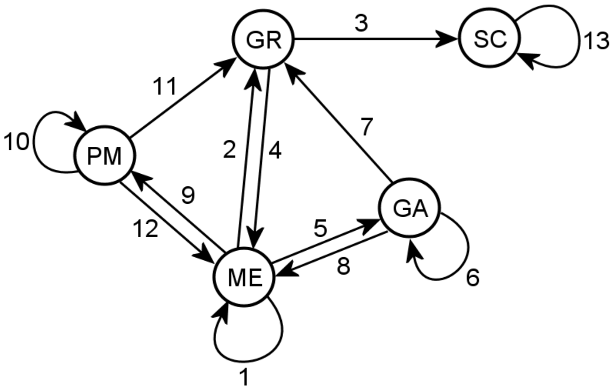

Figure 5: A high-level overview of state changes in the FSM. Actual state changes are also dependant on the conversational stance and relationship values that agents possess. Note that actions also happen when entering and leaving states; these actions may correspond to facial expressions, gestures etc.

6 Future Work

A number of future enhancements are possible for our model. One interesting area of research involves the inclusion of a shared attention module, or *SAM*, in our theory of mind module. This module is concerned with ones tendency to follow the line of sight of a person staring intensively at a particular object or location. Such a module would of use for creating situated environments in work with a similar emphasis to the Situated Chat project (Vilhjálmsson, 2003).

We are in the process of implementing the full model described in this paper for agents in the *Torque* engine (http://www.garagegames.com) and hope to test its effectiveness for automating more general social attention and inter-conversation behaviours.

7 Acknowledgements

The author would like to thank Catherine Pelachaud for her guidance and valuable suggestions throughout this work.

References

S. Baron-Cohen. How to build a baby that can read minds: cognitive mechanisms in mind reading. *Cahiers de Psychologie Cognitive*, 13:513–552, 1994.

J. Cassell, H.H. Vilhjálmsson, and T. Bickmore. Beat: the behavior expression animation toolkit. In *Proceedings of ACM Siggraph 2001*, pages 477–486, Los Angeles, August 2001.

E. Horvitz and T. Paek. A computational architecture for conversation. In *Proceedings of the Seventh International Conference on User Modeling*, pages 201–210, New York, June 1999.

K.K.W. Kampe, C.D. Frith, and U. Frith. “hey john”: Signals conveying communicative intention toward the self activate brain regions associated with “mentalizing”, regardless of modality. *Journal of Neuroscience*, 23(12):5258–5263, 2003.

A. Kendon. *Conducting interaction: patterns of behaviour in focused encounters*. Cambridge University Press, New York, 1990.

D.I. Perrett and N.J. Emery. Understanding the intentions of others from visual signals: neurophysiological evidence. *Current Psychology of Cognition*, 13:683–694, 1994.

C. Peters and C. O’ Sullivan. Synthetic vision and memory for autonomous virtual humans. *Computer Graphics Forum*, 21(4):743–753, 2002.

B. Scassellati. Investigating models of social development using a humanoid robot. In Barbara Webb and Thomas Consi, editors, *Biorobotics*. M.I.T. Press, 2000.

H.H. Vilhjálmsson. Autonomous communicative behaviors in avatars. Master’s thesis, Media Arts and Sciences, M.I.T. Media Lab, Cambridge M.A., 1997.

H.H. Vilhjálmsson. *Avatar augmented online conversation*. PhD thesis, Media Arts and Sciences, M.I.T. Media Lab, Cambridge M.A., 2003.

H.H. Vilhjálmsson and J. Cassell. Bodychat: Autonomous communicative behaviors in avatars. In *International Conference on Autonomous Agents*, pages 269–276, 1998.

M.B. Villamil, S.R. Musse, and L.P.L. de Oliveira. A model for generating and animating groups of virtual agents. In Thomas Rist, Ruth Aylett, Daniel Ballin, and Jeff Rickel, editors, *IVA*, pages 164–169, Irsee, Germany, September 2003.

B. Wicker, D.I. Perrett, S. Baron-Cohen, and J. Decety. Being the target of another’s emotion: a pet study. *Neuropsychologia*, 41:139–146, 2003.