eleventh
international
conference
on
autonomous
agents and
multiagent
systems

AAMAS 2012

4th- 8th June 2012
Valencia

W14
Workshop on Emotional and Empathic Agents (EEA)

Organising Committee

João Dias, INESC-ID and Instituto Superior Técnico, Portugal
Janneke van der Zwaan, Delft University of Technology, Netherlands
Jason Tsai, University of Southern California, USA

Senior Steering Committee

Ana Paiva, INESC-ID and Instituto Superior Técnico, Portugal
Catholijn Jonker, Delft University of Technology, Netherlands
Stacy Marsella, University of Southern California, ICT, USA
Virginia Dignum Delft University of Technology, Netherlands

Program Committee

Ruth Aylett, Heriot-Watt University, UK
Christian Becker-Asano, University of Freiburg, DE
Hana Boukricha, University of Bielefeld, DE
Joost Broekens, TU Delft, NL
Kerstin Dautenhahn, University of Hertfordshire, UK
Frank Dignum, Utrecht University, NL
Dirk Heylen, University of Twente, NL
Kate Hone, Brunel University, UK
Ian Horswill, Northwestern University, USA
Eva Hudlicka, Psychometrix Associates, USA
Toru Ishida, Kyoto University, JP
James Lester, North Carolina State University, US
Magalie Ochs, CNRS LTCI, TELECOM ParisTech, FR
Catherine Pelachaud, CNRS LTCI, TELECOM ParisTech, FR
Adriana Tapus, ENSTA ParisTech, FR

Preface

Creating characters and robots that give the illusion of life and allow for the user's suspension of disbelief is still a debated and fundamental goal in the area of virtual agents. Emotional and empathic relationships with these characters provide a path towards achieving this suspension. As such, this workshop will be a meeting point to discuss the creation of agents that are both empathic towards their users and foster empathic reactions towards them by their users. This workshop will be the third on this topic; the first one was organized in AAMAS 2004 in New York and the second in AAMAS 2009 in Budapest.

The main goal of this workshop is to bring together researchers from different disciplines to discuss the creation of what we call "empathic agents". Empathy has been associated with the processes that make a person to have "feelings that are more congruent with another's situation than with his own situation". Humans, when interacting with virtual agents or robots can be led to feel empathy, and experience a diverse set of emotional reactions. On the other hand, agents and robots can in a certain, perhaps limited way, also show certain emotions in reaction to human emotions, thus seemingly expressing empathy towards other agents and towards humans. Further, agents interacting in social simulation scenarios may react to the other agents in a way that is more congruent with the other's. Thus, by seeking inspiration in empathic relations established between humans and between humans and animals, in this workshop we expect to explore these dimensions of empathic agents.

9th, April, 2012

Contents

A BDI Dialogue Agent for Social Support: Specification and Evaluation Method

J.M. van der Zwaan
Delft University of Technology
Jaffalaan 5
2628 BX Delft, The Netherlands
j.m.vanderzwaan@tudelft.nl

V. Dignum
Delft University of Technology
Jaffalaan 5
2628 BX Delft, The Netherlands
m.v.dignum@tudelft.nl

C.M. Jonker
Delft University of Technology
Mekelweg 4
2628 CD Delft, The Netherlands
c.m.jonker@tudelft.nl

ABSTRACT

An important task for empathic agents is to provide social support, that is, to help people increase their well-being and decrease the perceived burden of their problems. The contributions of this paper are 1) the specification of speech acts for a social support dialogue agent, and 2) an evaluation method for this agent. The dialogue agent provides emotional support and practical advice to victims of cyberbullying. The conversation is structured according to the 5-phase model, a methodology for setting up online counseling for children. Before this agent can be used to support real children with real-world problems, a careful and thorough evaluation is of utmost importance. We propose an evaluation method for the social support dialogue agent based on multi-stage expert evaluation in which (adult) online bullying counselors interact with the system with varying degrees of freedom. Only when we are convinced that performance of the system is satisfactory, children will be involved, again in multiple stages and under the supervision of experts.

Categories and Subject Descriptors

I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems

General Terms

Design, Experimentation

Keywords

Conversational agents, Verbal and non-verbal expression, Modeling cognition and socio-cultural behavior

1. INTRODUCTION

Social support refers to communicative attempts to alleviate emotional distress and is aimed at increasing the well-being of people and decreasing the perceived burden of their problems. Recent developments in affective computing show that empathic agents are increasingly capable of complex social and emotional dialogues. However, these dialogues are predominantly task oriented, i.e. to help the user perform a concrete task, such as finding information or learning [15, 16].

Generally, giving social support is unrelated to this type of tasks; it is typically a non-task oriented effort. In our research, we are investigating how and to what extent Embodied Conversational Agents (ECAs) can provide social support. Recently, we proposed a design for an ECA that gives social support to children that are victims of cyberbullying [27]. Cyberbullying refers to bullying through electronic communication devices [17]. It is a complex problem that has a high impact on victims [18]. Research shows 40–60% of the victims is emotionally affected by incidents of cyberbullying [18, 20]. The anti-cyberbullying ECA tries to empower these victims by giving emotional support and practical advice.

The anti-cyberbullying agent implements different (verbal and non-verbal) strategies for giving social support. This paper is focused on the dialogue engine of the anti-cyberbullying agent, i.e. verbal strategies for social support. Therefore, the embodiment and non-verbal behavior of the agent are beyond the scope of this paper. In the remainder of this paper we use the term 'dialogue agent' to refer to the dialogue system and 'anti-cyberbullying agent' to refer to the complete system (the dialogue system combined with the embodiment).

Cyberbullying is a real problem, affecting real people. It is not our intention to present the anti-cyberbullying agent as a solution to cyberbullying. As mentioned before, our focus is on providing social support. Given the sensitivity of the topic and the vulnerability of the target audience (children), a careful and thorough evaluation is highly important. In this paper, we present 1) our implementation of different types of verbal social support and 2) our evaluation plan for the dialogue agent.

The paper is organized as follows. In section 2, we discuss related work on (embodied) conversational agents. In section 3 we operationalize social support for the anti-cyberbullying agent. Section 4 introduces the architecture of the dialogue agent. In section 5, we specify the social support types and explain how they were implemented in the prototype. Section 6 presents our plan for the evaluation of the dialogue system. Finally, in section 7, we present our conclusions.

2. RELATED WORK

Early work on affective computing demonstrated that agents are able to reduce negative emotions in users by addressing them [13]. Since then, emotional agents have been applied

Appears in: *Proceedings of the Workshop on Emotional and Empathic Agents, in the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, June, 4-8, 2012, Valencia, Spain.

predominantly in task oriented systems, i.e. systems that support users in performing concrete tasks, such as finding information. Examples include museum guide MAX that provides users with information about the museum and exhibitions [15], and agent GRETA that presents health information to the user [21]. Another popular application of emotional agents is responding to user emotions in e-learning and tutoring systems [8, 16, 26]. These so called pedagogical agents use different (emotional) strategies, such as displaying active listening behavior, encouragement and praise, to motivate the user and to make learning more engaging.

Cavazza et al. developed the 'How was your day' (HWYD) system, a non-task oriented ECA that allows users to talk about their day at the office [5, 23]. The system tries to influence the user's attitudes as a part of a free conversation on work related topics, such as office mergers, promotion and workloads. The system alternates between employing clarification dialogue (asking questions to find out details) and generating appropriate affective responses to the information gathered. The system allows users to speak uninterrupted for longer periods of time (utterances of > 30 words). In addition to short sympathetic responses to the user's input, the system may start a longer utterance to provides advice and support. These longer utterances are called comforting tirades. Comforting tirades are aimed at encouraging, comforting or warning the user. An important difference between the HWYD system and the anti-cyberbullying agent is the structure of the conversation. While the HWYD system incorporates social support into free conversation, the anti-cyberbullying agent imposes a structure on the conversation. This structure facilitates giving support, because the agent's verbal support actions are linked to this structure (see section 3).

Small talk is non-task oriented talk; it is not used for content exchange, but has a social function in the conversation. Giving social support has certain similarities with small talk. For example, almost all social support categories can be found in the small talk taxonomy presented by Klüwer [14]. However, small talk is also typically used in task oriented systems, for example real-estate agent REA uses small talk to make the user feel comfortable before asking questions about sensitive topics such as money [3]. For giving social support, our dialogue agent uses a sequence similar to the one defined by Schneider for small talk [22]:

1. A query from the dominant conversation partner (in our case, this is the dialogue agent),
2. An answer to the query,
3. A response to the answer, consisting of one of the following possibilities: echo-question, check-back, acknowledgement, confirming an unexpected response, positive evaluation,
4. An unrestricted number of null steps or idling behavior.

Generally, the dialogue agent will give support in step 3 of the model, for example by responding sympathetically to the user's answer to the query.

3. SOCIAL SUPPORT

Schneider's model specifies the dialogue agent gives support in response to the user, but it does not show what kind of social support is given. In this section, we provide a background on social support. The agent's verbal social support actions are based on a typology of social support in online settings [4]. This typology is relevant for the dialogue agent, because online communication is mostly textual and does not depend on additional communication channels (such as non-verbal behavior and auditory information). The typology consist of five main support categories [4]:

- Information support (messages that convey instructions),
- Tangible assistance (offers to take concrete, physical action in support of the recipient),
- Network support (messages that appear to broaden the recipient's social network),
- Esteem support (messages that validate the recipient's self-concept, importance, competence, and rights as a person), and
- Emotional support (attempts by the sender to express empathy, support the emotional expressions of the recipient or reciprocate emotion)

Each category breaks down into multiple subtypes. From these subtypes, 5 that occurred frequently in counseling conversations by chat [10] were selected to be implemented in the dialogue agent, that is sympathy (emotional support), compliment (esteem support), encouragement (emotional support), advice (information support) and teaching (information support). Table 1 lists descriptions and examples of these support types.

Support type	Description	Example
Sympathy	Express feelings of compassion or concern	How awful that you are being bullied!
Encouragement	Provide recipient with hope and confidence	I know you can do it!
Compliment	Positive assessments of the recipient and his or her abilities	Good of you to have told your parents!
Advice	Suggestions for coping with a problem	Perhaps you should tell your parents.
Teaching	Factual or technical information	You can block a contact by clicking the 'block' button

Table 1: The types of social support implemented in the conversational agent.

To facilitate giving social support, the conversation between the user and the dialogue agent is structured according to the 5-phase model. The 5-phase model was developed as a methodology to structure counseling conversations via telephone and chat [2]. The five phases of a conversation are:

1. Warm welcome: the counselor connects with the child and invites him to explain what he wants to talk about

2. Clarify the question: the counselor asks questions to try to establish the problem of the child
3. Determine the objective of the session: the counselor and the child determine the goal of the conversation (e.g., getting tips on how to deal with bullying)
4. Work out the objective: the counselor stimulates the child to come up with a solution
5. Round up: the counselor actively rounds off the conversation

The 5-phase model thus a template for the conversation. Even though multiple conversation objectives are possible, we assume the user wants to get advice on how to deal with a cyberbullying incident. Therefore, the third conversation phase has a trivial implementation; the objective of the conversation is fixed to 'get advice on how to deal with cyberbullying'. The 5-phase model assumes the child itself can come up with a solution. Since our goal is to demonstrate how a conversational agent can give verbal social support, we relax this responsibility and have the dialogue agent take the lead in phase 4. Additionally, to simplify the model, we assume certain types of support only occur in certain phases: sympathy, compliment and encouragement can occur in phase 2; advice and teaching only occur in phase 4.

4. ARCHITECTURE

Figure 1 shows the different components of the dialogue agent's architecture. This architecture is based on the generic architecture for companion agents and robots by Steunebrink et al. [24]. The reasoning engine is modeled according to the Belief-Desire-Intention (BDI) paradigm [6]. This means the dialogue agent has beliefs (e.g., about what advice to give in which situations), goals (e.g., to give social support), and plans (e.g., the 5-phase model). Grey boxes indicate components of the dialogue agent that have not been implemented in the prototype, i.e. the input interpretation and utterance formulation modules, the user profile and the emotional module. These components will be added to the dialogue in the future. Components that have been implemented are discussed next.

4.1 Input/Output

The agent and the user communicate through natural language text messages. Given the complexity of interpreting and generating natural language, in the current system, text interpretation and generation have not been implemented. Instead, the input and the output of the prototype consists of FIPA-ACL communicative acts [9]. The communicative used by the dialogue agent are inform and request. Inform is used to inform the receiver that a given proposition is true. Request is used by the sender to request the receiver to perform some action, for example to perform another communicative act (i.e. to answer a question). An example of a social support communicative act is:

```
send(user, inform, compliment( incident(response,
    confronted_bully), courageous) )
```

This communicative act represents a compliment given to the user for being courageous because he confronted the bully. A translation of this communicative act to natural language could be: *I think it was very brave of you to confront the bully!*

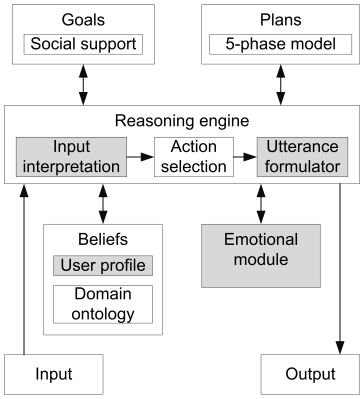


Figure 1: The architecture of the social support dialogue agent. Output is produced by the action selection engine based on the input and the agent's beliefs.

4.2 Beliefs

The dialogue agent's beliefs are stored in the belief base. The dialogue agent has beliefs regarding the domain (e.g., what questions to ask the user and what advice to give in different situations), social support (e.g., when to give which type of social support), and conversation management (e.g., how to open and close conversations). Additionally, the dialogue agent keeps up its beliefs about the current conversation phase, for example

```
conversation(phase, welcome)
```

and facts about the incident the buddy has learned from the user, for example

```
incident(incident_type, cyberbullying)
```

The contents of the speech acts (and thus of the conversation) are defined by the contents of the belief base. To enable reuse in other domains, the knowledge in the belief base is kept as generic as possible. This is achieved by separating dialogue management rules from domain specific knowledge. The action selection engine requests and updates information from the belief base.

4.3 Reasoning Engine

As mentioned before, the reasoning engine is based on the BDI paradigm. In the reasoning engine, beliefs are combined to select actions, which, in case of the dialogue agent, are speech acts. The main goal of the dialogue agent is to give social support. Giving social support is operationalized as completing the conversation with the user. The dialogue agent has a single plan to reach this goal, that is the 5-phase model. Beliefs about the conversation phase trigger subgoals and subsequently the dialogue agent's actions. In phase 1 (welcome), the goal is to have greeted the user. In phase 2 (gather information), the dialogue agent has the goal of knowing certain facts about the cyberbullying incident. Established facts (i.e., the user's answers to the dialogue agents's questions) may trigger speech acts to give different types of social support. The implementation of social support types is explained further in section 5. The third phase of the 5-phase model (determine conversation objective) is

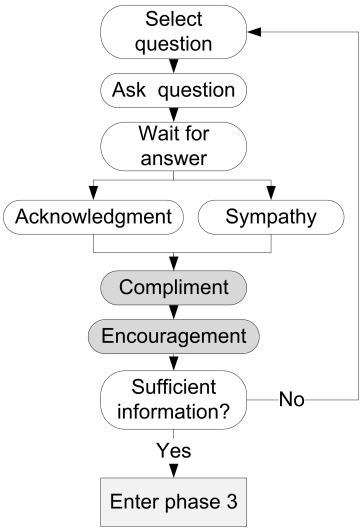


Figure 2: Social support in phase 2 (gather information). Darker grey boxes represent optional steps.

assumed to be fixed (and has a trivial implementation). In phase 4 (work out objective) the dialogue agent's goal is to have delivered all relevant advice. The advice is based on the information the dialogue agent gathered in the second phase. Finally, in phase 5 (close conversation), the dialogue agent has the goal to have said goodbye to the user.

5. SPECIFICATION OF SOCIAL SUPPORT TYPES

Here we specify the social support types that were selected in section 3. A prototype of the social support dialogue agent was implemented in GOAL, a high level agent programming language [12]. We assume that sympathy, compliment and encouragement only occur in the second conversation phase (gather information), and advice and teaching only in phase 4 (work out objective).

After greeting the user in conversation phase 1 (welcome), the second conversation phase (gather information) starts. Figure 2 gives an overview of phase 2. Phase 2 consists of a recurring pattern of the dialogue agent selecting and asking a question, the user answering that question, and the dialogue agent acknowledging the answer. An acknowledgement is either neutral (e.g., *I see*, or *Okay*) or sympathetic. In addition to acknowledging the input (either neutrally or sympathetically), the dialogue agent optionally compliments the user or encourages him. If the dialogue agent has gathered sufficient information (what is sufficient depends on domain knowledge), it enters the third conversation phase (determine conversation objective), which, in the prototype, has a trivial implementation; the dialogue agent assumes the user wants advice on how to deal with cyberbullying. The advice is delivered in phase 4 (work out objective), which is illustrated in figure 3. After selecting a piece of advice, the dialogue agent presents it to the user. If the dialogue agent advises the user to perform a task that requires technical knowledge, it will follow up with the question whether the user wants him to explain how to perform the task. If the user confirms, the dialogue agent explains how to perform

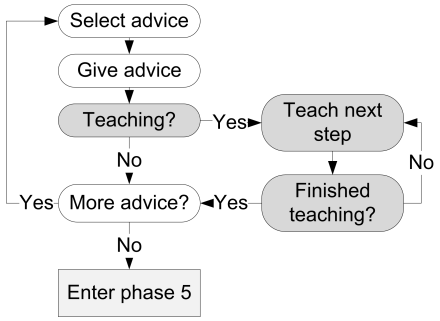


Figure 3: Social support in phase 4 (work out objective). Darker grey boxes represent optional steps.

the task step by step. If the dialogue agent has given all relevant advice, the fifth phase (round off) is entered and the dialogue agent says goodbye to the user.

5.1 Sympathy

Sympathy expresses feelings of compassion or concern. During the information gathering phase, the dialogue agent may respond sympathetically to answers given by the user. The dialogue agent expresses sympathy if it follows from his beliefs sympathy is applicable, otherwise it plays safe by staying neutral. The implementation of sympathy is illustrated in the following example:

Dialogue agent: *Can you tell me what happened?*
Child: *Someone is calling me names on msn*

The child's utterance causes the addition of the following `incident` facts:

```
incident(type_cb, name_calling).
incident(method_cb, msn).
```

to the belief base of the dialogue agent. Since the belief base also contains the following fact:

```
sympathetic_acknowl(type_cb, name_calling) :-
    incident(type_cb, name_calling).
```

the agent responds sympathetically to the user:

Agent: *That's awful!* (sympathy)

Absence of the `sympathetic_acknowl` rule would have resulted in a neutral acknowledgement of the user's input:

Dialogue agent: *I see* (acknowledgment)

5.2 Compliment

Compliments are positive assessments of the recipient and his abilities. In the context of a social support dialogue about a specific event, there are two possibilities for the dialogue agent to give compliments: 1) the user tells the dialogue agent he performed a constructive, positive or otherwise positive action (e.g., in response to being bullies, the user didn't retaliate), and 2) the user performs well as a dialogue partner (e.g., the user gives a clear explanation of something). Currently, only the first type of compliment

is implemented. The following example illustrates how the dialogue agent compliments the user.

Dialogue agent: *How did you respond when you were being called names on msn?*
Child: *I told him to stop, but he didn't listen*

The child's utterance causes the addition of

```
incident(response, confronted_bully).
```

to the beliefs of the dialogue agent. Additionally, the belief base contains the following information:

```
quality(courageous).

characteristic_of(confronted_bully, courageous).

compliment(Fact, Value, Quality):-
    incident(Fact, Value),
    characteristic_of(Value, Quality),
    quality(Quality).
```

The `quality` fact states courageousness is a quality and the `characteristic_of` fact links the user response to the quality. The `compliment` rule combines the `incident` fact with the quality and the user response. This enables the agent to compliment the user:

Dialogue agent: *I see.* (acknowledgment)
Dialogue agent: *That was very brave of you!* (compliment)

In case multiple compliments are triggered by an `incident` fact, the dialogue agent randomly selects one. This procedure will be extended in future work.

5.3 Encouragement

Encouragement is about providing the recipient with hope and confidence. The process of encouraging the user closely resembles the implementation of giving compliments. Again, we assume that encouragement is always given in response to a user utterance. Utterances indicating the user's situation is severe trigger encouragement. The circumstances under which a situation can be considered severe depend on domain knowledge. The implementation of encouragement is illustrated by the following example:

Dialogue agent: *Has he bullied you before?*
Child: *Yes, all the time*

The child's response results in the addition of

```
incident(bullied_before, often).
```

to the beliefs of the dialogue agent. Based on the following rule in the belief base:

```
encouragement(bullied_before, often):-
    incident(bullied_before, often).
```

encouragement is triggered and the dialogue agent encourages the user:

Dialogue agent: *I'm sorry to hear that* (sympathy)
Dialogue agent: *Let's try to stop the bullies!* (encouragement)

5.4 Advice

In phase 4, the dialogue agent gives advice on how to deal with cyberbullying. Which advice is given depends on domain knowledge and the specific situation of the user. The domain specific rules that trigger pieces of advice also provide a reason for giving the advice. The reason is added to the speech act to allow the dialogue agent to justify its advice. For example, if the belief base contains the following information:

```
incident(bully, classmate).

advice(talk_to_teacher, bully, classmate):-
    incident(bully, classmate).
```

, the advice `talk_to_teacher` is triggered by the `incident` fact. And the dialogue agent can say something like:

Dialogue agent: *Since you are bullied by a classmate, it might be a good idea to talk to your teacher.* (advice)
Child: *Okay*

In case the same piece of advice is associated with multiple explanations, the dialogue agent randomly picks one. In future work, we will implement more sophisticated means of choosing between multiple explanations, e.g. by selecting an explanation that has been successful before. After giving a piece of advice, the dialogue agent waits for confirmation from the user. Once the user has confirmed, it moves on to the next piece of advice, or phase 5 when all advice has been given.

5.5 Teaching

Teaching concerns conveying factual or technical information. In the domain of cyberbullying, many measures the user can take to protect himself require technical know-how, such as blocking users, changing accounts, and reporting inappropriate content. The implementation of teaching is based on the work by Vergunst [25]. The dialogue agent uses a list of instructions to explain how to perform the task. After the explanation of a step, the dialogue agent waits for confirmation from the user before moving on to the next step. The following example illustrates how teaching is implemented in the prototype. The belief base of the dialogue agent contains facts describing which pieces of advice require teaching and a list of instructions to perform the teaching:

```
advice(block_msn_contact, bully, anonymous):-
    incident(method_cb, msn),
    incident(bully, anonymous).

teaching(block_msn_contact,
  [login_msn, locate_contact, right_click_contact,
  choose_block_contact]).
```

The `advice` rule triggers the advice to be given and the `teaching` fact states this piece of advice might require teaching and provides the list of instructions.

Dialogue agent: *Because the bully is anonymous, blocking this contact on msn might stop him.* (advice)
Dialogue agent: *Do you want me to explain to you how to do that?* (teaching)

Child: *Yes, please!*
Dialogue agent: *First, you login to msn.* (teaching)
Child: *Yes*
Dialogue agent: *Next, locate the contact you want to block in your contact list.* (teaching)
Child: *OK*
...
Dialogue agent: *Okay, now you know how to block a contact in msn.* (teaching)

If the user would have answered *No, thank you* in the second turn, the dialogue agent would have aborted the teaching and moved on to the next piece of advice.

6. EVALUATION PLAN

The dialogue system described in the previous sections is a component of an ECA that should provide emotional support and practical advice to victims of cyberbullying. In addition to the dialogue system, the embodiment of the ECA is another main component of the anti-cyberbullying agent. We propose to evaluate the anti-cyberbullying agent in three stages: first, the dialogue system and the embodiment should be evaluated separately. This can be done in parallel. After that, the complete anti-cyberbullying agent, i.e. the combination of the dialogue system and the embodiment, should be evaluated. The advantage of evaluating the separate components before the complete system is that we will be able to clearly demonstrate the contribution of individual components to the results of the final system.

This section will describe our evaluation plan for the dialogue agent (the dialogue component of the anti-cyberbullying agent). The goal of the evaluation is to determine the extent to which users experience social support when interacting with the dialogue agent. This will be measured with a questionnaire that was used by Fukkink and Hermanns in a qualitative content analysis of support provided by a Dutch child helpline [11]. Prior to interacting with the dialogue system, participants will indicate on a 9-point scale how they feel (well-being) and how severe their problem is (perceived burden of the problem). These questions will be asked again after the interaction. In addition, participants will also rate (again on a 9-point scale) to what degree they felt supported, whether they now knew what to do, if they felt they had been taken seriously, whether they had been made to feel at ease, and if they understood the dialogue system's messages. Finally, participants will be asked to rate the trustworthiness of the dialogue agent. The perceived social support will be compared to perceived social support in conversations with human counselors.

The evaluation plan consists of multiple, incremental stages in which the dialogue system is improved based on the feedback from the previous stage before moving on to the next. If performance of the dialogue agent is unsatisfactory, the current stage will be repeated after incorporating the feedback. The different stages of the evaluation plan are listed in table 2.

Participants	Experiment
Online counselors	Dialogue fragments WOZ with scenarios Dialogue system with scenarios Dialogue system with free input
Children	WOZ with scenarios Supervised dialogue system with scenarios
Cyberbullying victims	WOZ with free input Supervised dialogue system with free input

Table 2: Overview of the multi-step evaluation plan for the dialogue agent.

6.1 Expert Evaluation

Since we are dealing with a sensitive topic (cyberbullying) and a vulnerable target audience (children), we need to know how good the system is before we involve children in the evaluation process. Therefore, we first perform an expert evaluation with adults trained to hold counseling conversations with children about different topics, including bullying[1]. These experts will be asked to interact with the dialogue agent from a children's perspective.

Before allowing the experts to interact with the dialogue agent, they will be asked judge fragments of social support conversations. This is done to make sure the dialogue system's messages are clear and understandable. The fragments will be similar to the example dialogues in section 5 and created from counselor utterances found in actual chat conversations. The experts will assess the fragments on understandability for children, recognizability and relevance of social support types, and the extent to which the formulation is consistent with the experience of the target audience. In addition, they can suggest alternative formulations. At the end of this stage we will have gathered a validated library with conversation fragments for the dialogue agent.

For the next stage, we will design scenarios of frequently occurring cyberbullying situations. In this stage, experts will interact with a Wizard of Oz (WOZ) system based on these scenarios. In a WOZ experiment, a human experimenter selects the utterances of the dialogue agent. Participants first read the situation description from one of the scenarios and put themselves in the shoes of the main character. Next they fill out the pre-test questionnaire, interact with the WOZ system and fill out the post-test questionnaires. Finally, the participants are asked to give feedback on how the conversation went. They will be asked to elaborate on what went well, what could be improved, and to what extent the conversation similar was to a conversation with a human counselor. Based on the feedback, the dialogue agent will be improved.

For the next experiment we follow the same procedure. However, instead of interacting with the WOZ system, participants interact with the actual dialogue agent. After processing the feedback and updating the dialogue system, participants will interact with the dialogue agent based on free input. This means the counselors can come up with situations based on their experience and ask the dialogue agent for advice.

6.2 Involving Children

If the previous experiments have been completed successfully, we can start to involve children in the evaluation pro-

[1]For the development and evaluation of the anti-cyberbullying buddy we are cooperating with psychologists from the Open University (the Netherlands) and (online) counselors from Pestweb (www.pestweb.nl).

cess. All experiments in which children participate will be conducted in cooperation with and under the supervision of experts (i.e. online counselors and/or psychologists). The first stage in the evaluation with children is a WOZ experiment with the scenarios from the second experiment of the expert evaluation. Because we use scenarios, there is no need to recruit children that have experience with being cyberbullied. The wizard will be played by an online counselor. The dialogue agent will be improved based on the feedback from the experimenter.

In the next stage, children will interact with a supervised dialogue agent. This means the dialogue agent will suggest an utterance that will be send to the participant only if the experimenter (which is again an online counselor) approved it. Additionally, if the experimenter does not approve of the suggested utterance, she can send a custom message to the participant (just as she normally does during counseling via chat). The participant will be asked to complete the questionnaires as described previously. In addition, we will take into account the number of human interventions. Finally, feedback from the experimenter is gathered: what went well and what does still need improvement? This experiment will be repeated with new participants and updated versions of the dialogue agent, until the number of interventions is acceptably low (what is acceptable will be discussed with the experts).

6.3 Involving Cyberbullying Victims

In final stage, the previous two stages are repeated, but the dialogue agent responds to actual experiences of cyberbullying victims. First, victims interact with a WOZ and after successful completion of that stage, victims interact with the supervised dialogue agent, so the experimenter can intervene at any moment. Performance is measured with the questionnaires, the number of human interventions and feedback from the experimenter.

If in this stage of the evaluation the number of human interventions is acceptably low (again, what is acceptable will be discussed with the experts) and if scores on the social support questionnaire, scores for well-being and perceived burden of the problem are close to scores obtained by human counselors the evaluation of the dialogue agent is complete. If the embodiment has been evaluated successfully, we can move on to the evaluation of the complete anti-cyberbullying agent.

7. DISCUSSION AND CONCLUSION

In this paper, we specified 5 verbal social support types: sympathy, compliment, encouragement, advice, and teaching and, inspired by a model for small talk, implemented these in a BDI dialogue agent. The dialogue agent structures the conversation according to the 5-phase model: in phase 1, the agent welcomes the user; in phase 2, the agent gathers information about the incident; phase 3 (determine objective of the conversation) has a trivial implementation in which the conversation objective is always 'get advice on how to deal with cyberbullying'; in phase, 4 the agent gives advice; and in phase 5, the conversation is rounded off. Sympathy, compliment and encouragement are always given in response to user input. Advice and teaching are offered pro-actively.

Additionally, we presented an evaluation method for the dialogue agent. Because cyberbullying is a sensitive topic and children are a vulnerable target audience, we will start with an expert evaluation and create scenarios of common cyberbullying situations for indirect evaluation. After multiple experiments and incremental improvements on the dialogue agent we intend to involve children in the evaluation process. Experiments in which children participate will be conducted always in cooperation with and under the supervision of psychologists and online counselors. Performance of the dialogue agent will be measured with questionnaires on perceived social support and trustworthiness of the agent.

Braithwaite's typology of social support contains more support types that can be implemented in the dialogue agent. In particular empathy is relevant for the anti-cyberbullying agent, because being empathic is important in supportive communication [7]. To appear empathic, the agent needs the capability to reason about emotions. Therefore an emotional module will be added to the anti-cyberbullying agent (see figure 1). We also plan to extend the dialogue agent with additional conversation techniques online counselors use to actively manage conversations, including requesting feedback (e.g., *Is that right?*), summarizing (e.g., *So, you are being bullied in school and via msn and you haven't told anybody because you are embarrassed?*), and verbalizing feelings (e.g., *You sound disappointed, are you?*).

The dialogue agent specified in this paper is part of an embodied agent. The embodiment is currently under development and will allow the anti-cyberbullying agent to give non-verbal feedback in addition to verbal feedback. The non-verbal channel will be mainly used for the expression of (empathic) emotions. Related work on empathic agents shows that text-only agents are outperformed by embodied agents [1, 13, 19]. Therefore, we expect the perceived social support will increase when a virtual character displaying appropriate emotional expressions is added to the system's interface.

The anti-cyberbullying agent is an application that addresses a real world problem. We would like to emphasize that a lot more than satisfactory performance in laboratory experiments is needed before the application can be introduced into a real world setting. Many additional criteria play a part in the feasibility and acceptability of software applications, such as the protection of privacy and other ethical and legal issues. At the very least the anti-cyberbullying agent should be able to detect and deal with cases it can not handle, either by referring the user to a specialized helpline, or call in a human counselor that takes over the conversation.

8. ACKNOWLEDGEMENTS

This work is funded by NWO under the Responsible Innovation (RI) program via the project 'Empowering and Protecting Children and Adolescents Against Cyberbullying'.

9. REFERENCES

[1] D.C. Berry, L.T. Butler, and F. de Rosis. Evaluating a realistic agent in an advice-giving task. *International Journal of Human-Computer Studies*, 63(3):304–327, 2005.

[2] A. de Beyn. *In gesprek met kinderen:de methodiek van de kindertelefoon*. SWP, 2003.

[3] T. Bickmore and J. Cassell. 'how about this weather?' social dialog with embodied conversational agents. In

Proceedings of the American Association for Artificial Intelligence (AAAI) Fall Symposium on "Narrative Intelligence", pages 4–8, 2000.

[4] D.O. Braithwaite, V.R. Waldron, and J. Finn. Communication of social support in computer-mediated groups for people with disabilities. *Health Communication*, 11(2):123–151, 1999.

[5] M. Cavazza, C. Smith, D. Charlton, N. Crook, J. Boye, S. Pulman, K. Moilanen, D. Pizzi, R. de la Camara, and M. Turunen. Persuasive dialogue based on a narrative theory: An eca implementation. In T. Ploug, P. Hasle, and H. Oinas-Kukkonen, editors, *Persuasive Technology*, volume 6137 of *Lecture Notes in Computer Science*, pages 250–261. Springer Berlin / Heidelberg, 2010.

[6] P.R. Cohen and H.J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2-3):213 – 261, 1990.

[7] H. Cowie and P. Wallace. *Peer Support in Action: From Bystanding to Standing By*. Sage Publications Ltd, 2000.

[8] S. D'Mello, B. Lehman, J. Sullins, R. Daigle, R. Combs, K. Vogt, L. Perkins, and A. Graesser. *Intelligent Tutoring Systems*, volume 6094 of *Lecture Notes in Computer Science*, chapter A Time for Emoting: When Affect-Sensitivity Is and Isn't Effective at Promoting Deep Learning, pages 245–254. Springer Berlin / Heidelberg, 2010.

[9] Foundation for Intelligent Physical Agents. Fipa communicative act library specification. http://www.fipa.org/specs/fipa00037/SC00037J.html, 2002.

[10] R. Fukkink. Peer counseling in an online chat service: A content analysis of social support. *Cyberpsychology, Behavior, and Social Networking*, 14(4):247–251, 2011.

[11] R. Fukkink and J. Hermanns. Counseling children at a helpline: chatting or calling? *Journal of Community Psychology*, 37(8):939–948, 2009.

[12] K.V. Hindriks. Programming Rational Agents in GOAL. In A. El Fallah Seghrouchni, J. Dix, M. Dastani, and R.H. Bordini, editors, *Multi-Agent Programming:Languages, Tools and Applications*, volume 2, pages 119–157. Springer US, 2009.

[13] K. Hone. Empathic agents to reduce user frustration: The effects of varying agent characteristics. *Interact. Comput.*, 18(2):227–245, 2006.

[14] T. Klüwer. "I Like Your Shirt" - Dialogue Acts for Enabling Social Talk in Conversational Agents. In H. Vilhjálmsson, S. Kopp, S. Marsella, and K. Thórisson, editors, *Intelligent Virtual Agents*, volume 6895 of *Lecture Notes in Computer Science*, pages 14–27. Springer Berlin / Heidelberg, 2011.

[15] S. Kopp, L. Gesellensetter, N.C. Krämer, and I. Wachsmuth. *Intelligent Virtual Agents*, chapter A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application, pages 329–343. 2005.

[16] T.-Y. Lee, C.-W. Chang, and G.-D. Chen. Building an interactive caring agent for students in computer-based learning environments. In *Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on*, pages 300–304, 2007.

[17] Q. Li. New bottle but old wine: A research of cyberbullying in schools. *Computers in Human Behavior*, 23(4):1777–1791, 2007.

[18] S. Livingstone, L. Haddon, A. Görzig, and K. Ólafsson. Risks and safety on the internet: The perspective of european children. initial findings. http://www2.lse.ac.uk/media@lse/research/EUKids Online/EUKidsII%20%282009-11%29/home.aspx, 2010.

[19] R. Looije, M.A. Neerincx, and V. de Lange. Children's responses and opinion on three bots that motivate, educate and play. *Journal of Physical Agents*, 2(2):13, 2008.

[20] J.W. Patchin and S. Hinduja. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice*, 4(2):148–169, 2006.

[21] C. Pelachaud, V. Carofiglio, B. De Carolis, F. de Rosis, and I. Poggi. Embodied contextual agent in information delivering application. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, AAMAS '02, pages 758–765, New York, NY, USA, 2002. Acm.

[22] K.P. Schneider. *Small Talk: Analyzing Phatic Discourse*. Marburg: Hitzeroth, 1988.

[23] C. Smith, N. Crook, J. Boye, D. Charlton, S. Dobnik, D. Pizzi, M. Cavazza, S. Pulman, R. de la Camara, and M. Turunen. Interaction strategies for an affective conversational agent. In Jan Allbeck, Norman Badler, Timothy Bickmore, Catherine Pelachaud, and Alla Safonova, editors, *Intelligent Virtual Agents*, volume 6356 of *Lecture Notes in Computer Science*, pages 301–314. Springer Berlin / Heidelberg, 2010.

[24] B.R. Steunebrink, N.L. Vergunst, C.P. Mol, F.P.M. Dignum, M.M. Dastani, and J.-J.C. Meyer. A generic architecture for a companion robot. In J. Filipe, J.A. Cetto, and J.-L. Ferrier, editors, *Proc. 5th Int. Conf. on Informatics in Control, Automation and Robotics (ICINCO'08)*, pages 315–321, 2008.

[25] N.L. Vergunst. *BDI-based Generation of Robust Task-Oriented Dialogues*. PhD thesis, Utrecht University, 2011.

[26] K. Zakharov, A. Mitrovic, and L. Johnston. *Intelligent Tutoring Systems*, volume 5091 of *Lecture Notes in Computer Science*, chapter Towards Emotionally-Intelligent Pedagogical Agents, pages 19–28. Springer Berlin / Heidelberg, 2008.

[27] J.M. van der Zwaan, V. Dignum, and C.M. Jonker. Simulating peer support for victims of cyberbullying. In *Proceedings of the 22st Benelux Conference on Artificial Intelligence (BNAIC 2010)*, 2010.

Agent Communication for Believable Human-Like Interactions between Virtual Characters

Joost van Oijen
Utrecht University
PO Box 80.089, 3508 TB
Utrecht, the Netherlands
J.vanOijen@uu.nl

Frank Dignum
Utrecht University
PO Box 80.089, 3508 TB
Utrecht, the Netherlands
F.P.M.Dignum@uu.nl

ABSTRACT

Virtual characters in games or simulations are increasingly required to perform complex tasks in dynamic virtual environments. This includes the ability to communicate in a human-like manner with other characters or a human user. When applying agent technology to create autonomous, goal-directed characters, interactions have to be generated at runtime. In this paper we propose a model balancing efficient agent communication on one hand and believable realizations of human-like interactions on the other hand.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Intelligent Agents, Multiagent Systems*
; I.6.3 [**Simulation and Modeling**]: Applications

General Terms

Design, Human Factors

Keywords

Agent Communication, Intelligent Virtual Agents

1. INTRODUCTION

As the technology to create more realistic, complex and dynamic virtual environments advances, there is an increasing interest to create intelligent virtual agents (IVAs) to populate these environments for the purpose of games, simulations or training. The use of agent technology in the form of multi-agent systems (MASs) seems a good fit to realize the cognitive and decision-making aspects of an IVA. One of the problems one faces when applying a MAS to control the behavior of virtual characters is how to deal with agent communication in the MAS: agents now become embodied in a real-time virtual environment and have to communicate through the environment to simulate believable interactions. Additionally, MASs often do not have to deal with human-like aspects like emotions or empathy and thus standards developed for agent communication (e.g. FIPA) typically do not support other kinds of communicative intents besides performative acts (e.g. the ability to communicate an affective state or to associate emotion with a message).

In current commercial 3D video games or game-based training applications, human-like interaction between virtual characters has hardly been employed. When it is, it is often realized during so-called cut scenes or in specific situations that are known to occur by design (e.g. scripts). Here, the believability of the graphical and audible realization of an interaction can be of a reasonably good level (e.g. encompassing conversational gestures or emotional expressions). Since the dialog acts and context in which the interaction takes place are fully known beforehand, realization can be crafted in detail at design time.

Now when we turn to agent technology to design autonomous, goal-directed agents, the context in which they might communicate cannot be known beforehand. Hence, communicative behavior has to be generated dynamically at runtime and achieving the same level of believability becomes more difficult to realize. This requires fine-grained multimodal control over an agent's embodiment, believable perception of multimodal behavior in the environment and models for generating outgoing and processing incoming communicative intents.

The use of multi-agent systems to control virtual characters has been considered before [13] and successful attempts have been made to demonstrate its potential. Here, agents are usually integrated in a game engine using a custom developed connection between a specific game engine and MAS [5] or making use of available technologies allowing access to a certain game engine [1, 6]. Looking at the facilities for agents to exhibit human-like communicative abilities in these systems, they fall short on delivering the necessary interfaces for agents to express and perceive communicative behaviors, due to the limitations of the underlying intermediate software [1] or game engines they were dependent on.

In this paper we present design issues for realizing believable human-like communication between virtual agents situated cognitively in a MAS and physically in a virtual environment. A model is proposed to tackle these issues allowing agents to effectively communicate any intent at the cognitive level while realizing this in a believable manner at the physical level. By not restricting ourselves to specific intents or intent representations, we leave designers the choice to decide which type of signals agents should be allowed to communicate, whether they are speech acts (e.g. performatives), meta-conversational signals (e.g. turn-taking) or affective signals (e.g. emotional state). This flexibility allows agent designers to employ our model to deal with additional

Appears in: *Proceedings of the Workshop on Emotional and Empathic Agents, in the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, June, 4-8, 2012, Valencia, Spain.

human-like abilities (e.g. flexible interaction management or empathic processing). By discussing a use case scene we aim to show that our model provides an infrastructural basis for realizing the agent communication occurring in such a scene while maintaining a suitable balance between efficiency and believability.

The paper is organized as follows. In section 2 related research areas are discussed. Section 3 addresses issues for realizing human-like interactions using agent communication. A model is proposed in section 4, followed by an implementation and evaluation in sections 5 and 6. Finally, in section 7 and section 8 we conclude.

2. RELATED WORK

In MAS research there is an increasing need for more *open systems* in which agents are situated in more dynamic environments, carrying out complex interactions. In such systems, the situations in which communication can take place is not fully known beforehand and agents require more knowledge about the specific meaning or underlying goal of a message in order to properly deal with it. In [3], research directions in agent communication are presented concerning these issues. For example, the use of *social semantics* is discussed, ascribing meanings to messages based on social concepts such as commitments or conventions. Compared to human-like interactions in virtual worlds, similarities can be drawn: virtual worlds are becoming more complex and dynamic whereas human-like interactions are by definition flexible and full of social semantics. Therefore simulating human-like interactions with agents, the use of standards like FIPA ACL and corresponding fixed protocols is not enough. We really have to make use of rich communication semantics in line of what is discussed in [3].

Considering the simulation of human-like communication by virtual agents, this is the research focus of the ECA (Embodied Conversational Agent) community. Here, frameworks and computational models are proposed for a wide range of aspects of human-like communication (e.g. multimodal communicative behaviors [9, 11], conversation modeling [16] and topics dealing with personality or cultural factors in ECAs [19]. There are several reasons why this research is not always directly applicable for our purposes: the focus is often on interaction with a user situated in the real world where little research is found on human-like interactions between two virtual agents, especially on the perception side. Further, most research focuses on a single aspect and is rarely seen within the scope of a full agent architecture (except for specific instances like [8]). Last, their possible employment for use in real-time games is often not a priority such that important aspects like practicality and efficiency are less focused upon.

Finally we consider the work on the integration of multi-agent platforms or other decision-making systems in virtual environments and look at the communicative abilities of these virtual agents. In [5], the cognitive BDI-architecture of CoJACK was used to control characters in VBS2, a 3D training environment used in military domains. Pogamut [6] is designed as a mediation-layer between a game engine and an agent's decision-making system to bridge the "representational gap". In [18], the agent programming language GOAL was used to integrate BDI agents in the UT game engine using both Pogamut and EIS. The latter is a proposal for an environment interface standard for MAS agents and has been advertised for use in agent platforms including 2APL, Jadex or Jason [2]. In these systems agents have very poor communicative abilities, caused by the employed game engines which offer very limited facilities for expressing and perceiving communicative behaviors. Further, considering multi-agent platforms, this raises the question of how communication should be handled: i.e. when connected to a game engine, communication can be accomplished through the virtual environment. Does this make a platform's communication mechanism obsolete? Or in what situations should agents still use direct communication within the MAS? Such questions have not been addressed in related work.

3. CONCEPTUAL GAP

Imagine a scene from a game-based training application for firefighters where each virtual character is controlled by a fully autonomous agent: *"A fire has been reported in a residential home, thought to be uninhabited. A team of firefighters arrive at the scene. The team leader assesses the situation and calls for a command huddle. While the leader is giving out orders to each team member to attack the fire, an injured woman stumbles out of the burning house. She is in a panicked state and screams something to the firefighters while pointing to a window on the first floor. Because of an explosion occurring simultaneously, the fire fighters fail to hear the woman but realize something is wrong based on the woman's expressions and gestures. The team leader interrupts the huddle and rushes to the woman who explains that her child is still in the house. A police officer nearby overhears the woman and calls for medical services. Some bystanders get hold of this development and spread the information to others. Meanwhile, the team leader reassesses the new situation, returns to his team and gives out new orders to first save the child and then attack the fire."*

Now when we consider using MAS technology to control human-like characters as illustrated in the example scene, one has to bridge the inherent conceptual gap between typical agent communication in MASs and human-like communication by characters in virtual environments. In the remainder of this section we will discuss concrete issues for realizing human-like interactions using agent communication.

3.1 Issues at the Mind-Body Interface

The first category of issues relate to the technical issues of applying agent communication to simulate character interactions in virtual environments.

3.1.1 Embodiment

Simulating human-like communication, agents should not be allowed to communicate directly with each other within a MAS but resort to expressing communicative behavior through their embodiment, separating *what* is communicated by an agent's mind from *how* this is realized by its embodiment. One aspect involves the use of multiple modalities to express a communicative intent (e.g. speech, gestures, gazing, facial expressions, etc). E.g. referring to our example scene, when the team leader is giving out orders, he may accompany his verbal acts with gestural body movements clarifying the meaning of a task being ordered. Another aspect concerns the *choice* of behavior realization. A similar intent may be communicated many different ways depending on factors like personality, culture, interaction partners

or social setting. E.g. in our example, more introvert team members may use less expressive gestures during communication than other members.

The same aspects concern for the perception of intents: since intents are communicated using multimodal behavior, they also need to be perceived through the observation of this behavior, requiring an inference step to assign a meaning to the observed behavior. Where the expression of communicative intents has gotten a lot of attention in research on virtual humans (e.g. in [9]), the corresponding perception largely remains untackled for communication between virtual agents.

3.1.2 Environment

Unlike in typical MAS applications, agents required to represent human-like characters have to deal with a different kind of environment, namely a *real-time* and *virtual* environment. This introduces several issues.

First of all, unlike in typical MAS environments, actions now become durative and the successful execution of an act is not immediately known (e.g. speech can last a number of seconds to realize). Now both the environment and the agent's cognitive state may change during the realization of a communicative act and could result in a realization failure or a desire for the agent to interrupt the ongoing realization respectively (e.g. in our example scene, the team leader interrupts his communication when he hears the screaming woman). Further, the perception communicative behavior also becomes a durative process. Even though not fully perceived an intent, one may still require an agent to be aware of ongoing communicative behavior (e.g. for the purpose of situation awareness or providing backchannel feedback).

Second, where MASs usually provide a reliable communication mechanism for inter-agent communication, successful communication realized in a virtual environment depends on the sensory capabilities of the agents and the simulated laws of physics. For example, in our example scene, a bystander walking past the incident may not have perceived the screaming woman simply because she was out of sensory range. And although the team leader perceived the woman's nonverbal behavior, he did not fully understand her because the explosion distorted proper perception of her verbal message. Further, agents within the vicinity are able to overhear communication even if not directed towards them (e.g. the police officer overhearing the screaming woman).

The last issue concerns factors of success. Typically in MASs, semantics for communication success or failure are trivial: either the message was successfully delivered or not. In virtual agent communication, failure can occur at difference conceptual levels: an act was scheduled for realization but there was a problem to physically express it; an act was successfully realized but not perceived; or the act was perceived but not properly interpreted.

3.2 Issues on the Agent-Side

The second category of issues relate to more conceptual issues within the agent itself for simulating human-like interactions. Unlike the technical issues described above, these issues will not be tackled explicitly by our model proposed in the following section. Rather, we aim to summarize the aspects to consider in agent design and the impact it may have on application design.

Communicative Functions Communication in MASs typically involves the use of performative acts (e.g. FIPA ACL) to effectively allow agents to exchange information or delegate tasks. Signals communicated between virtual agents required to exhibit human-like communicative abilities are much richer. For example, in [14], a taxonomy of communicative functions is given for human communication and amongst others include functions related to conversation management (e.g. turn-taking), meta-cognitive signals, deictic references and emotional expressions. Developing agent frameworks supporting such functions is an active area of research (e.g. [7, 10]), though, they often impose strong requirements on the design of the agent. Considering them for use in real-time games, a tradeoff is in place between desired believability and design complexity.

Emotions Emphasizing a category of communicative functions are affective functions, mandatory for agents required to cope with aspects like emotion or empathy. Computational models and frameworks have been proposed based on theories of appraisal and emotion (e.g. [4, 12]). Since emotional factors may impact agent processes like belief formation, deliberation or intent realization, a more complex agent design is required. Employing such affective agents in games, the challenge is deciding why, when and what kind of emotional signals must be communicated between agents and how this can be realized.

Conversation Modeling In MASs, conversations between two or more agents are often regulated by fixed interaction protocols where each agent takes on a predefined role, either as the initiator or participant (e.g. FIPA *Query* or *Contract Net* protocols). Natural human-like conversations tend to be more flexible and dynamic: participants may take, request or give the turn at any point in time; they can join or leave a conversation any time and may take on different roles (e.g. side participant or overhearer). Human-like conversation modeling has been addressed in previous research on virtual humans (e.g. in [16]). The challenge here is to integrate such models in the deliberation process of a MAS agent coexisting with non-communicative behavioral models (e.g. BDI reasoning on an agent's task model).

Listening Behavior In MAS communication a message is either delivered as a whole or not at all. In virtual agent communication, performing an intent may take some time and for an addressee, it can look unnatural to restrain from expressing any behavior while the speaker is talking. Here, listening behaviors and backchannel feedback can be used for showing attention or for grounding purposes and are typically expressed as head nods, gaze behavior or short verbal utterances. Research is available proposing models for listening feedback (e.g. [21]). Although increasing believability, such models can add considerably in design complexity: proper listening feedback requires partial understanding of content being communicated.

4. A MIDDLEWARE APPROACH

We present a model for virtual agent communication employing a middleware approach to fill the gap between agent communication in a MAS and its realization in a virtual environment. It builds upon our previous effort of designing a middleware bridging the conceptual gap between agent and game engine technology [20].

In figure 1, our model is illustrated and addresses the is-

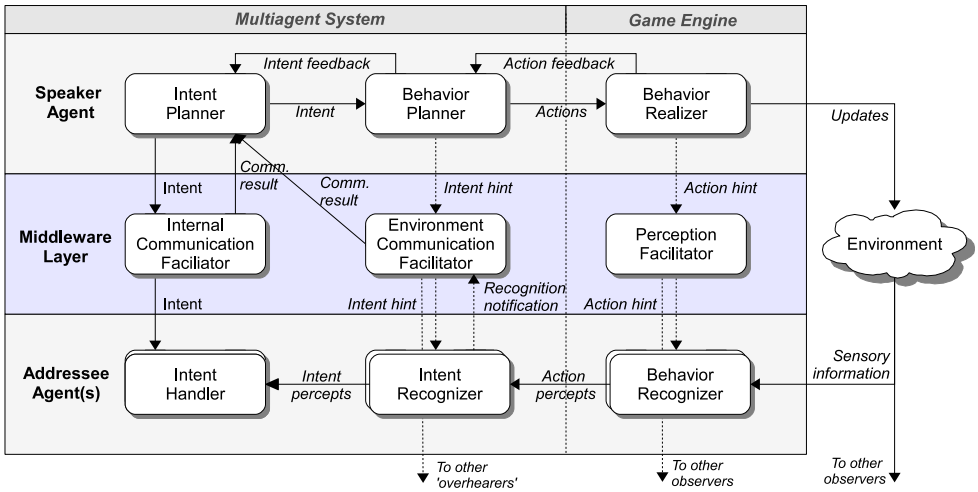


Figure 1: Communication Model

sues described in section 3.1. More concretely it deals with (1) separating *mind* and *body* allowing agents in a MAS to communicate with each other using multimodal communicative behaviors, (2) separating intent from behavior planning, and behavior from intent recognition respectively, allowing agents to express behaviors and interpret intents depending on contextual factors, (3) durative expression and observation of communicative intents, allowing agents to monitor and interrupt scheduled communication and (4) believable perception based on sensory capabilities and environment physics. It provides an infrastructural basis for dealing with agent-side aspects as presented in section 3.2. Next, we describe the model in more detail.

4.1 Communication Expression

The upper part of the model is responsible for realizing a communicative intent using multimodal behavior expressions. The stages shown are conceptually similar to the behavior generation stages part of the SAIBA framework [9], though, since our goal deviates from the SAIBA initiative, we do not focus on standard data representations used between the stages. First, the *Intent Planner* generates communicative intents a speaker agent wishes to express. Next, the *Behavior Planner* translates an incoming intent to a schedule of communicative actions where each action represents a single modality (e.g. a speech action, gesture or facial expression). Last, the *Behavior Realizer* executes communicative actions for realization in the game engine.

At each stage, feedback information about the progress of a realization is sent to previous stages allowing an agent to monitor the execution of its intent. Feedback about actions is used to determine the feedback to be generated for scheduled intents (e.g. started, finished, failed or aborted). Further, at any point in time, an agent may abort a scheduled intent resulting in the abortion of all scheduled actions.

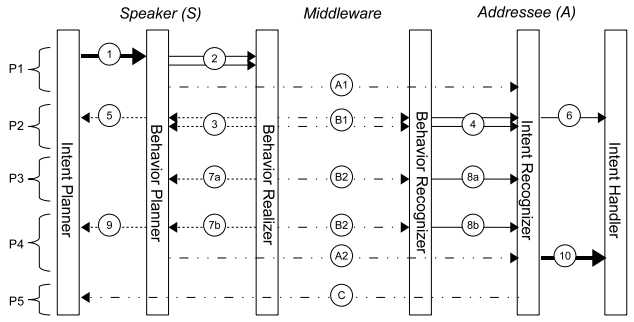
4.2 Communication Perception

Next, the lower part of the model deals with the perception of communicative intents by an addressee agent. Similar stages are identified as above. First, in reverse order, the *Behavior Recognizer* interprets communicative signals (actions) based on perceived sensory information from the environment. It represents a physical process where the ability to interpret signals is limited by the sensory capabilities of an agent. For example, an agent could recognize a *head nod* performed by a speaker based on head bone positions observed over time. Next, the *Intent Recognizer* assigns a meaning to 'recognized' signals, possibly representing the original intent the speaker agent tried to convey. It represents a cognitive inference process influenced by contextual factors. To give an example, different meanings can be assigned to an observed *head nod*. In some situations it could be interpreted as an *acknowledgement*, in other situations as a form of *greeting*. Last, the *Intent Handler* receives inferred intents for further processing.

4.3 Middleware Facilities

Looking at the software engineering aspects of the perception stages described above, both are computationally heavy processes and contribute to design complexity: the stage of *behavior recognition* requires observations over time to recognize communicative signals like speech (e.g. stream of sound waves) or gestures (e.g. motion of bones). The stage of *intent recognition* can be seen as a pattern matching problem where a set of multimodal communicative signals have to be matched to an intent (taking into account both the type and timing of signals). Although this approach results in a fully autonomous process for the perception of communicative intents, we believe it is not very practical to implement and is overly complex for use in real-time games. As an alternative, we propose a design approach employing a middleware layer to simplify the perception processes, making a tradeoff between efficiency and believability.

Since the data representations for communicative intents and actions that need to be recognized are already available within the speaker agent, we propose to employ this information during the corresponding perception of these actions and intents. First, the *Perception Facilitator* allows agents to perceive communicative actions directly. It simplifies the process of behavior recognition where actions do not have to be interpreted from sensory information. Instead, it is reduced to a query whether an action that was just expressed can be perceived by an observer based on its current sensory capabilities. Here, the middleware provides observing agents



#	Agent	Activity	Message
1	S	schedule intent	*communicate(id=i1,content=inform_child_in_house)*
2	S	schedule action	*speech(id=a1,resource=child_in_house.mp3)*
3	S	receive action feedback	*action_feedback(action=a1,state=started)*
4	A	perceive action	*action_percept(action=a1,state=started)*
5	S	receive intent feedback	*intent_feedback(id=i1,state=started)*
6	A	perceive intent (no content)	*intent_percept(state=started)*
7	S	receive action feedback	*action_feedback(action=a1,state=finsihed)*
8	A	perceive action	*action_percept(action=a1,state=finished)*
9	S	receive intent feedback	*intent_percept(id=i1state=finished)*
10	A	perceive intent	*intent_percept(id=i1,state=ended)*

#	Middleware	Activity	Message
A1	Comm. Facilitator	send intent hint	*intent_hint(intent=i1, actions=a1;a2)(state=started)*
B1	Perception Facilitator	send action hint	*action_percept(action=a1,state=started)*
B2	Perception Facilitator	send action hint	*action_percept(action=a1,state=finished)*
A2	Comm. Facilitator	send intent hint	*intent_hint(intent=i1)(state=finished)*
C	Comm. Facilitator	send communication result	*communication_result(id=i1,observed_by=A)*

Figure 2: Communication Example

with *action hints* which they can use to create percepts (after successfully passing the query). Next, the *Environment Communication Facilitator* facilitates the process of intent recognition within an observing agent by providing a *hint* about the communicative intent currently being expressed by a speaker agent. This hint not only contains the original intent, but also the actions used by the speaker to realize this intent. This reduces the problem of pattern matching to a matter of comparing recognized actions to *expected* actions where the corresponding *expected* intent can be immediately inferred. With this approach, perception can be performed efficiently, though still in a believable manner bounded by environment physics. Also, agents not only perceive the end of an action or intent respectively, but also the beginning, allowing an agent to recognize an intent being communicated *while* the speaker is expressing it (though without full semantics for believability).

To clarify the communication process in our model, figure 2 illustrates the successful communication of a single communicative intent, realized using multimodal behavior consisting of two actions. Note that the focus of the model is on the semantics of the communicated data and not on specific data representations (shown messages have been simplified and *ids* are used to denote a corresponding intent or action). The upper table left of the diagram illustrates information being communicated between components *within* the speaker and addressee agent in a time-ordered fashion. The bottom table shows information being send *between* the agents using the middleware's facilities. Referring to the diagram, in phase 1 (P1), the speaker schedules the intent along with the realization actions. In phase 2, the speaker receives feedback stating the realization of the intent has started while the addressee perceives the beginning of the intent (though without actual content). Phase 3 represents the ongoing process of expressing and recognizing actions. In phase 4, the speaker receives feedback about the successful completion of the intent while the addressee perceives the full intent. Finally, phase 5 provides the speaker with feedback about the successful perception of an intent by the addressee. This last phase is optional and is explained next.

4.3.1 Successful Communication

The success of a non-communicative action can be validated in the physical environment inside the game engine. For example, the success of an action like *open door* can be checked based on the values of certain game state parameters (e.g. status property of the door). But determining whether a communicative intent was successfully communicated cannot easily be done and depends on whether the intent was successfully perceived and interpreted. This would require inspection of agent parameters which are not externally accessible by the speaker agent.

To support an agent in reasoning about the success or failure of communication, in our model we provide feedback about the success or failure of the delivery of an intent to the addressee(s) (i.e. if the corresponding communicative behavior was perceived and interpreted correctly as the original intent). This facility is provided by the *Environment Communication Facilitator*. After successful execution of a communicative intent, this component will inform the speaker agent whether its message has been properly received and recognized and by which addressees. It accomplishes this based on received *recognition notifications* sent by the *Intent Recognizer* from addressee agents.

4.3.2 Direct Communication

One can think of situations where agents may require communication to exchange information or coordinate their actions but where it is not relevant for a human user to notice this during game play. In this situation, instead of realizing this in the environment, direct communication may be more efficient. Our model provides an *Internal Communication Facilitator* allowing agents to send messages directly to any other agent within the MAS. It fulfills the same task as the typical communication mechanism in an agent platform.

4.4 Discussion

Comparing our model to typical agent communication like FIPA ACL, the main difference can be found in the lower-level protocols and medium used to communicate. Where FIPA communication deals with communication over a network medium using a protocol like TCP/IP, communication between virtual agents requires a more complex medium that deals with (1) the cognitive abilities of agents to express and interpret intents, (2) the physical abilities of agents to express and perceive behavior (through actuators and sensors) and (3) a transportation medium represented by a virtual environment. Looking at the protocol from figure 2, FIPA would merely cover line 1 and 10: sending and receiving communicative intents. All the other lines can be seen as a necessary extension to achieve proper realization, efficient interpretation and believable transportation. A concrete application of this protocol is outlined in the following sections.

5. IMPLEMENTATION

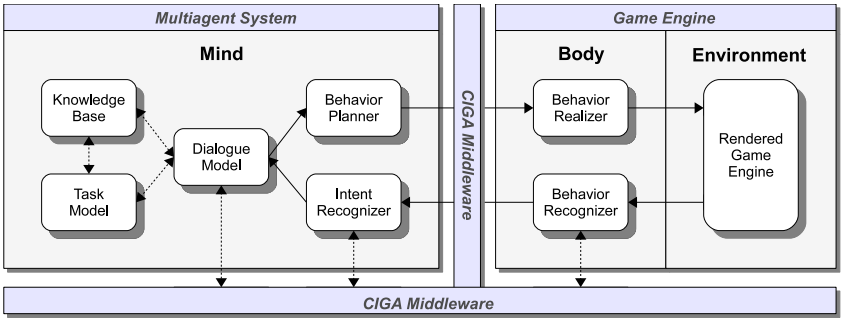


Figure 3: Agent Architecture in System Design

In this section we discuss the implementation of a full system design that will be used for further evaluation. It shows a possible interpretation of the communication model from figure 1. The design is illustrated in figure 3 and is made up of a MAS and a game engine, coupled by a middleware. It focuses on implemented agent components related to communication (i.e. not a full agent architecture). Current implementation does not support the communication of signals other than performative acts (e.g. no turn-taking, backchanneling or emotions), though this is not a necessity for a proper evaluation. Rather, it should make clear how the proposed communication infrastructure supports the use of additional communicative functions.

As for the MAS and game engine, in-house developed systems have been employed [1]. Further, the system includes a middleware (named *CIGA*) which has been developed as a generic solution to facilitate the coupling between a MAS and game engine. The middleware layer proposed in the communication model has been integrated within this middleware. A more elaborate description of CIGA and its motivation falls outside the scope of this paper and can be found in [20]. Below, the agent's components in its cognitive and physical layer are described shortly.

5.1 Cognitive Layer

[**Knowledge Base**] A storage for propositions that can be accessed during deliberation. Propositions can be created from perceived sensory information (not shown in design) or received information through communication.

[**Task Model**] Deliberation and decision-making rules for non-communicative behavior. Rules according to the BDI-paradigm can be implemented within a task hierarchy representing a certain role for an agent. We have previously experimented with the reasoning engines Jadex and 2APL, though currently a behavior tree implementation is employed which suffices for our evaluation.

[**Dialogue Model**] Deliberation and decision-making rules for communicative behavior. Both the *Intent Planner* and *Intent Handler* from figure 1 are included in this model. To support flexible interactions, the model was implemented as an information state-based dialogue system, inspired by the theory in [17]. To give an example, an incoming dialogue act updates the model's information state (e.g. an obligation to address an act). At the next deliberation cycle the new state is inspected to determine the next dialogue move to perform. This allows flexible interactions not based on specific protocols. Currently dialogue moves are supported for conversation management and for a limited set of core speech acts (inform, enquiry and order).

[**Behavior Planner**] Realizes a communicative intent scheduled by the *Dialogue Model*. An intent is represented by a dialogue act with one or more intended receivers. Custom defined *mapping rules* are used to map an intent to a schedule of actions including instructions for different modalities. For example, a *greeting* could be mapped to an approach behavior, including speech, gaze and an appropriate gesture. Rules can be based on context variables covering aspects like cultural background, relationship with the interlocutor or the current social setting. During realization, feedback information is sent to the *Dialogue Model* whenever an intent was started, finished, aborted or failed its realization (e.g. an intent is started as soon as the first corresponding action is started). A scheduled intent can be aborted at any time and results in the abortion of all corresponding actions that have been scheduled.

[**Intent Recognizer**] Manages intent observations based on *action percepts* received from the *Behavior Recognizer*. From the middleware, this component receives information about the intent being communicated together with the actions used for its realization (the *intent hint*). Based on this information together with the received action percepts, it can determine the progress of an intent observation. *Intent percepts* are then sent to the *Dialogue Model* whenever an intent observation was started or has fully been recognized. In the latter case, the original dialogue act expressed by the actor is included in the percept. In this way an agent could perform a certain listening behavior like gazing knowing its interlocutor has started expressing an intent. Next, after being informed about the intent's full recognition, the agent's *Dialogue Model* can decide on a next course of action.

5.2 Physical Layer

[**Behavior Realizer**] Realizes scheduled communicative actions. Parameterized actions have been defined as control instructions for individual modalities. For concrete action implementations, game engine functionality is used to control and monitor the realization within the virtual character. For example, a gesture expression requires access to the animation engine, gazing requires specific bone control while locomotion requires path finding and collision avoidance. During execution, feedback information about progress is sent to the agent's mind.

[1] www.vstep.nl

[**Behavior Recognizer**] Manages action observations from actions expressed by other agents. This component is dependent on the middleware from whom it receives information about actions expressed by other agents as *action hints*. Upon receival, it checks whether or not the specific action can be perceived. For example, for a speech action this involves querying if the corresponding sound in the environment is observable based on the agent's auditory sensor, loudness and distance towards the source and possible interferences. If observable, an *action percept* containing the action's original representation and its current progress (e.g. started or ended) is generated and sent upstream to the *Intent Recognizer* for further processing.

6. EVALUATION

We evaluate the communication model at the functional level by realizing a set of scenarios for different interactional situations. Based on the implemented system described in the previous section, a simple base scenario has been developed in which the required functionality can be demonstrated. The base scenario concerns a two-turn conversation between two IVAs where small variations to this scenario are run for testing different aspects of the model.

Below we list the requirements that will be evaluated for the implemented system and correspond to the issues outlined in section 3.1.

1. Context-dependent multimodal behavior expressions
2. Multimodal behavior perception
3. Monitoring durative intent realizations
4. Monitoring durative observation of communicative intents
5. Interruption of scheduled communication
6. Believable perception based on sensory capabilities and environment physics

An impression of the implemented base scenario and its variations is illustrated in figure 4. The base scenario demonstrates a successful conversation covering requirements 1 through 4; variation (a) demonstrates interruption of communication (requirement 5); remaining variations relate to requirement 6 and demonstrate failed communication because of no perception (out of range) or partial perception (variation b and c respectively) concluding with overhearing of communication by bystanders (variation d). Due to space limitations, below we only describe the base scenario and variation (a).

Base scenario: An agent starts a conversation with a passerby agent and asks for the current time. The participant answers by giving the time after which they both terminate the conversation and resume their way. This base scenario illustrates a primitive successful conversation. From the communication model, it covers all stages for behavior generation and recognition. Middleware facilitators allow for efficient behavior and intent recognition for the initiator and participant where a speaker is notified about the successful delivery of its intent.

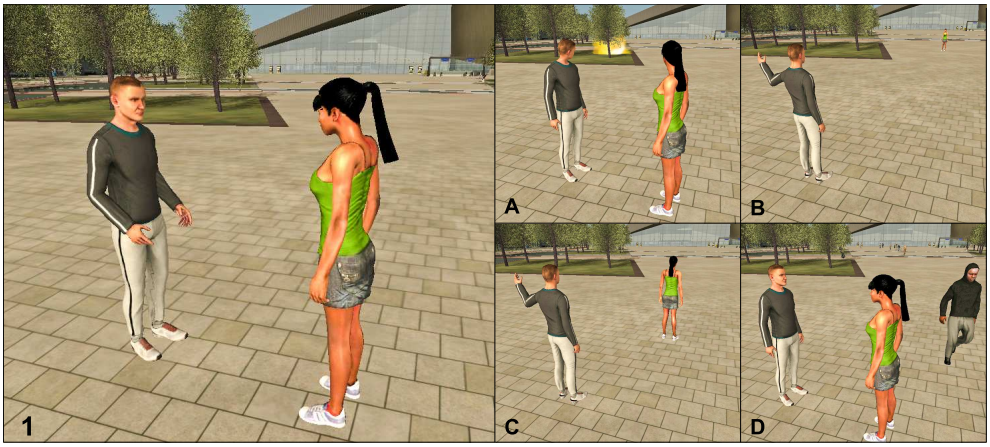
Variation a: The same situation is simulated though here while the initiator is asking for the time, both agents hear an explosion and notice a fire starting. The current speaker interrupts its current communicative intent while the addressee interrupts its listening behavior. Both agents end the conversation implicitly and pursue a new goal to deal with the situation. This variation illustrates agents coping with external events in the middle of a dialogue. Referring to the communication model, the speaker aborts its intent realization (and therefore behavior realization) from the *Intent Planner*; correspondingly the addressee's *Intent Handler* is informed that the intent was not fully perceived (i.e. was aborted). In the implemented system, the dialogue model (encompassing both the above components), at both agents, decides to end the conversation caused by a higher priority goal originating from the task model.

The base scenario and its variations have been realized successfully in our system. Although the involved agent components were implemented in an ad-hoc manner based on simple rules and policies, it suffices in demonstrating the basic principles of our model. More complex scenarios can be created based on the same principles but using more complex rules and policies. These may support for example: the use of more complex (multi-party) dialogues; the use of a richer context for expressing an intent; dealing with more believable ways of reacting on partial intent observations; or a more generic and context dependent way of decision-making for overheared communications. Often such aspects relate to different research areas as mentioned in section 3.2.

7. DISCUSSION

In this paper we have focused mainly on directed communication of dialogue acts between two or more agents. However, as we have already shortly mentioned in section 3.2, there are more types of meanings that human-like agents could communicate through signals, concerning information on the speaker's mind like beliefs, goals or emotions [15]. For example, through specific verbal or nonverbal signals information can be conveyed about the speaker's meta-cognitive or affective state. Our communication model does not restrict one to use specific types of meaning involved in communication. As long as they can be expressed and observed through the agents actuators and sensors, they can be represented as a communicative intent and processed by the middleware's facilitators. This allows designers to develop their own intents geared towards their specific needs. For example, for agents required to be emotional or empathic one could design affective signals to be communicated. Such signals could then be accompanied by a directed dialogue act, but also used as undirected intents (i.e. where there is no specific addressee). E.g. one might use undirected intents to support leaked emotions or 'communicate' an agent's mood expressed through certain postures or facial expressions. The ability to observe such signals would be helpful as input for an agent's empathic processing.

Now when we consider using our model to communicate intents not directed towards a specific agent, this raises the question of whether the model could also be used to efficiently 'communicate' non-communicative intents and if this would be desirable from an conceptual point of view. Knowing another's agent's intent could ease the realization of certain social behaviours. To give an example, consider an agent walking towards a door with the intent to open it. Another agent observing this intent could assist the agent by



1 successful communication; (a) interruption; (b) out of range; (c) partial observation; (d) overhearing

Figure 4: Scenario Impressions

informing it that the door is locked. The possible advantages and disadvantages of using the model for non-communicative intents is currently being investigated.

8. CONCLUSION

In this paper we proposed a design approach for modeling agent communication in a MAS to be represented in a human-like manner in a game engine. We focussed on the benefits of employing a middleware layer to facilitate perception and decision-making aspects involved in communication. The middleware layer allows IVAs to (1) communicate intents efficiently at the cognitive level on the MAS side and (2) realize this at the physical level in the game engine through the expression and perception of multimodal communicative behaviors. This is accomplished by the middleware's communication protocol which couples the cognitive and physical channels of communication between sender and receiver agents. Here, the perception stages of receiver agents do not require fully autonomous processes for recognizing communicative actions and intents (which are computationally expensive). Further, decision-making in dialogues can be handled more efficiently based on the acquired knowledge of the success or failure of communication (provided to sender agents by the middleware layer). Although requiring a more complex protocol for agents to adhere to (compared to FIPA), it does not enforce any specific implementation for any involved agent component. Nor does it enforce any specific data representation for communicative intents and actions used between agent components or channeled between agents themselves.

We believe with this more practical approach, one can achieve a proper balance between believability and efficiency for simulating human-like interactions (e.g. suitable for real-time games). The proposed model provides an infrastructure one can build upon to implement additional aspects of human-like communication like described in section 3.2. It therefore provides a first stepping stone to realize the example scene described in the beginning of this paper.

9. REFERENCES

[1] R. Adobbati, A. N. Marshall, A. Scholer, and S. Tejada. Gamebots: A 3d virtual world test-bed for multi-agent research. In *In Proceedings of the Second International Workshop on Infrastructure for Agents, MAS, and Scalable MAS*, 2001.

[2] T. Behrens, K. Hindriks, and J. Dix. Towards an environment interface standard for agent platforms. *Annals of Mathematics and Artificial Intelligence*, pages 1–35, 2010.

[3] A. K. Chopra, A. Artikis, J. Bentahar, M. Colombetti, F. Dignum, N. Fornara, A. J. I. Jones, M. P. Singh, and P. Yolum. Research directions in agent communication. *ACM Transactions on Intelligent Systems and Technology*, To Appear.

[4] J. Dias, S. Masceranhas, and A. Paiva. Fatima modular: Towards an agent architecture with a generic appraisal framework. In *Proceedings of the International Workshop on Standards for Emotion Modeling*, 2011.

[5] R. Evertsz, M. Pedrotti, P. Busetta, H. Acar, and F. Ritter. Populating VBS2 with realistic virtual actors. In *Proceedings of the 18th conference on Behavior Representation in Modeling and Simulation*, pages 1–8, 2009.

[6] J. Gemrot, C. Brom, and T. Plch. A periphery of pogamut: From bots to agents and back again. In F. Dignum, editor, *Agents for Games and Simulations II*, volume 6525 of *LNCS*, pages 19–37. 2011.

[7] D. Heylen, S. Kopp, S. C. Marsella, C. Pelachaud, and H. Vilhjálmsson. The next step towards a function markup language. In *Proceedings of the 8th international conference on Intelligent Virtual Agents*, pages 270–280. Springer-Verlag, 2008.

[8] P. Kenny, A. Hartholt, J. Gratch, W. Swartout, D. Traum, S. Marsella, and D. Piepol. Building interactive virtual humans for training environments. In *The I/ITSEC*, volume 2007, 2007.

[9] S. Kopp, B. Krenn, S. Marsella, A. Marshall, C. Pelachaud, H. Pirker, K. Thórisson, and H. Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In J. Gratch and et al., editors, *Intelligent Virtual Agents*, LNCS, pages 205–217. 2006.

[10] J. Lee, D. DeVault, S. Marsella, and D. Traum. Thoughts on FML: Behavior generation in the virtual human communication architecture. In *Proceedings of The 1st Functional Markup Language Workshop*, 2008.

[11] J. Lee and S. Marsella. Nonverbal behavior generator for embodied conversational agents. In J. Gratch and et al., editors, *Intelligent Virtual Agents*, volume 4133 of *LNCS*, pages 243–255. Springer Berlin, 2006.

[12] S. C. Marsella and J. Gratch. EMA: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70 – 90, 2009.

[13] E. Norling and L. Sonenberg. Creating interactive characters with BDI agents. In *IE2004: Proceedings of the Australian Workshop on Interactive Entertainment*, pages 69–76, 2004.

[14] I. Poggi. Mind markers. In I. N. T. M. Rector, Poggi, editor, *Gestures, Meaning and use*. University Fernando Pessoa Press, Oporto, Portugal, 2003.

[15] I. Poggi, C. Pelachaud, F. Rosis, V. Carofiglio, and B. Carolis. Greta. a believable embodied conversational agent. In O. Stock, M. Zancanaro, and N. Ide, editors, *Multimodal Intelligent Information Presentation*, volume 27 of *Text, Speech and Language Technology*, pages 3–25. Springer Netherlands, 2005.

[16] D. Traum and J. Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of AAMAS'02*, pages 766–773, 2002.

[17] D. Traum, W. Swartout, J. Gratch, and S. Marsella. A virtual human dialogue model for non-team interaction. In L. Dybkjær and et al., editors, *Recent Trends in Discourse and Dialogue*, volume 39, pages 45–67. Springer Netherlands, 2008.

[18] K. v Hindriks, B. van Riemsdijk, T. Behrens, R. Korstanje, N. Kraayenbrink, W. Pasman, and L. de Rijk. Unreal GOAL Bots. In F. Dignum, editor, *Agents for Games and Simulations II*, LNCS, pages 1–18. Springer-Verlag, Berlin, 2011.

[19] M. Vala, G. Blanco, and A. Paiva. Providing gender to embodied conversational agents. In H. Vilhjálmsson and et al., editors, *Intelligent Virtual Agents*, volume 6895 of *LNCS*, pages 148–154. Springer, 2011.

[20] J. van Oijen, L. Vanhée, and F. Dignum. CIGA: A Middleware for Intelligent Agents in Virtual Environments. In *Proceedings of the 3rd International Workshop on the uses of Agents for Education, Games and Simulations*, 2011.

[21] Z. Wang, J. Lee, and S. Marsella. Towards more comprehensive listening behavior: Beyond the bobble head. In H. Vilhjálmsson, S. Kopp, S. Marsella, and K. Thórisson, editors, *Intelligent Virtual Agents*, volume 6895 of *LNCS*, pages 216–227. 2011.

Preliminary Exploration of Agent-Human Emotional Contagion via Static Expressions

Jason Tsai[1], Emma Bowring[2], Stacy Marsella[3], Wendy Wood[1] Milind Tambe[1]

[1]University of Southern California, Los Angeles, CA 90089
{jasontts, wendy.wood, tambe} @usc.edu
[2]University of the Pacific, Stockton, CA 95211
ebowring@pacific.edu
[3]USC Institute for Creative Technologies, Playa Vista, CA 90094
marsella@ict.usc.edu

ABSTRACT

In social psychology, emotional contagion describes the widely observed phenomenon of one person's emotions mimicking surrounding people's emotions [13]. While it has been observed in human-human interactions, no known studies have examined its existence in agent-human interactions. As virtual characters make their way into high-risk, high-impact applications such as psychotherapy and military training with increasing frequency, the emotional impact of the agents' expressions must be accurately understood to avoid undesirable repercussions.

In this paper, we perform a battery of experiments to explore the existence of agent-human emotional contagion. The first study is a between-subjects design, wherein subjects were shown an image of a character's face with either a neutral or happy expression. Findings indicate that even a still image induces a very strong increase in self-reported happiness between Neutral and Happy conditions with all characters tested and, to our knowledge, is the first ever study explicitly showing emotional contagion from a virtual agent to a human. We also examine the effects of participant gender, participant ethnicity, character attractiveness, and perceived character happiness and find that only perceived character happiness has a substantial impact on emotional contagion.

In a second study, we examine the effect of a virtual character's presence in a strategic situation by presenting subjects with a modernized Stag Hunt game. Our experiments show that the contagion effect is substantially dampened and does not cause a consistent impact on behavior. A third study explores the impact of the strategic decision within the Stag Hunt and conducts the same experiment using a description of the same strategic situation with the decision already made. We find that the emotional impact returns again, particularly for women, implying that the contagion effect is substantially lessened in the presence of a strategic decision.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence–*Intelligent agents*

Appears in: *Proceedings of the Workshop on Emotional and Empathic Agents, in the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, June, 4-8, 2012, Valencia, Spain.

General Terms

Human Factors

Keywords

Virtual Agents, Emotional Contagion, Social Influence

1. INTRODUCTION

Emotional contagion is defined as the tendency to catch the emotions of other people [13]. While initial work focused on documenting its existence, recent research has moved to understanding its impacts on everyday life. In the workplace, researchers have examined its influence on promoting employee efficiency and client happiness [11, 22]. Research in administrative sciences has shown emotional contagion to improve cooperation, decrease conflict, and increase perceived task performance in groups and organizations [2]. Small et al. have shown substantial impacts on charitable donation amounts with only a still image [25]. Though its effects are often felt, in-depth understanding of emotional contagion remains an open area of research.

A variety of hypotheses regarding factors that influence emotional contagion have been explored in social psychology. A popular one examines differences in the strength of emotional contagion felt by men and women, with many researchers finding that women are significantly more responsive to emotional contagion than men [8, 27]. Researchers have also found that contagion increases in cases where the subject shares the same ethnicity as the stimulus [8] and when the expression is stronger [30]. Finally, attraction to the stimulus has been shown to have a positive effect on the contagion experienced by subjects [27].

The vast majority of emotional contagion research, however, has come from the social sciences and examines the spread of emotions from humans to other humans. Emotional contagion's impact in virtual agents' interactions with humans, however, is a largely untouched area of research. Specifically, while many researchers have worked to understand immersion, rapport, and influence in other contexts [12, 18], far fewer have looked into the emotional impact that the mere presence of virtual character emotions can have on people. The effects are assumed to either be nonexistent and therefore overlooked entirely or to mimic human-human emotional influences. However, as this work demonstrates, these are both poor assumptions to make and can be harmful to users in sensitive domains. As virtual agents enter high-risk and emotionally delicate applications such as virtual psychotherapy [23, 24], for example,

researchers must be cognizant of all potential emotional influences characters can have on users.

Attempting to confirm the aforementioned social psychology findings in agent-human emotional contagion forms the basis of this work. Pursuant of this goal, three sets of studies are conducted. The first study examines the pure contagion case by simply showing subjects a still image of a virtual character with either a happy expression or a neutral expression and then assessing the subject's mood thereafter. The use of a still image as a manipulation follows from previous studies in emotional contagion [19, 25, 26, 30].

The second study adds the presentation of a game-theoretic situation known as a Stag Hunt along with the character image to assess both the contagion the behavioral impact of the virtual character in a strategic setting. While studies have shown that emotional contagion can impact one's propensity to trust and enhance perceived cooperation among other findings [2, 9], there has been far less work showing behavioral impacts in strategic situations. Although people may report themselves to be more trusting, for example, this may not result in any meaningful impact on behavior in a strategic situation. Thus, we also attempt to examine whether behavioral impacts arise in strategic situations from agent-human contagion to better understand its potential impacts in real-world agent applications. Finally, the third study examines the post-hoc hypothesis that the presentation of a decision to the user dampens the emotional contagion effect. Specifically, we present the same strategic situation as in the second study, but with the decision already made for the subject. These studies present the first attempt to assess emotional contagion from virtual characters to human users.

In this work, we begin by providing, to our knowledge, the first confirmation of emotional contagion between virtual agents and humans. Evidence shows a very large increase in happiness from only adding a smile to an otherwise identical still image of a virtual character. We then examine the details of the contagion, finding no support for the hypothesis that women are generally more strongly influenced by emotional contagion than men. Neither the perceived attractiveness nor the perceived ethnicity of the character used appear to affect the contagion consistently either. However, the perceived happiness of the character has a very high correlation with participant happiness. In the second study, when the character is placed in the context of a strategic decision, both subject behavior and subject emotions are only impacted significantly by one character. The last study, which removes the user's decision from the previous experiment, finds that the character's expression's effect on emotion returns significantly, showing that a strategic decision posed to users will dampen the emotional contagion effect beyond only reading about a situation. Finally, post-hoc analysis suggests that emotional contagion with women may be more resilient to the cognitive load dampening effects of reading about a situation.

2. RELATED WORK

Emotional contagion research in the agents literature falls primarily into three categories: models of emotional contagion, creating rapport between virtual agents and humans, and the impact of agent mood expressions on behavior. Models of emotional contagion have been explored in a computational context that focus on crowd or society simulation. For example, [4, 10, 21] each present alternative models of emotional contagion in agent crowds, while [28] proposes a comparison technique to evaluate such models. Bosse et al. [4] attempt to model the phenomenon of emotions in a crowd spiraling out of control. Durupinar [10] instead uses emotional contagion as a component in a crowd simulation to aid in creating natural variation in crowd types. Pereira et al. [21] model the incorporation of individual susceptibilities and biases into the computation of emotional contagion. This body of work is an attempt to mimic human-human contagion and not an exploration of agent-human contagion which we seek to understand here.

There also exists a large body of work on the interaction between virtual agents and humans [5, 12, 29]. The entire area of virtual rapport [12, 29], for example, focuses on user opinions of the virtual agents and their interaction. The primary goal is to create agents that users enjoy, appreciate, and relate to. Recent work has looked at the impact of agent expressions in a strategic negotiation setting [5] as well. However, their work focuses on the behavioral impact of varying the intent of agent expressions on user behavior without examining the emotional impact or the mechanism by which the change is induced. Neither of these works explicitly examine the impact of virtual character expressions on the emotions of subjects.

3. THEORETICAL BACKGROUND

In the social sciences, the literature on emotional contagion is far more expansive. Hatfield et al. [13] popularized the area by compiling a plethora of situations in which the phenomenon had been observed in their work as well as the work of other researchers. Follow-up research by the co-authors as well as researchers in related fields such as managerial and occupational sciences [2, 11, 22, 25] continued to detail the effects of the phenomenon in new domains. Recently, there have been works beginning to quantify emotional contagion and explore cross-cultural variations in attributes that affect emotional contagion [7, 20].

In light of the extensive evidence of emotional contagion's effects in human-human interactions, our work extends the understanding of this phenomenon into the realm of agent-human interactions. While some studies have been conducted with real people as the stimulus [2, 22], a large body of social psychological studies of emotional contagion features an image or video of only a person's face as the origin of the contagion [14, 25, 30]. With the rapid improvements in virtual agent facial displays, and the accepted assumption that the facial display of emotion plays a key role in emotional contagion, we would expect to see a contagion of emotions from an image of a virtual agent's face to humans. Thus, the primary hypothesis of this work is:

HYPOTHESIS 1. *The facial display of an emotion by a virtual character will result in emotional contagion with a human.*

A directly related hypothesis also presented by Hatfield et al. [13] states that the strength of the expression will be correlated with the degree of emotional contagion. This was explored by Wild et al. [30] who tested four degrees of expressions for four different expressions (happiness, sadness, surprise, and pleasure), but found *no* significant systematic effect of expression strength. We examine a similar hypothesis in a virtual character context:

HYPOTHESIS 2. *The perceived happiness of the virtual character's expression will be correlated with the degree of change in the happiness of the human viewer.*

While many recent pursuits in emotional contagion research have looked into the mechanism causing the contagion [14, 15], our focus is on its existence in agent-human interactions. Previous work explored differences in the effect of emotional contagion by gender, and found that women were significantly more strongly impacted than men [8, 27]. Researchers also found that contagion increased in cases where the subject shared the same ethnicity as the stimulus [8]. Finally, attraction to the stimulus was shown to have a positive effect on the contagion experienced in subjects [27]. These results yield the following set of hypotheses:

HYPOTHESIS 3. *Women will experience a stronger contagion effect with a virtual character's facial expression than men will.*

HYPOTHESIS 4. *People will experience a stronger contagion effect with a virtual character's facial expression if the character is perceived to be more attractive.*

HYPOTHESIS 5. *People will experience a stronger contagion effect with a virtual character's facial expression if the character is of the same ethnicity.*

4. PURE CONTAGION STUDY

In this study, we test the existence of and factors contributing to emotional contagion between an image of a virtual character's facial expression and a human subject. The experiment setup involved a still image of a character, a self-report of emotion, and a character assessment. Participants were randomly assigned to see one of the images shown in Figure 1, and participants were informed that they would be questioned about the character later. Thus, the study was a 4 (characters) $\times$ 2 (expressions) between-subjects design.

Each character was shown with either a happy or neutral expression. Ellie is part of the SimCoach[1] project, while Utah is part of the Gunslinger[2] project. Dia was taken from screenshots from Final Fantasy XIII.[3] Finally, Roy was taken from screenshots of the game L.A. Noire.[4]

In the self-report of emotion, we asked subjects how strongly they felt each of 8 emotions on a 0-8 Likert scale: angry, joyful, upset, sad, happy, gloomy, irritated, and calm. Only the measure of Happy was used as the other emotions were only included for compliance checking. Specifically, participants that rated both Angry and Joyful higher than 5 and participants that rated Happy and Joyful more than 3 points apart were considered not in compliance.

Finally, a 15-question survey was administered to gauge subjects' perception of the characters shown. Attributes were drawn primarily from the BSRI [3] and included: Aggressive, Affectionate, Friendly, Attractive, Self-Reliant, Warm, Helpful, Understanding, Athletic, Gentle, and Likable. Every question was asked on a 0-8 Likert scale. Compliance tests included duplicating the Attractiveness question and ensuring both occurrences were within 2 points of each other, an Unattractiveness question which could not exceed 5 if Attractiveness exceeded 5, and finally a question that simply asked participants to 'Pick number eight'. Participants were also asked to rate how happy the character seemed.

A total of 415 participants that responded to the experiment, conducted on Amazon Mechanical Turk, passed the compliance tests. Participants were required to be over 18 years of age and were compensated $0.25. The gender distribution was approximately one-third female and two-thirds male, and approximately two-thirds of respondents indicated their ethnicity as Indian.

4.1 Results

We examined whether the facial emotion expressed affected subjects' self-report of emotion. For each of the characters used, participants rated the image used in the Happy condition as significantly happier than the image used in the Neutral condition ($p < 0.001$ for all characters). Thus, according to Hypothesis 1, participants should report greater happiness in the Happy condition compared to the Neutral condition.

[1]http://ict.usc.edu/projects/simcoach
[2]http://ict.usc.edu/projects/gunslinger/
[3]www.finalfantasyxiii.com
[4]www.rockstargames.com/lanoire/

	Condition	Mean	SD	n	p
Utah	Neutral	3.96	2.54	57	< 0.001
	Happy	5.60	2.12	52	
Roy	Neutral	4.00	2.45	45	< 0.001
	Happy	5.75	1.86	55	
Dia	Neutral	4.04	2.26	46	< 0.001
	Happy	5.96	2.19	47	
Ellie	Neutral	4.49	2.37	66	< 0.001
	Happy	5.27	2.10	47	

Table 1: Happiness statistics for Pure Contagion Study

Figure 2 shows the happiness reported for each character, with dark bars indicating responses in the Neutral condition and light bars indicating responses in the Happy condition. Table 1 shows the means, standard deviations, sample size, and p-values for each experiment. As can be seen, greater happiness was reported in the Happy condition for every character and one-way ANOVA tests revealed significance in every case. This supports Hypothesis 1's prediction that an image of a virtual character will cause emotional contagion with a human viewer, since the display of happiness resulted in reports of higher happiness in subjects as compared to the neutral display.

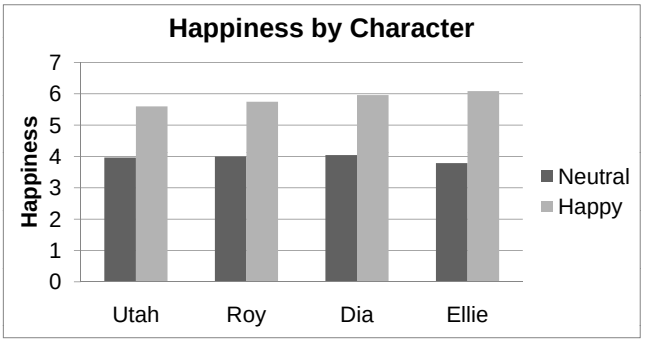


Figure 2: Happiness by character, Pure Contagion Study

4.2 Gender Effects

Hypothesis 3 predicts that women will experience a stronger contagion effect than men. In this context, this suggests that female subjects will report a greater difference in happiness between Neutral and Happy conditions as compared to male subjects. We breakdown the previous results and list the average differences in happiness reported by each gender for each character in Figure 3. The y-axis now shows the difference in participant happiness from the Neutral to Happy condition and the x-axis shows the character. The dark bars represent the increase in the average happiness of men while the light bars show the same measure for women. Therefore, Hypothesis 3 suggests that the bars for female subjects should always be taller than the bars for male subjects.

As can be seen, there was a greater increase in happiness for females in Utah and Roy, but the opposite was true for Dia and Ellie. This does not support Hypothesis 3, but post-hoc analysis suggests a clear cross-gender effect. None of the 11 character attributes surveyed in this study nor the 7 attributes surveyed in the third study showed the same cross-gender trend as exhibited in Figure 3. However, analysis of the perceived happiness of the character shown reveals an alternative explanation.

4.3 Perceived Happiness Effects

Figure 1: Characters used, neutral and happy expressions (color)

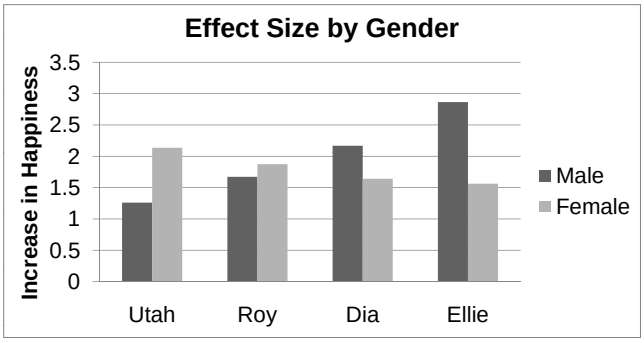


Figure 3: Effect size by gender, Pure Contagion Study

Hypothesis 2 suggests that the perceived happiness of a character will be correlated with the self-report of happiness by subjects. Wild et al. [30] do *not* find systematic support for this hypothesis across the four expressive strengths they tested. In our experiments, however, a Pearson's product moment correlation test reveals that perceived happiness of the character is highly correlated with the self-report of happiness of the participant ($p < 0.001$, $r = 0.6826$). Next, we examine the perceived happiness data on aggregate for each character.

Figure 4 shows the average differences in perceived happiness of the character between Neutral and Happy conditions. If perceived happiness of the characters are highly correlated to respondents' self-reports of happiness, we would expect the exact same trend from Figure 3 to be replicated here, with high increases in perceived happiness occurring with high increases in subject happiness. As can be seen, this is very nearly the case. The trend is identical for female subjects, with light bars exhibiting the same pattern as they do in Figure 3. With the exception of Utah, it is the same for male subjects as well. This suggests that the 'cross-gender' trend seen in Figure 3 may actually be caused by variations in perceived happiness of the characters instead of by gender biases. Of course, the differences in perceived happiness of the characters appears to originate from gender-based effects, but we leave further exploration of this subject to future work.

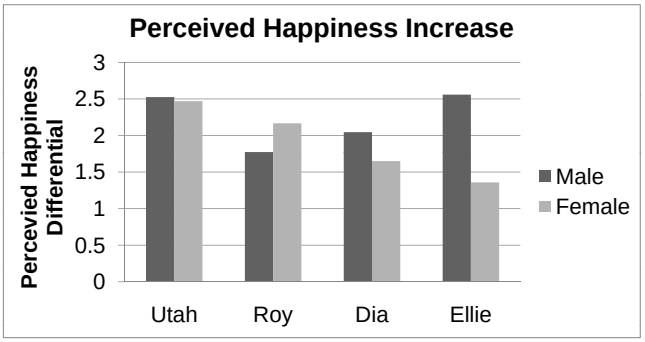


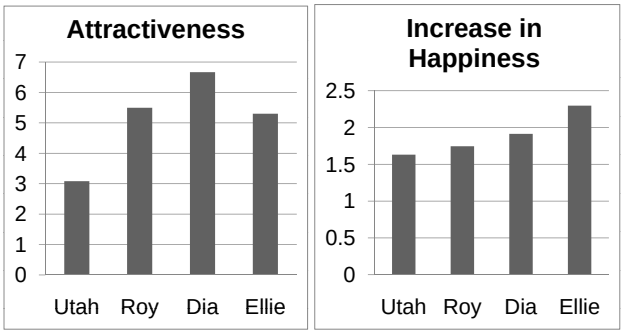
Figure 4: Increase in perceived character happiness by gender

4.4 Attractiveness Effects

Hypothesis 4 suggests that perceived attractiveness of the character should contribute to emotional contagion. A Pearson's product-moment correlation reveals a significant but mild correlation ($p < 0.001$, $r = 0.3918$) between the happiness of participants and the perceived attractiveness of the character shown. For further support, we look to an aggregate analysis of the data, grouping the attractiveness data by character.

Figure 5a shows the average attractiveness rating for each character. As can be seen, Dia is the most attractive, statistically significantly more so than Ellie ($p < 0.001$). Figure 5b shows the increase in respondent happiness between Neutral and Happy con-

ditions. Although Utah is the least attractive and does indeed cause the lowest increase in happiness as per Hypothesis 4, Ellie is actually the character that induces the greatest increase, with Dia substantially lower. This suggests that the attractiveness of the character alone does not provide a strong enough mediating effect in this context to support Hypothesis 4.



(a) Attractiveness (b) Increase in happiness

Figure 5: Pure Contagion Study

4.5 Ethnicity Effects

The study also asked subjects for their ethnicity. In light of the large Indian population, the ethnicities included: Caucasian, Asian (exc. Indian), Indian, African / African-American, Other. However, since the subject pool only contained substantial numbers of Caucasian and Indian respondents ($n > 10$), we restrict analysis to these two ethnic groupings only.

To assess each character's perceived ethnicity, the character assessment also included an element asking the user to respond on a 0-8 Likert scale of '0 - Do not agree at all' to '8 - Very strongly agree' with the statement: The Character is the same ethnicity as I am. As can be seen in Figure 6, all characters were rated much more similar to Caucasians than Indians, with especially large differences for Roy and Utah. All differences were statistically significant ($p < 0.01$ for all but Dia, which was $p = 0.02843$).

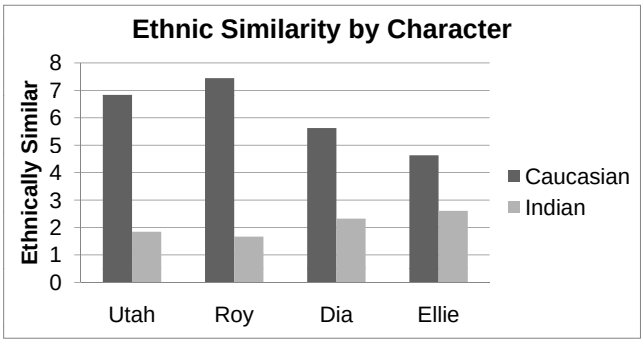


Figure 6: Ethnic similarity for Caucasians and Indians

Figure 7 shows the increase in happiness between Neutral and Happy conditions, broken down by self-reported ethnicity. Hypothesis 5 suggests that since Caucasians find all the characters more similar to themselves, Caucasian respondents should report a greater effect of contagion than Indians. This difference should be especially large for Utah and Roy. However, no such trend emerges. Caucasian subjects show a smaller increase in happiness for Utah than Indian subjects, but also show a larger increase for Roy. Thus, we do not find support for Hypothesis 5 in this experiment.

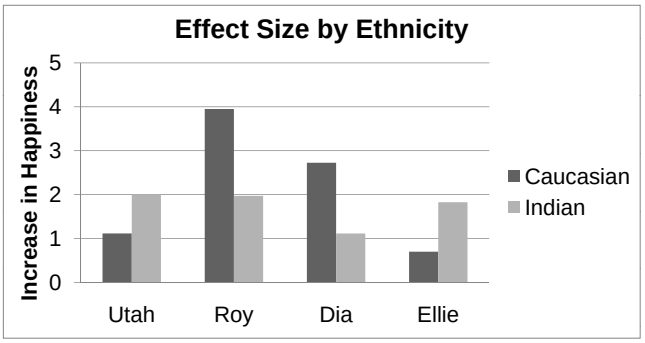


Figure 7: Effect size by ethnicity, Pure Contagion Study

5. STRATEGIC SITUATION STUDY

Having established the existence of agent-human emotional contagion, we extend the research to include a strategic interaction. Studies into the effects of emotional contagion have primarily been in mimicry, self-reports of emotion, and other non-decision-based effects such as changes in trust inventory responses and judge ratings of 'cooperativeness' [2, 9, 14]. While there has been some work in behavioral changes due to emotional contagion, such as its impact on donation amounts [25], our work is the first to consider impacts in a strategic context.

The experimental setup involved a still image of a character along with the presentation of a strategic situation for which a decision must be made, followed finally by a self-report of emotion. The same self-report of emotion used in the first study is employed here.

We used a cooperation situation based on the standard game-theoretic Stag Hunt situation. Originally posed by Jean-Jacques Rousseau, the original story involves two individuals going out on a hunt. Each can commit to hunting a stag or a rabbit and must do so without knowing the other player's choice. An individual can successfully catch a rabbit alone, but the rabbit is worth less than the stag. However, in order to successfully hunt a stag, both hunters must commit to hunting stag. This situation resembles the well-known Prisoner's Dilemma, but differs in that the highest reward comes from both players cooperating. In the Prisoner's Dilemma, the highest reward is achieved by the defector if the other player choose to cooperate. Thus, rational play depends on beliefs about the other player in a Stag hunt, whereas defecting is strictly dominant in a Prisoner's Dilemma.

The actual story used in this experiment casts the Stag Hunt scenario in a modern, less outlandish context in which the subject and a coworker he/she has never met are tasked with decorating specific rooms in the office and can either choose to work separately (taking more time) or work together through both of their assigned rooms (taking less time). The amount of time it would take to perform the decoration task was not explicitly stated. The coworker in question was the character whose image is presented with the situation. Subjects were asked how likely they were to help the character with the task on a 0-8 Likert scale.

A total of 572 participants responded to the experiment, which was again conducted via Amazon Mechanical Turk, passed the compliance tests. Participants were required to be over 18 years of age and were again compensated $0.25 for compliant participation. The gender distribution was once more approximately one-third female and two-thirds male, with approximately two-thirds of

respondents were from India.

5.1 Decision Results

In light of the very strong contagion effect in the Pure Contagion Study and reports of emotional contagion of happiness leading to more trust [9], we expected to see increased happiness in Happy conditions lead to increased likelihood of cooperation. Indeed, we do find a tight link between likelihood of cooperation and participant happiness as shown in Figure 8. The x-axis plots the happiness rating, and the y-axis indicates the average likelihood of cooperation for all respondents with the given happiness rating across all conditions. As the regression's very high R-squared of 0.852 indicates, the two measures are very tightly linked.

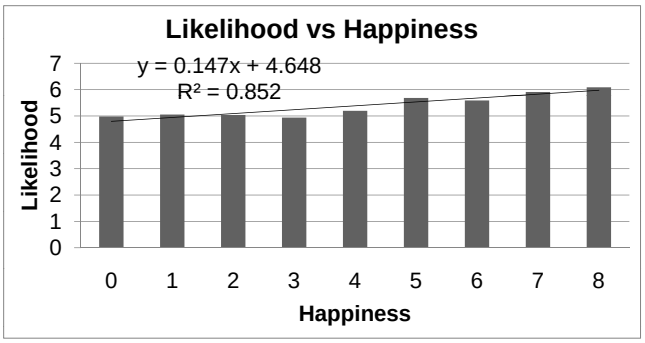


Figure 8: Likelihood of cooperation versus happiness

However, only the experiment with Dia yielded a statistically significant change in responses. This suggests that the change results from a character-specific attribute and not simply an expression-based mechanism. The lack of effect for the other characters is due partially to the regression's low coefficient of 0.147, which implies that huge changes in happiness are required to induce changes in the likelihood of cooperation. However, the Pure Contagion Study *did* find very large changes in happiness that should have been sufficient. A closer look at the emotional influence of our manipulation reveals the second half of the story.

5.2 Contagion Results

While the Pure Contagion Study reported astoundingly large effects of a smile in a still image of a virtual character, the addition of a strategic situation and decision may have altered the contagion effect. Thus, we examine them in this experiment again. We summarize the overall results for each character in Figure 9. Each character is shown on the x-axis, with the happiness reported on the y-axis. The dark bars indicate the average happiness reported by subjects who viewed the specified character with a neutral expression while the light bars indicate the average happiness for viewers of the happy expression.

As before, we expect subjects in the Happy condition to report higher happiness than subjects in the Neutral condition across all characters. This was indeed the case, as evidenced by the light bars always being higher than the corresponding dark bars. However, the difference between the bars are much smaller than in the Pure Contagion Study and, in fact, statistical significance was found only in the experiment using Dia, indicating that something character-specific is allowing her to retain more of her emotional impact while all other characters experienced a much greater dampening of emotional impact. In exploring the attributes surveyed in this work (11 in the Pure Contagion Study, 7 in the Strategic Decision Study), no candidate for a consistent explanatory variable was found. The

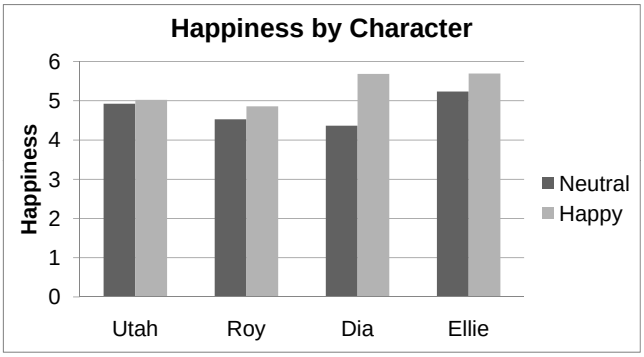


Figure 9: Happiness by character, Strategic Situation Study

	Condition	Mean	SD	n	p
Utah	Neutral	4.92	2.56	105	0.7638
	Happy	5.02	2.48	125	
Roy	Neutral	4.53	2.38	36	0.2098
	Happy	4.86	2.76	49	
Dia	Neutral	4.37	2.57	41	0.019
	Happy	5.68	2.30	38	
Ellie	Neutral	5.24	2.59	93	0.2231
	Happy	5.69	2.39	85	

Table 2: Happiness statistics, Strategic Situation Study

full table of statistical test results can be seen in Table 2.

These results suggest that the presentation of a strategic situation and a trust-based decision dampens the emotional contagion effect. This is actually in line with findings by researchers in social psychology [25, 31] that found that deliberative thinking can dampen emotional influences. However, in light of the tight correlation between the decision and reported happiness, we hypothesize that the decision itself contributes to the dampening effect beyond the impact of simply reading about the situation.

6. STRATEGIC DECISION STUDY

This study was pursued to disentangle the novel effect of making a strategic decision from the previously confirmed effect of reading a situation description [25, 31]. It presents subjects with the same situation as in the Strategic Situation Study but removes the decision element from it and simply states that the subject will be cooperating with the character shown to complete the office decoration task. We again specify that the character's room will be decorated first to minimize confounding factors.

In addition to this, we also conducted a second character assessment to target attributes that may contribute to cooperation in the office decoration task to aid in post-hoc analysis. This was done after the self-report of emotion, so it did not impact the original intent of the experiment. A 10-question survey, primarily a subset of the survey used in [16], was administered using a 0-8 Likert scale for each question. Attributes included: Competent, Trustworthy, Knowledgable, Hard-Working, Enthusiastic, Fun, and Artistic. Compliance tests for the character assessment included duplicating the Competence question and ensuring ratings for both occurrences were within 2 points of each other, a Laziness question which could not exceed 5 if Hard-working exceeded 5 as well, and finally a question that simply asked participants to 'Pick number seven'.

In Table 3, the overall results of the experiment are shown, with significance again calculated using a one-way ANOVA. As would

	Condition	Mean	SD	*n*	*p*
Utah	Neutral	4.04	2.67	27	0.1329
	Happy	5.09	2.63	32	
Roy	Neutral	4.83	2.33	24	0.2247
	Happy	5.66	2.53	29	
Dia	Neutral	5.88	2.11	48	0.3485
	Happy	6.28	2.08	46	
Ellie	Neutral	4.76	2.33	46	0.008
	Happy	5.95	1.77	41	

Table 3: Happiness statistics, Strategic Decision Study

	Condition	Mean	SD	*n*	*p*
Utah	Neutral	2.69	2.21	13	0.0302
	Happy	5.00	2.96	14	
Roy	Neutral	4.78	2.44	9	0.1054
	Happy	6.40	1.43	10	
Dia	Neutral	4.94	2.54	16	0.1081
	Happy	5.85	2.54	21	
Ellie	Neutral	4.80	2.27	15	0.1206
	Happy	6.00	1.59	12	

Table 4: Happiness, female subjects, Strategic Decision Study

be expected following findings in social psychology that even reading additional material can dampen emotional influence [25, 31], the effect observed in the Pure Contagion Study has not returned in full force. However, the average happiness reported by participants shows a much larger differential than in the Strategic Situation Study, supporting the hypothesis that the decision itself contributed substantially to the dampening of emotional contagion.

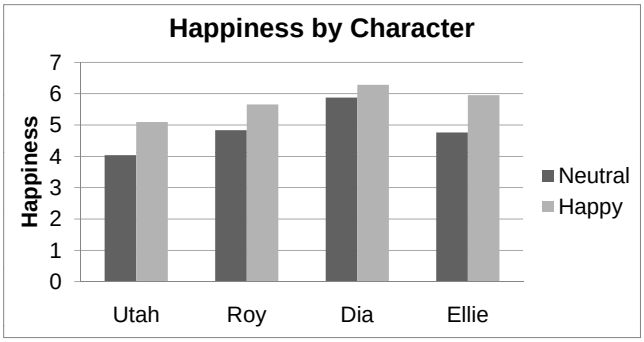


Figure 10: Happiness by Character, Strategic Decision Study

A closer look reveals that gender plays a large role in this study. Figure 11 shows the increase in happiness between Neutral and Happy conditions, split for female and male participants. Notice that for each character, the difference between conditions for females (light bars) is always very high, whereas for men (dark bars), this only occurs with Ellie. It is also interesting to note that the same trend seen in Figure 3 is evident here for women as well. Specifically, greater increases in happiness occurred for characters that were perceived to have a greater increase in happiness between neutral and happy expressions. Table 4 shows the detailed statistical results for female subjects. Notice that all results are either significant or very nearly so. The equivalent table for men, Table 5, reveals that only with Ellie do men have anywhere near a statistically significant response to the stimulus used.

Notice that the effect sizes for women are nearly the same as in

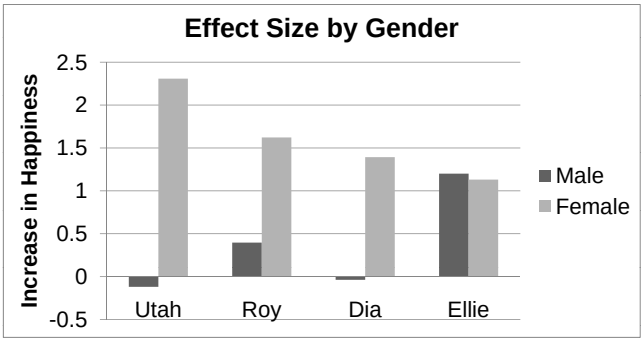


Figure 11: Happiness by gender, Strategic Decision Study

	Condition	Mean	SD	*n*	*p*
Utah	Neutral	5.29	2.49	14	0.8933
	Happy	5.17	2.43	18	
Roy	Neutral	4.87	2.36	15	0.6632
	Happy	5.26	2.90	19	
Dia	Neutral	6.44	1.63	32	0.9322
	Happy	6.40	1.66	25	
Ellie	Neutral	4.8	2.41	30	0.0487
	Happy	5.93	1.87	29	

Table 5: Happiness, male subjects, Strategic Decision Study

the Pure Contagion Study and, in fact, exhibit the exact trend from Figure 3. This supports a variation of Hypothesis 3 that emphasizes resilience of emotional contagion as opposed to magnitude of effect as has been previously reported. Specifically, it appears that emotional contagion to women is less dampened by reading a situation description than for men. However, since this is a post-hoc hypothesis, we leave further exploration of this to future work.

7. CONCLUSION

In this work, we provide the first ever examination of agent-human emotional contagion. We confirm its existence with a pure contagion study with astoundingly strong results. We find no support for gender differences in emotional contagion strength despite numerous studies in human-human contagion in support of the hypothesis [6, 17]. The attractiveness of the character also does not appear to affect the contagion effect, although its perceived happiness does. In a second study, a strategic decision is added that greatly dampens the contagion effect and, with one exception, did not impact behavior. The final study, which removes the user's decision from the previous experiment, finds that the emotional contagion effect returns significantly. This shows that a strategic decision posed to users will dampen the emotional contagion effect beyond the dampening effect of reading the situation itself. In addition, we find evidence of a gender-based difference in susceptibility to cognitive load's dampening effect on emotional contagion.

Our findings suggest a number of key recommendations for virtual agent researchers. First, emotional contagion with virtual agents is very substantial and applications need to accurately account for it. We have shown that in some domains, even a still image can have a huge emotional effect, but more work must be done to delineate these domains with greater clarity. Second, considering the number of unsupported hypotheses found in this work, researchers should be wary about assuming that human-human social psychology will directly translate into agent-human interactions. Finally, our work has looked at smiles that are perceived as happy, but there

are different types of smiles and not all smiles reflect positive emotional states [1]. Further investigations should be carried out to understand the different effects of character expressions. As virtual agent applications extend beyond entertainment into emotionally-charged domains with very serious repercussions such as psychotherapy and military training, researchers must be ever-vigilant of the emotional impacts their characters might have on users.

8. ACKNOWLEDGEMENT

This research is supported by the United States Department of Homeland Security through Center for Risk and Economic Analysis of Terrorism Events (CREATE). This research was sponsored in part by the U.S. Army Research, Development, and Engineering Command (RDECOM) Simulation Training and Technology Center (STTC). The content or information presented does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

9. REFERENCES

[1] Z. Ambadar, J. F. Cohn, and L. I. Reed. All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior*, 33(1):17–34, 2009.

[2] S. G. Barsade. The Ripple Effect: Emotional Contagion and Its Influence on Group Behavior. *Administrative Science Quarterly*, 47:644–675, 2002.

[3] S. L. Bem. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 42:155–162, 1974.

[4] T. Bosse, R. Duell, Z. A. Memon, J. Treur, and C. N. V. D. Wal. A Multi-Agent Model for Emotion Contagion Spirals Integrated within a Supporting Ambient Agent Model. In *PRIMA-09*, 2009.

[5] C. de Melo, P. Carnevale, and J. Gratch. The Effect of Expression of Anger and Happiness in Computer Agents on Negotiations with Humans. In *AAMAS-11*, 2011.

[6] U. Dimberg and L.-O. Lundquist. Gender differences in facial reactions to facial expressions. *Biological Psychology*, 30(2):151–159, 1990.

[7] W. Doherty. The Emotional Contagion Scale: A Measure of Individual Differences. *Journal of Nonverbal Behavior*, 21(2), 1997.

[8] W. Doherty, L. Orimoto, T. M. Singelis, E. Hatfield, and J. Hebb. Emotional Contagion: Gender and Occupational Differences. *Psychology of Women Quarterly*, 19(3):355–371, 1995.

[9] J. R. Dunn and M. E. Schweitzer. Feeling and Believing: The Influence of Emotion on Trust. *Psychology of Personality and Social Psychology*, 88(5):736–748, 2005.

[10] F. Durupinar. *From Audiences to Mobs: Crowd Simulation with Psychological Factors*. PhD thesis, Bilkent University, 2010.

[11] A. A. Grandey. Emotional regulation in the workplace: A new way to conceptualize emotional labor. *Journal of Occupational Health Psychology*, 5(1):95–110, January 2000.

[12] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. J. van der Werf, and L.-P. Morency. Virtual Rapport. In *IVA-06*, 2006.

[13] E. Hatfield, J. T. Cacioppo, and R. L. Rapson. *Emotional Contagion*. Cambridge University Press, 1994.

[14] U. Hess and S. Blairy. Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. *International Journal of Psychophysiology*, 40(2):129–141, 2001.

[15] U. Hess, P. Philippot, and S. Blairy. Facial Reactions to Emotional Facial Expressions: Affect or Cognition? *Cognition and Emotion*, 12(4):509–531, 1998.

[16] L. Hoffmann, N. C. Krämer, A. Lam-chi, and S. Kopp. Media Equation Revisited: Do Users Show Polite Reactions towards an Embodied Agent? In *IVA-09*, 2009.

[17] A. M. Kring and A. H. Gordon. Sex differences in emotion: Expression, experience, and physiology. *Journal of Personality and Social Psychology*, 74(3):686–703, 1998.

[18] P. Kulms, N. C. Krämer, J. Gratch, and S.-H. Kang. It's in Their Eyes: A Study on Female and Male Virtual Humans' Gaze. In *IVA-11*, 2011.

[19] D. A. Lishner, A. B. Cooter, and D. H. Zald. Rapid Emotional Contagion and Expressive Congruence Under Strong Test Conditions. *Journal of Nonverbal Behavior*, 32:225–239, 2008.

[20] L.-O. Lundqvist. Factor Structure of the Greek Version of the Emotional Contagion Scale and its Measurement Invariance Across Gender and Cultural Groups. *Journal of Individual Differences*, 29(3):121–129, 2008.

[21] G. Pereira, J. Dimas, R. Prada, P. A. Santos, and A. Paiva. A Generic Emotional Contagion Computational Model. In *ACII-11*, 2011.

[22] S. D. Pugh. Service with a smile: Emotional contagion in the service encounter. *Academy of Management Journal*, 44(5):1018–1027, 2001.

[23] G. Riva. Virtual Reality in Psychotherapy: Review. *CyberPsychology & Behavior*, 8(3):220–230, 2005.

[24] A. Rizzo, J. Pair, P. J. McNerney, E. Eastlund, B. Manson, J. Gratch, R. W. Hill, and W. Swartout. *Development of a VR Therapy Application for Iraq War Military Personnel with PTSD*, volume 111 of *Medicine Meets Virtual Reality*, pages 407–413. IOS Press, 13th Annual Medicine Meets Virtual Reality Conference, Long Beach, CA, 2005.

[25] D. A. Small and N. M. Verrochi. The Face of Need: Facial Emotion Expression on Charity Advertisements. *Journal of Marketing Research*, 46(6):777–787, December 2009.

[26] M. Sonnby-Borgström, P. Jönsson, and O. Svensson. Gender differences in facial imitation and verbally reported emotional contagion from spontaneous to emotionally regulated processing levels. *Scandinavian Journal of Psychology*, 49(2):111–122, 2008.

[27] N. Stockert. *Perceived Similarity and Emotional Contagion*. PhD dissertation, University of Hawaii at Manoa, Dept. of Psychology, 1994.

[28] J. Tsai, E. Bowring, S. Marsella, and M. Tambe. Empirical Evaluation of Computational Emotional Contagion models. In *IVA-11*, 2011.

[29] N. Wang and J. Gratch. Rapport and Facial Expression. In *ACII-09*, 2009.

[30] B. Wild, M. Erb, and M. Bartels. Are emotions contagious? evoked emotions while viewing emotionally expressive faces: quality, quantity, time course and gender differences. *Psychiatry Research*, 102(2):109–124, 2001.

[31] T. D. Wilson, S. Lindsey, , and T. Y. Schooler. A Model of Dual Attitudes. *Psychological Review*, 107(1):101–126, 2000.

Modeling emotional contagion based on experimental evidence for moderating factors

Rene Coenen[1] and Joost Broekens[2]

[1]Leiden University, Mediatechnology

[2]Delft University of Technology, Interactive Intelligence

{renecoenen, joost.broekens}@gmail.com

ABSTRACT

There is a lot of evidence for the phenomenon describing the spread of emotion from one person to another, called emotional contagion. Although there is a large body of research on this topic, research containing evidence for factors that moderate the process of emotional contagion, is limited and inconclusive. Furthermore most of these studies are done in a dyadic lab-setting and consequently little is known about emotional contagion in groups. This paper presents, for the first time, a dynamic computational model of contagion in groups of agents based on factors that moderate contagion. These factors are strictly based on experimental evidence in the psychological literature. In this paper we first present our review of the psychological literature. We then present our computational model as well as a pilot study investigating several group contagion cases showing the flexibility and potential of this strategy.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence–Intelligent agents

General Terms

Human Factors.

Keywords

Emotional Contagion, agent based modeling, multi-agent systems

1. INTRODUCTION

If we perceive another person's emotional expressions, for example seeing a happy person smile, we tend to suddenly find ourselves also smiling and sharing this person's happiness without ever having intended to do so. This phenomenon of catching each other's emotions is called emotional contagion.

A specific, and predominant definition of emotional contagion describes it as the tendency to automatically mimic and synchronize facial expressions, vocalizations, postures and movements with another person's and, consequently, to converge emotionally[13]. This definition is based on the theory of primitive emotional contagion [12] and follows one of two predominant perspectives regarding possible mechanisms behind emotional contagion. The perspective emphasizes a subconscious level on which emotional contagion occurs. Some research suggests that emotional contagion is directly induced by the activation of neural representations of similar emotions in the observer. Emotional Contagion can also occur through a more conscious process. Following this perspective some studies suggest that emotional contagion can happen through social comparison processes, in which people evaluate their affective state in comparison with that of other people in their environment and respond according to what seems appropriate [11,28].

Appears in: *Proceedings of the Workshop on Emotional and Empathic Agents, in the 11th International Conference On Autonomous Agents and Multiagent Systems (AAMAS2012), June, 4-8, 2012, Valencia, Spain.*

Most of the research on emotional contagion to date has been done in a dyadic setting and therefore little is known about the occurrence of emotional contagion within groups, of which the existence is also known [2]. To better understand contagion in groups, computational modeling can be used in the same way as it is used in other dynamic systems, especially if moderating factors for emotional contagion can be systematically varied in a computational agent. In terms of more concrete application value, our work can contribute to the development of virtual characters; especially VC's that need to show behavior that is emotionally plausible at the group level [33], or in dyadic setting involving a VC and a user.

To date there are only a few studies concerning computational models specifically for emotional contagion. Tsai et al. [31] present an interesting empirical evaluation of several recent computational emotional contagion models. In their evaluation they compare two models which differ substantially in the underlying modeling-approach. They find a thermodynamics based model created by Bosse et al. [4] generates superior results when compared to a model by Durupinar [9] which is based on an epidemiological process; implying that to date first mentioned specific underlying modeling-strategy is best suited to represent the process of emotional contagion. Bosse's model is inspired on recent studies involving group affect. However, the model approaches the dynamic nature of contagion in a relatively abstract manner; taking into account only the basic aspects necessary for contagion. Other studies utilize a similar abstraction strategy. For example Bispo and Paiva [3] based their model on the emotional contagion scale; a measurement instrument for susceptibility to emotional contagion. They take into account the specific emotions and susceptibility to emotional contagion as the only moderating factors.

We present a novel approach of computational modeling of emotional contagion solidly based on psychological evidence for contagion moderating factors. While we share a dynamical system approach with Bosse et al. [4], our model explicitly simulates the effect of individual moderating factors. We first review experimental evidence for moderating factors based on the psychological literature, then we present our computational model and pilot study showing its potential.

2. REVIEW METHODOLOGY

Although there is a large body of research on the topic of emotional contagion, research directly studying emotional

contagion containing evidence for potential factors that influence the process of emotional contagion is limited and inconclusive, illustrating the importance of a structured review of the evidence found in this area.

First an exploratory search was performed using Google Scholar in September 2011. Search-terms included emotional contagion, affective contagion, mood contagion and affect contagion. All types of relevant studies in English were taken into account. Possibly relevant studies were added based on the references in these studies.

In the second phase the final corpus was obtained with an additional exhaustive search in October 2011 using the following EBSCOhost online databases: PsycINFO, PsycARTICLES, Acedemic Search Premier and the Psychology and Behavioral Sciences Collection. Basic exact-phrase searches were done consisting of the following relevant keywords: emotional contagion, affective contagion, affect contagion and mood contagion. The resulting collection of articles was compared with previous findings and relevant articles were added to the collection.

During the last phase the resulting collection was filtered based on the following criteria. The study contains results based on quantitative studies directly involving psychologically healthy subjects. The study explicitly presents evidence for factors that can, or evidently cannot possibly moderate emotional contagion.

3. RESULTS REVIEW

Emotional contagion moderators can be categorized in three categories (Table 1.) and these categories will be described in more detail in the rest of this section:

- *individual differences* accounting for factors such as personality and gender
- *interpersonal factors* comprising for example similarity and group membership, and
- *miscellaneous* factors.

Table 1. moderating factors for emotional contagion.

				Susc-eptibility	Cont-agability
Individual differences	emotion related trait	sender			+
		receiver		+	
	gender	female		+	+
interpersonal factors	similarity	attitudinal similarity		+ *	
		situational similarity		+	
			group membership	+ **	
	social power			+/- ***	+/- ***
	intimacy			+	
miscellaneous	pre-existing positive mood			+	

** effect only found for positive emotions; ** non-group membership induces opposite emotion (divergence); *** effect found in both directions*

We further found it useful to separate the moderation effect of factors in a moderation of susceptibility to contagion ("in") and a moderation of what we call contagability ("out"). Concretely this means that a factor can influence the susceptibility of a person but also the contagability. The overall contagion one person (let's say *Marie*) experiences is thus determined by that person's (*Marie's*) susceptibility ("in") and by the other's (*Bob's*) contagability ("out"). This is different from a person's sender and receiver traits. These two factors contribute to contagability and susceptibility respectively but are not equivalent to these constructs, as will become clear in the review.

3.1 Individual Differences

3.1.1 Emotion Related Trait

The theory of primitive emotional contagion of Hatfield et al. implies that a differentiation can be made between people who are strong transmitters of emotions and people who are strong receivers (catchers) of emotions. Hatfield et al. state that contrary to the often charismatic, entertaining or dominant people who by their innate bodily circuitry communicate their emotions more strongly to others, the people who are especially susceptible to emotional contagion are those who pay close attention to others and are therefore more likely to read and mimic other people's emotional expressions. Consequently their emotional experience is more influenced by the afferent feedback, which results in stronger emotional convergence [12,13]. In theory strong transmitters of emotion demonstrate insensitivity to the emotions of others compared to strong receivers. However, Hatfield et al. suggest that these characteristics are not mutually exclusive. We now review studies concerning sender and receiver differences.

3.1.1.1 Senders

Sullins [28] found evidence for individual differences in emotional (nonverbal) expressiveness. Based on social comparison theory, their study additionally provides evidence for the relationship between these differences and the ability to infect another person with emotions. Subjects who scored high in nonverbal expressiveness had more influence on the emotions of the unexpressive subjects in a dyadic setting than vice versa.

3.1.1.2 Receivers

Doherty [6] attempted to develop a measure of individual differences in susceptibility to emotional contagion. The study resulted in the now commonly used Emotional Contagion Scale. One of the methods they used for the validation of the Emotional Contagion Scale was a comparison with other measurements of potentially related psychological concepts. This analysis showed that susceptibility to emotional contagion was positively associated with amongst other things emotionality and sensitivity to others and negatively associated with self-assertiveness and emotional stability.

A study done by Laird et al.[18] demonstrates a relation between individual differences in so called 'cue responsiveness' (the degree to which a person is affected by his/her own expressions) and emotional contagion. Participants that were more responsive to self-produced cues proved to be more likely to feel the emotions of those they mimic and thus were more susceptible to emotional contagion. Additional support for this effect was found in a different study by Doherty [8]. More recent research done by Papousek et al. [24] corresponds with- and complements most of these findings. They used self-reports for emotional contagion and

measured physiological indicators for emotional arousal (cardiovascular measures). Interesting is that both methods had almost identical results; participants who were strong emotion regulators and weak at emotion perception showed the weakest emotional contagion to sad emotions. Participants who were weak emotion regulators and were good at emotion perception showed the strongest responses to happy (cheerful) emotions.

These findings are in line with the idea that there is a difference between people with a strong tendency to regulate one's emotions reducing one's susceptibility to emotional contagion and people strong in perceiving emotions of others reacting more to these emotions resulting in high contagion susceptible.

3.1.2 Gender

In light of the findings that individuals differ in the degree to which they are good senders of emotion and the degree to which they are good receivers of emotion, it is also found that there is a difference between men and women regarding this construct. One example is a study of gender differences in facial reactions to facial expressions [5] by Dimberg and Lundqvist. They found differences in facial expressiveness of men and women when reacting to emotional stimuli. In this EMG-study women showed stronger imitative responses to angry and happy facial expressions than men, indicating that woman are more facially reactive than men are. Similar results were also found in a study by Lundqvist who complements these findings by investigating them in the context of primitive emotional contagion [22] and providing evidence consistent with the theory that facial emotional expressions are contagious.

Hatfield et al. theorize that females tend to be more susceptible to emotional contagion than males are and that this is amongst other thing due to traditional gender roles; woman are taught by the way they are socialized to be more sensitive to others' emotional displays as compared to males. The following studies provide supporting empirical evidence regarding this theory.

In the context of primitive emotional contagion Doherty et al. [6] found compelling evidence that women are more susceptible than men to the emotions of others and thus to emotional contagion, for both positive and negative emotions. Women from a variety of occupations reported being more susceptible to emotional contagion then men. These results were consistent with the findings in a second study where they used judges' ratings to measure the actual responsiveness to other's emotions. The judges rated women as displaying more emotional contagion than did men. A study by Stockert [27] came up with similar results additional showing that women also reported more intense emotion than men after watching emotional videos, significantly so for happiness.

A number of studies regarding the adaptation of the emotional contagion scale within a different culture resulted in additional compelling evidence supporting the theory that gender is a moderating factor for emotional contagion; Most of them conclude with almost identical findings as found with the original version. [7,16,20,21]

Within the data used for this review we found one interesting result by Hsee et al. [14] that is not in conformity with previously mentioned findings. Although gender was not included in the original design in their study but later taken into account, they found no significant gender differences in emotional contagion when showing participants another person's happy and sad expressions.

In conclusion, and in line with Kevrekidis [16] we can state that that although more research is needed to explore if gender differences in emotional contagion exist, these differences must be taken into account when studying emotional contagion. Woman tend to be better in transmitting and receiving emotions as compared to men, and therefore are more susceptible to emotional contagion.

3.2 Interpersonal Factors

3.2.1 Similarity

Perceived similarity is a factor moderating contagion. A basis for the assumption of this effect can for example be found in the social comparison theories, for it is known that emotional contagion can happen through social comparison processes where people evaluate their affective state in comparison with that of other people and their environment to come with an appropriate response [11,28]. Although similarity as a single construct has been shown to influence contagion in a study by Paukert et al. [25], other studies indicate that a differentiation can be made with regards to types of similarity.

3.2.1.1 Attitudinal Similarity

Stockert [27] specifically researched perceived similarity and emotional contagion. She investigated similarity in a attitudinal context using attitude questionnaires and assigned subjects to either a similar or dissimilar partner, hypothesizing that similarity will lead to increased emotional contagion. Additionally she took dissimilarity into account, questioning whether this would have the opposite effect or maybe would even reflect in the induction of opposite (discordant) emotions. She proposed that similarity would have a positive effect on emotional contagion regarding happiness and sadness and that dissimilarity would not lead to discordant emotions within the research setting; subjects in the dissimilar condition were hypothesized to show less emotional contagion then subjects in the neutral condition and the subsequent similar condition. The results partially supported the proposition. Although there were a number of seemingly random effects hampering theoretical interpretations, a significant positive relation was found both by judge's ratings of facial expressions and subject's self reports between perceived attitudinal similarity (and subsequent attraction) and the contagion of happiness. However the results did not support the same effect for the contagion of negative emotions (the sadness condition). In this context identical results were also found by Paukert et al. [25]. Regarding dissimilarity it was found that although dissimilar subjects tended to catch more emotion then expected (the results were close to the controls in one of the measures), the overall results show that in this research setting the dissimilar subjects did not experience discordant emotions compared to subjects in the similar condition.

3.2.1.2 Situational Similarity

Sullins' [28] focused on similarity and contagion in specifically a situational context. One of the conditions incorporated in their 3x3 study design was the pairing of the participant with a relevant other; a person who they believed was going to engage in a similar situation, opposed to the irrelevant other condition where the participant was paired with a person whom they believed was there for a different reason. The results indicate that the moods of

participants who were experiencing the same situation as their partner were most likely to converge compared to the participants in the irrelevant other condition or control group. The latter two did not show significant differences in their scores, which can be interpreted as an absence of the reversed effect; dissimilarity having a negative effect on emotional contagion.

Interesting is the fact that in a later study done by Gump et al.[11] threat and perceived situational similarity was manipulated to investigate affiliation and emotional contagion specifically in threatening situations. They predicted that threat would increase the tendency for people's emotions to be influenced by the emotions of others, especially when facing the same situation. Although the presence of emotional contagion was ascertained, and the predictions were confirmed regarding the results for mimicry, they found no evidence that either threat or situational similarity was a significant moderator for emotional contagion. They conclude with the suitable statement that: 'although it would be premature to conclude that perceived situational similarity of the other's situation plays no role in emotional contagion, the importance of such perceptions may be less fundamental than has been assumed by social comparison theorists.' The aforementioned results of Stockert in a sense support these findings. Although similarity was addressed within a slightly different context and the studies focus on different emotions Stockert provides evidence for the presence of a difference in strength of the effect of different moderating factors for emotional contagion, by showing that for example susceptibility had more impact on emotional contagion compared to perceived similarity.

3.2.1.2.1Group membership

Situational similarity can also be interpreted more specifically in terms of group membership; questioning whether a person belongs to the same group or not. Van der Schalk et al. specifically investigated if group membership moderates emotional mimicry and contagion [32]. They found that expressions of anger and fear were mimicked to a greater extend by subjects in the in-group condition then subjects in the out-group condition. Interesting is the fact there was no such effect found for the mimicry of happiness. And although they offer some strong possible reasons for the lack of this effect it is interesting to note that in this context these results prove similar to for example the results found by Stockert. Van der Schalk et al. furthermore found some evidence for a divergence effect. Although these results were somewhat weaker than those for the convergence effect they found in one study that the expression of angry emotions in the out-group condition resulted in more self-reported fear and that the expression of fear in the out-group condition resulted in the experience of aversion which was found both in the subjects' self reports and their emotion display. An interesting observation was that although the effect was found for the mimicry of emotional expressions, they found no significant correlation between mimicry and self-reported emotions; thus for emotional contagion. Nevertheless they argue that the 'research shows emotional convergence is more likely to occur when individuals share a group membership.'

More indicative evidence for this divergence effect of dissimilarity was also found by Epstude et al. [10] They utilized the concept of similarity both in the context of group membership and in a context where subjects were primed to specially look for similarities or dissimilarities. Within both these contexts they found evidence confirming their hypothesis; subjects focusing on similarities experienced more concordant emotions when being confronted with pictures of a person pre-rated as conveying a specific (positive, neutral or negative affect), while subjects focusing on dissimilarities experienced more discordant mood in the same condition.

Although the amount of evidence is limited, overall these studies provide evidence that contagion is stronger in the in-group condition. Furthermore they show that emotional divergence is also a possible effect.

3.2.2 Social power

In the earlier mentioned historical review on social contagion [19] Levy et al. argue based on indicative evidence that contagion in the context of social status is most likely to happen in a top down fashion; from high status individuals to low status individuals.

Anderson et al investigated emotional convergence in the context of relationships [1]. Two studies provided similar results; one with partners in romantic relationships and one with college roommates. Examining amongst other things personality and emotional experiences, during two laboratory sessions they found that the low power subject was influenced to a greater extent by the emotions of his/her partner, then vice versa.

On first sight the statement by Sy et al. [29] that their findings are 'consistent with recent research showing that high status individuals are more likely to transmit their moods to low status individuals than vice versa' seems to support the findings of Anderson et al. They investigated the effect of a leader's mood on that of members of the group by priming a leader with a positive or negative mood before engaging in a complex group-task. Nevertheless, as they also state with regards to limitations of the study, the fact is that they only investigated contagion in the direction of a high power condition to a low power which consequently makes conclusions about the moderating effect of power on emotional contagion impossible. They propose that it is very possible that contagion can also happen in the opposite direction and with a different effect-strength.

Contrary to aforementioned research Sestak et al. [26] investigated the influence of social status on emotional contagion in a direct manner explicitly testing for a moderating effect. They collected trait based data from a number of dyads consisting out of a supervisor and subordinate working in a global manufacturing company which subsequently provided data regarding amongst other things their emotional state over a period of two weeks. In general their observations support the theory that the direction of emotional contagion in a group possibly goes from a high power to low power; at least within subordinate-supervisor context. Early research done by Hsee et al. also directly focused on the assumed relation between power and emotional contagion [14]. Test subjects were assigned to the role of teacher or the role of learner. The latter representing the powerless condition. Subjects were led to believe that the teacher had to teach the learner a list of words and had the power to punish the learner by administering an electric shock when he or she saw fit. They theorized that subjects in the low power condition would be more affected by the emotions of the other (powerful) subjects then vice versa. They found no evidence for this effect examining the subject's self-reports of the experienced emotion. However it is interesting that the results of the judges' ratings showed an significant effect in the opposite direction. Seemingly the powerful were more susceptible to the emotions of the subjects in

the low power condition. A possible discrepancy between the subjects self-reports and the judges' ratings further support the findings (the subjects' self reports seemed to be less reliable as a measurement for their feelings).

Kimura et al. [17] found almost identical results; participants were more susceptible to the emotions of juniors whose social power was low then to seniors representing high social power. However their results are somewhat debatable due to the lack of an initial manipulation check for social status and a questionable method.

Nevertheless it is interesting that both studies directly addressing the effect of social power on emotional contagion following a similar hypothesis, report similar findings in terms of a inverse effect, suggesting high social power increases susceptibility to emotional contagion. Overall, evidence indicates that social power is a moderating factor concerning for emotional contagion. However due to the fact that there is indicative evidence for two different directions of this effect it is hard to make any sound conclusions concerning whether it contagion is more likely to go from low power individuals to high power individuals or vice versa, of which the latter effect is predominantly theorized.

3.2.2Liking and Intimacy

Hatfield et al. theorized that emotional contagion is most likely to occur in relationships involving power or love. The latter concept is closely related with liking and intimacy which some studies indeed suggest can be a moderating factor for emotional contagion. Kimura et al [17] successfully manipulated intimacy by making the subjects assume one of the following roles: friend, acquaintance, senior junior and found evidence suggesting that participants were more susceptible to those with whom they shared the highest degree of intimacy. The effect was only found for experiences of positive emotions, but they argue that it is plausible that the absence of this effect for negative emotions might be due to Japanese display rules.

Another study done by McIntosh [23] provided similar results regarding mimicry of facial emotional expressions. However their results did not show that the evoked emotional expressions directly caused the found contagion and therefore conclusions about the effect of liking on contagion cannot be made. Nevertheless it is interesting to note that the results partially support previously mentioned findings.

3.3 Miscellaneous

3.3.1Pre-existing mood

A lot of experimental studies regarding emotional contagion utilize emotional priming as a means of control for a specific emotion. A logical continuation of this idea can be that mood can moderate emotional contagion. The following study by Hsee et al. specifically focused on this research-question. Hsee et al. investigating the impact that pre-existing mood has on an individuals susceptibility to emotional contagion [15]. Participants were primed with a happy, neutral or sad mood by letting them recall a series of events consistent with the specific condition after which they were asked to view a happy or sad video. The results suggest that pre existing mood can have a minimal impact on emotional contagion. Evaluating the judges' ratings of the facial expressions they found borderline significance. Subjects in the happy condition showed more attention to the emotions of the target person and were more likely to mimic the expressions of the target person. This can be interpreted as weak evidence for their hypothesis that people are most susceptible to emotional contagion when they are happy.

4. MULTI-AGENT BASED MODEL

To investigate emotional contagion within groups, we have developed a multi-agent simulation in which the agents influence each other based solely on the factors found in the review. The simulation system itself is in essence a simple dynamical system composed of the individual agents that populate a continuous 2D space. Time increments, *dt,* advance the state of the simulation. This section will discuss this agent model in more detail.

Behavior model.

Each agent has a behavior model that defines how it moves through space. Initially an agent is set at a starting location. When the simulation starts, the behavior model defines how the coordinates of an agent change. Currently we have only one model implemented that represents a simple way of wandering through space based on constant movement with a random change in direction induced by a set timer or a collision with another agent.

The emotion model.

In the greater part of the studies the investigated contagion factors are limited to two basic emotions; namely positive emotions and negative emotions, without clear differentiation. As such, we have used a factor-based emotion representation based on the Pleasure Arousal Dominance factor model. Each factor, *P*, *A,* and *D* can have a value between -1 and 1. The emotional state decays in a linear fashion based on a constant change towards (0,0,0).

The contagion model.

To minimize assumptions around contagion, a direct interpretation of the factors and their effects on contagion was used. We assume that emotional contagion flows in the commonly theorized direction from high social power subjects to those with a lower social power [1,24]. The model is a description of what kind of effect a specific factor has on emotional contagion; a positive, negative or null effect. In this context contagion is defined by the effect of a specific factor on susceptibility and on so-called contagability of an agent. Nevertheless one factor had to be assumed, i.e., distance. Just like the other factors, the importance of distance as a factor can also be defined per agent in its personality.

Personality.

For the model to allow easy configuration of agents, separate of the definition of contagion factors and their effects, each agent has a personality type. In essence a personality is simply a vector of contagion factor weights with some additional agent variables such as power and group belonging, needed to calculate the effect of factors like social power and group membership. This enables us to vary the size of a specific factor's effect per individual agent. For the sake of clarity we call this a personality type.

Table 2. Overview of contagion factors used in the simulation model. Columns refer to factor effects on susceptibility and contagability, as well as three example male personalities for a high power leader ($Pers_a$), medium power leader ($Pers_b$) and a follower ($Pers_c$), as used in the pilot study described below.

Factor	Susc	Cont	$Pers_a$	$Pers_b$	$Pers_c$
Transmitter	0	1	1.0	1.0	0.5
Reciever	1	0	0.0	0.0	0.5
Female	1	1	0.0	0.0	0.0
Group_membership	1	0	0.5	0.5	0.5
Low_power	1	0	1.0	1.0	1.0
High_power	0	1	0.0	0.0	0.0
Distance	1	0	0.1	0.1	0.1
POWER			1.0	0.5	0.0
GROUP			A	A	A

Contagion

Contagion occurs from an agent *a* to an agent *b* after each *dt* and only if the distance $d_{ab} < maxViewingDistance$ as follows:

$$c_{ab} = \frac{\sum_F Succ_f \bullet Pers_{a(f)} \bullet State_{a(f)}}{|F|} \bullet \frac{\sum_F Cont_f \bullet Pers_{b(f)} \bullet State_{b(f)}}{|F|}$$

Succ and *Cont* refer to susceptibility and contagability relations between factors (see table), while *Pers* refers to the personality weights of an agent (see table for example personalities). The *State* defines to what extent factors play a role in the current situation of the agent. Currently all factors always play the same role (i.e., $State_f=1$), except for the following. If both agent *a* and *b* are in the same group:

if $Group_a=Group_b$ then $State_{a(group)}=State_{b(group)}=1$

Closer agents show stronger contagion. Actual distance between agents influences contagion as follows:

$$State_{a(distance)}=State_{b(distance)}=1-(d_{ab}/maxViewingDistance)$$

High and low social power are each others inverse in our current setting and are calculated as follows:

$$State_{a(high\_power)}=if\ (power_a>power_b)\ power_a-power_b\ else\ 0$$

$$State_{a(low\_power)}=if\ (power_a<power_b)\ power_b-power_a\ else\ 0$$

Eventually, the emotional state of agent *a* is influenced using a simple update function:

$$Emo_{a(t+dt)}=Emo_{a(t)}+dt*c_{ab}*(Emo_{b(t)}-Emo_{a(t)})$$

Obviously this is done for each agent pair *a* and *b*, and for each timestep in the simulation. By controlling *dt*, the resolution and speed of the simulation can be varied.

5. PILOT STUDY

For a first test-case we chose to simulate the spread of elatedness (intensely arousing and positive affective state) in the context of a recreational environment filled with students, induced by one individual. The almost instant spread of laughter and unrest amongst students after for example a funny remark by one individual is a phenomenon well known by teachers. Following Hatfield's reasoning in their theory of primitive emotional contagion regarding individual differences in emotional traits this specific initiator is likely to be a person who is very good at transmitting emotions and consequently has insensitivity to the emotions of others. Based on this reasoning the only effective factors varied between the two types of individuals is the tendency to be transmitter or a receiver. The simulation is constructed with the personality $Pers_a$ (the initiator) and $Pers_c$ (the other students). The recreational room is 10 x 10 meters, and is filled with 10 students at random locations and one initiator. The *maxViewingDistance* is set to 3 meters.

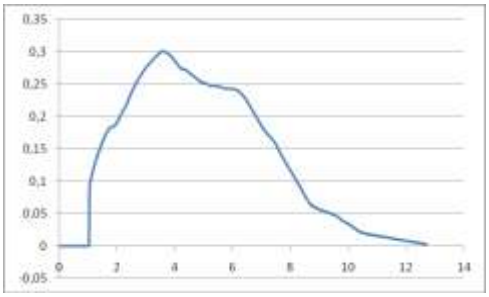


Figure 1. Initiator added in the middle. On the x-axis: time, on the y-axis: mean group Pleasure (*P*) intensity.

Around *t*=1 the initiator becomes happy (high *P*, *A* and *D* affective factors). When the initiator is in the middle of the group, as expected the results show a fast increase in the mean group happiness, continued by a gradual decrease of the emotion until all agents including the initiator reach the starting emotional state which is neutral. This is due to natural emotional decay.

In a second test run we generated the same situation but now the initiator was placed in a more secluded area of the room. Again the results show a similar pattern compared to previous test-run with the only difference that the overall mean scores for group happiness are lower, as expected due to less contagion induced by the increased distance between the initiator and the students.

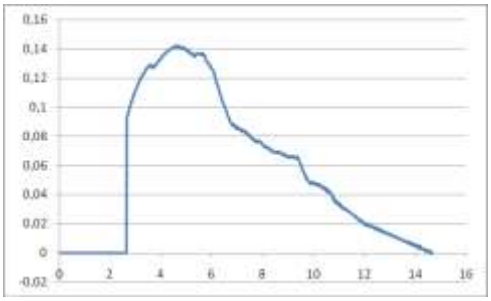


Figure 2. Initiator added in secluded area, axes same as above.

Running the same test but now with multiple initiators present resulted in a short increase of the emotion every time the initiator was added although with a smaller maximum intensity for every new initiation. The first contagion event (*t*=1) is similar to Figure 1 as it is a similar setup (10 students, 1 initiator). The second event (*t*=13) results in less contagion due to the presence of two initiators of whom only one becomes happy. The third shows the same effect when three initiators are present of which one becomes happy. This result can be easily explained. Neutral initiators are still bad receivers and are not influenced by an initiator who is happy. This means the neutral initiators influence the group with a neutral state functioning as a "resistor". It therefore becomes less likely for contagion to happen by happy initiator.

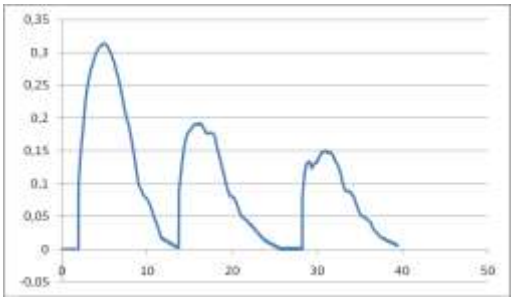


Figure 3. Multiple initiators present, axes as above.

In the above settings, contagion is a result the factors transmitter/receiver and a difference in social power between, although similar results would be obtained if the factor power was dropped as in the current setting power only makes contagion stronger (the transmitter is also the high power individual).

To expand the test-case we investigated the factor power in more detail. In addition to the initiator, who has leader characteristics (high power) we have a sub-leader personality $Pers_b$ (medium power). All other factors are the same for both personalities.

In a simulation similar to previous one, but now with the sub-leader becoming happy in the presence of an initiator, the effect of the sub-leader is strongly reduced (Figure 4, second contagion event) compared to when the sub-leader would be present alone (Figure 5, first contagion event). However, the effect of contagion of the initiator is amplified in the presence of a sub-leader (Figure 5, second contagion event) compared to in the presence of another initiator (Figure 3, second contagion event). The explanation is that the initiators and sub-leaders are sensitive to power. The sub-leader is influenced by the leader effectively functioning as an amplifier for the group, but the initiator is not influenced by the sub-leader still functioning as a resistor.

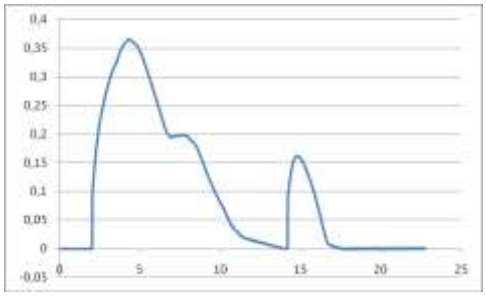


Figure 4. Initiator and sub-leader added respectively.

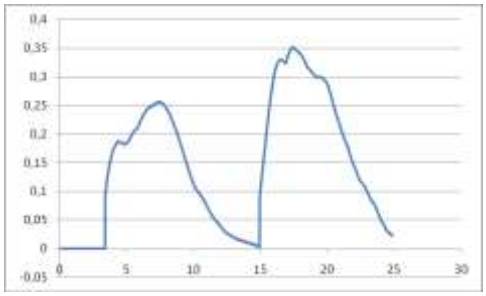


Figure 5. sub-leader and leader added respectively.

To expand this test-case even further within the context of students in a recreation room, simulated an annoyed teacher who enters the room after the initiator becomes happy. We used the same personality for the teacher and for the initiator ($Pers_a$). However, the teacher's initial emotional state is either negatively calm or negatively aroused to simulate a calm negative reaction and a very angry reaction.

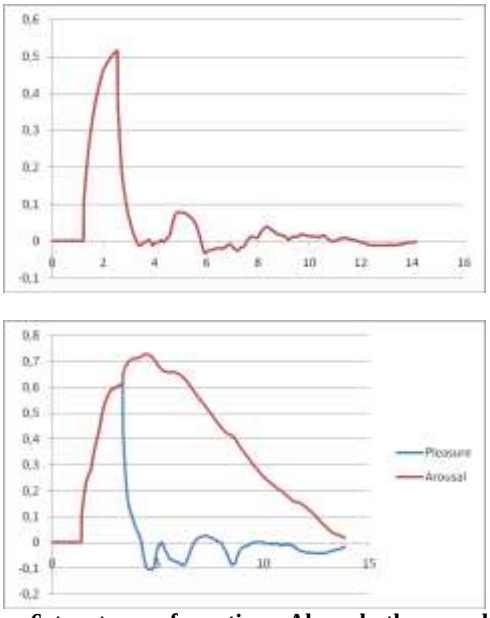


Figure 6. two types of reactions. Above both arousal and valence are reduced to neutral (in exactly the same way) due to calm negative teacher intervention. Below, arousal increases due to angry teacher intervention.

An interesting and explainable observation can be made. When the annoyed teacher reacts by expressing his negative emotions in a calm manner we can see that this results in a quick nullification of the spread of the initiated positive emotion and arousal. A negative but aroused reaction however, results in a nullification of the effect on pleasure spread, but not on arousal. The situation has not calmed down, only made less positive.

6. DISCUSSION

In this paper we show for the first time that a straightforward factor-based model of contagion can be used to study the details of how and due to which factors contagion spreads through a group. Current efforts focus on a validation of our approach. To this end we are in cooperation with social psychologists in order to investigate the model and use of the simulation system for both hypothesis testing and generation. We feel the strength of the system is the small number of additional assumptions needed to study contagion, other than those based on psychological findings. Although of preliminary nature, the pilot study is a clear example of the many potential settings in which our approach can be used to model and study emotional contagion. Other than simulating contagion in a multi-agent setting for the sake of understanding emotional contagion on a psychological level, we feel our review of factors is an important basis for the development of artificial agents that make use of or take into account contagion between agents and humans, such as the work recently published by Tsai et al [30], and Bispo and Paiva [3]. The novelty of our modeling approach is, when comparing it to existing models, that we are able to systematically vary moderating factors for contagion while other address the process of contagion in a relatively abstract manner. Further, we only introduce those factors that have shown to be moderators according to actual psychological experiments.

Our results show that modeling emotional contagion based on experimental evidence from psychology can give insight in the dynamics of emotional contagion within a group.

7. REFERENCES

[1] Anderson, C. et al. 2003. Emotional convergence between people over time *Journal of Personality and Social Psychology*. 84, 5 (2003), 1054–1068.

[2] Barsade, S. 2002. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*. 47, 4 (2002), 644–675.

[3] Bispo, J. 2009. A model for emotional contagion based on the emotional contagion scale. In Proc. Of Affective Computing and Intelligent Interaction 2009.

[4] Bosse, T., Duell, R., Memon, Z., Treur, J., & van der Wal, C. 2009. A Multi-agent Model for Emotion Contagion Spirals Integrated within a Supporting Ambient Agent Model. In: LNCS 5925, 48-67.

[5] Dimberg, U., & Lundquist, L.-O. 1990. Gender differences in facial reactions to facial expressions. *Biological Psychology*, 30(2), 151-159.

[6] Doherty, R. et al. 1995. EMOTIONAL CONTAGION: Gender and Occupational Differences. *Psychology of Women Quarterly*, 19, 3, (1995), 355-371.

[7] Doherty, R. W. 1997. The Emotional Contagion Scale: A Measure of Individual Differences. *Journal of Nonverbal Behavior*, 21(2), 131-154.

[8] Doherty, R. W. 1998. Emotional Contagion and Social Judgment. *Motivation and Emotion*, 22(3), 187-209.

[9] Durupinar, F. 2010. From audiences to mobs: Crowd simulation with psychological factors. PhD Thesis, BILKENT Univ.

[10] Epstude, K. and Mussweiler, T. 2009. What you feel is how you compare: How comparisons influence the social induction of affect *Emotion*. 9, 1 (2009), 1–14.

[11] Epstude, K. and Mussweiler, T. 2009. What you feel is how you compare: How comparisons influence the social induction of affect *Emotion*. 9, 1 (2009), 1–14.

[12] Hatfield, E., Cacioppo, J and Rapson, R.. 1992. Primitive emotional contagion. *Emotion and social behavior.Review of personality and social psychology*, 14, 151-177.

[13] Hatfield, E., Cacioppo, J and Rapson, R.. 1994. Emotional contagion. Cabridge Univ. Press.

[14] Hsee, C. et al. 1990. The effect of power on susceptibility to emotional contagion. *Cognition and Emotion* (1990).

[15] Hsee, C. , Hatfield, E., Carlson, JG. 1991. Emotional contagion and its relationship to mood. *Technical Document. Univeristy of Honolulu*.

[16] Kevrekidis, P. et al. 2008. Adaptation of the Emotional Contagion Scale (ECS) and gender differences within the Greek cultural context. *Annals of General Psychiatry*. 7, 1 (2008), 14.

[17] Kimura, M. and Daibo, I. 2008. The study of emotional contagion from the perspective of interpersonal relationships. *Social Behavior and Personality: an International Journal*, 36, 1, (2008), 27-42.

[18] Laird, J.D. et al. 1994. Individual differences in the effects of spontaneous mimicry on emotional contagion. *Motivation and Emotion*. 18, 3 (Sep. 1994), 231–247.

[19] Levy, D.A. and Nail, P.R. 1993. Contagion: A theoretical and empirical review and reconceptualization. *Genetic, Social & General Psychology Monographs, May93, Vol. 119 Issue 2, p235, 50p, 5 Charts*. (1993).

[20] Lundqvist, L.-O. and Kevrekidis, P. 2008. Factor Structure of the Greek Version of the Emotional Contagion Scale and its Measurement Invariance Across Gender and Cultural Groups. *Journal of Individual Differences*. 29, 3 (Jul. 2008), 121–129.

[21] Lundqvist, L.-O. 2006. A Swedish adaptation of the Emotional Contagion Scale: Factor structure and psychometric properties. *Scandinavian Journal of Psychology*. 47, 4 (Aug. 2006), 263–272.

[22] Lundqvist, L.O. 1995. Facial EMG reactions to facial expressions: a case of facial emotional contagion. *Scandinavian Journal of Psychology*. 36, 2 (Jun. 1995), 130-141.

[23] McIntosh, D. 2006. Spontaneous facial mimicry, liking and emotional contagion. *Polish Psychological Bulletin*. (2006).

[24] Papousek, I. et al. 2008. The interplay of perceiving and regulating emotions in becoming infected with positive and negative moods. *Personality and Individual Differences*. 45, 6 (Oct. 2008), 463–467.

[25] Paukert, A.L. et al. 2008. The Role of Interdependence and Perceived Similarity in Depressed Affect Contagion. *Behavior Therapy*. 39, 3 (Sep. 2008), 277–285.

[26] Sestak, N. 2008. Psychological contagion within the supervisor-subordinate dyad: An experience sampling investigation of mood and job attitude contagion at work. (2008).

[27] Stockert, N. 1994. Perceived similarity and emotional contagion. Thesis (Ph. D.)--University of Hawaii at Manoa.

[28] Sullins, E.S. 1991. Emotional Contagion Revisited: Effects of Social Comparison and Expressive Style on Mood Convergence. *Personality and Social Psychology Bulletin*. 17, 2 (Jan. 1991), 166–174.

[29] Sy, T. et al. 2005. The Contagious Leader: Impact of the Leader's Mood on the Mood of Group Members, Group Affective Tone, and Group Processes *Journal of Applied Psychology*. 90, 2 (2005), 295–305.

[30] Tsai, J., Bowring, E., Marsella, S., Wood, W., Tambe, M. 2012. Emotional Contagion with Virtual Characters. Proc. of AAMAS 2012, in Press.

[31] Tsai, J., Bowring, E., Marsella, S., & Tambe, M. .2011. Empirical Evaluation of Computational Emotional Contagion Models. In Proceedings of IVA, LNCS 6895, 384-397.

[32] van der Schalk, J. et al. 2011. Convergent and divergent responses to emotional displays of ingroup and outgroup *Emotion*. 11, 2 (2011), 286–298.

[33] Broekens, J., Netten, N. & DeGroot, D. 2004. Consistent Dynamic-Group Emotions for Virtual Agents. Proc. of the 16th BNAIC, 99-106.

Moral Appraisal and Emotions

Cristina Battaglino
Dipartimento di Informatica
C.so Svizzera 185, Università
degli Studi di Torino, Italy
battagli@di.unito.it

Rossana Damiano
Dipartimento di Informatica
and CIRMA
C.so Svizzera 185, Università
degli Studi di Torino, Italy
rossana@di.unito.it

Leonardo Lesmo
Dipartimento di Informatica
C.so Svizzera 185, Università
degli Studi di Torino, Italy
lesmo@di.unito.it

ABSTRACT

In this paper, we propose a model of moral appraisal in which self and other's actions are evaluated according to the agent's values, making emotional states such as "pride" or "reproach" arise in response to situations in which the agent's values are put at stake.

In an interactive context, the relationship between values and emotions is highly relevant: for an agent to be really empathic with the user, it should be able to share the user's values and feel emotions accordingly in response to the actions performed during the interaction.

In order to exemplify the model and test its effectiveness, we resort to a well known narrative situation, annotated according to a BDI–based ontology of story and character. By encoding the model as a set of SWRL rules, then, we apply it to the example situation, showing the viability of the model.

Categories and Subject Descriptors

I.2.m [**ARTIFICIAL INTELLIGENCE**]: Miscellaneous; I.2.1 [**ARTIFICIAL INTELLIGENCE**]: Knowledge Representation Formalisms and Methods

General Terms

Languages, Theory

Keywords

moral values, emotion models, virtual agents

1. INTRODUCTION

In the last two decades, emotions have raised a wide interest in the agent community: mostly inspired by cognitive models of emotions [30, 23], various research efforts have explored the role of emotions in artificial agents, with perspectives that range from the social and relational aspects of emotions [18, 19, 33] to their expression [25].

Since the pioneering work by [1], stories have provided an ideal testbed for artificial models of emotions. This trend is in line with narratological and drama studies, that have often acknowledged the role of emotions in stories, from the Age of Enlightement [9] to contemporary film theory [42] and aesthetics [42, 17].

In this paper,[1] we focus our attention on the moral dimension of emotions, taking inspiration from the cognitive model of emotions by Ortony, Clore and Collins [30] (OCC). According to this model, the agent's "standards", i.e., the agent's "beliefs in term of which moral and other kinds of judgmental evaluations are made" affect the evaluation of self and others' actions. Actions that meet the agent's standards are deemed praiseworthy, and their execution triggers emotions like pride and admiration. Conversely, blameworthy actions trigger emotions like shame and reproach. These emotion types have received less attention than other emotions, since they require – beside the integration with the agent's beliefs, goals and intentions – the presence of some deontic component in the agent architecture.

Despite this difficulty, however, moral aspects – and their related emotions – are of paramount importance in stories. According to Bruner [4], "moral commitments" represent the privileged object of stories, in a theoretical framework where stories are intended as an instrument for the transmission of culture in a society. In narrative and drama theory, the notion of moral values, first stated in Egri's definition of "drama premise" [10], underpins most of the subsequent work conducted in scriptwriting [5], until the recent formulation stated by McKee [27] about cinematographic stories.

In this work, building on the assumption that stories provide a useful paradigm for agent theories and design, we propose an account of "moral appraisal", i.e, the process by which actions are appraised by an emotional agent according to her/his "standards", as defined in [30]. In particular, we model the praiseworthiness (and blameworthiness) of agent's actions by resorting to the notion of "moral value" [44] and apply this model on a well known example in the narrative domain.

Within a BDI architecture, values provide motivations for goal formation and selection [8], thus endowing agents with the capability to reason about their own and others' motivations for actions in moral terms. Here, we equate standards to values, thus obtaining an explicit connection between the values acknowledged by an agent and the 'moral' emotions she/he feels in response to the actions performed by her/himself or by other agents. Embedding this model into virtual agents enables to model the range of emotions elicited by shared or conflicting values, and to make interac-

Appears in: *Proceedings of the Workshop on Emotional and Empathic Agents, in the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, June, 4-8, 2012, Valencia, Spain.

[1]This work is part of the Cadmos project (Character-based Annotation of Dramatic Media ObjectS), Regione Piemonte, Poli di Innovazione, POR-FESR 2007-2013.

tive agents emotionally respond to the values expressed by the user, or encoded in some user model.

This paper is structured as follows. After surveying the related work (section 2), in Section 3 we describe how the notion of "standards" can be modeled through the operational account of values provided by [8]. Agent's values, that in [8] are linked to the formation and selection of goals, are extended to the evaluation of goals, yielding a model of moral appraisal. We then describe an implementation of the model as a set of SWRL rules (Section 4), built on the top of the Drammar ontology of character and story [7]. Finally, we exemplify the model on a well known story episode (Section 5). Discussion and conclusions end the paper.

2. RELATED WORK

In the last years, researchers have invested much effort in understanding what emotions are and what kind of relation they have with cognitive processes. Several authors argue that emotions play a fundamental role in the social and cognitive functions of the human brain [29] and that they are a necessary component of an intelligent system [28]. Consequently, the study of computational models for emotions in intelligent agents is crucial to create believable, empathetic and lifelike agents. Many works tried to integrate computational models of emotions in a cognitive architecture for intelligent agents [35] [12] [24] with the aim to insert emotions in BDI (Belief-Desire-Intention) agents [3].

In the field of psychology, there are many theories about emotions and we can differentiate three main approaches: physiological (where emotions are related with body changes [43] [11]), dimensional (in which emotion are conceived as '*core-affect*' [45] [36]) and appraisal (in which cognitive processes are involved in the generation of emotions [31][22][38]). Indeed, due to the fact that Appraisal theories focus on the relation between emotions and cognition, most computational models are based on the appraisal theory and try to integrate emotions in BDI agents. According to Appraisal theories, cognitive processes have the function of building a mental representation of the situation in which a person is involved. This representation, often termed *person-environment relation*, is not limited to the external environment, but includes also the internal disposition of a person as goals, desires, intentions, norms, moral rules. Emotions rise from appraisal of the person-environment relation according to appraisal dimensions that are defined in the theory (i.e. desirability of an event).

In the OCC theory [31], the person-environment relation is represented by goals, standards and attitudes; appraisal dimensions are represented by *desirability* (or undesirability) of an event, *praiseworthiness* (or blameworthiness) of an action, *liking* (or disliking) of an object. In OCC, emotions are defined as valenced reaction to events, agents and objects and they are divided in three basic classes:

- Event-based emotions, that arise from reactions to events (i.e. being pleased (or displeased) about the event with respect to goals).
- Attribution emotions, that arise from reactions to agents (i.e. approval (or disapproval) of an action performed by an agent with respect to standards).
- Attraction emotions: reactions to objects (i.e. liking (or disliking) of an object with respect to attitudes).

While in the OCC model goals are conceived as states of affairs that one wants to obtain, standards concern the state of affairs that one believes ought to obtain. Standards represent the beliefs in terms of which moral and other kinds of judgmental evaluations are made, like *you ought to have tried harder* or *you ought not to do things that upset other people*. In this paper, the concept of moral values is very similar with the concept of standards in OCC theory.

Many computational models don't take into account the link between emotions, standards and goals. For example, EM [35] adopts the OCC model of emotion over plan based agents using domain-independent approaches. EM is very close to OCC and the differences are due to implementation choices. For instance, appraisal of the person-environment relation with respect to an event is performed by checking whether a goal is met or not in the event. Standards are taken into account in EM, following the model of OCC, but their implementation is quite limited. The appraisal of praiseworthiness or blameworthiness of an action is limited to only two standards:*help-my-goals-to-succeed* and *do-not-cause-my-goals-to-fail*. These standards are related to goal and they do not cover moral standards. Moreover, it is not clear how other standards can be included in the model and how actions can be evaluated.
In ALMA [16], the Big Five theory of personality [21] is combined with OCC Model to generate both emotion and mood. However, appraisal is encoded in domain-dependent XML rules, thus failing to grasp general principles. Also in [12] domain-dependent rules are used for the appraisal of person-enviroment relation.
In EMA [24], the first fully-implemented framework for conversational agents, appraisal is formed by a set of independent processes that operate on a plan-based representation of person-enviroment relation, named *causal interpretation*. This work is mainly based on Smith and Lazarus theory [22], so standards are not modeled. The authors take into consideration as appraisal variables the responsibility and intention of the agent in performing an action. EMA is employed for only one agent mind, but in [40] the integration of EMA in Thespian [41] is described.
In WASABI [2], both primary and secondary emotions are modeled: primary emotions are 'core affect' [36], secondary emotions are obtained by high cognitive processes. For secondary emotions the authors use the OCC model, but they take into account only Prospect-based emotions, like hope and fear, based on expectation of an event.

In this paper, we focus on Attribution (of responsibility) emotions, that rise from the approval (or disapproval) of an action according to standards that characters have in their mental state. In this class, the OCC model defines four type of emotions:

1. *Pride* arises from the approval of one's own praiseworthy action (with respect to standards).
2. *Self-reproach* arises from disapproval of one's own blameworthy action (with respect to standards).
3. *Admiration* arises from approval of someone else's praiseworthy action (with respect to standards).

4. *Reproach* arises from disapproval of someone else's blameworthy action (with respect to standards).

We also consider the four compound emotions that are characterized by the conjunction of Attribution emotions and Event-based emotions, like Joy and Distress. Joy and Distress, called Well-being emotions, are defined as follows:

1. *Joy* arises from being pleased about a desirable event with respect to one's own goals.
2. *Distress* arises from being displeased about a undesirable event with respect to one's own goals.

The four compound emotions are:

1. *Gratification* arises from the approval of one's own praiseworthy action with respect to standards (pride) and from being pleased about a desirable event with respect to one's own goals (joy).
2. *Remorse* arises from the disapproval of one's own blameworthy action with respect to standards (self-reproach) and from being displeased about a undesirable event with respect to one's own goals (distress)
3. *Gratitude* arises from the approval of someone else's praiseworthy action with respect to standards (admiration) and from being pleased about a desirable event with respect to one's own goals (joy).
4. *Anger* arises from the disapproval of someone else's blameworthy action with respect to standards (reproach) and from being displeased about a unsiderable event with respect to one's own goals (distress).

Finally, the "Fortunes of Others" emotion type is related to the appraisal of desirable (on undesirable) events for other agents: based on her/his belief of other agents' goals, an agent feels "happy-for" or "sorry-for" another agent when event occurs that is desirable or undesirable given the other agent's goals.

3. A MODEL OF MORAL APPRAISAL

According to OCC, "attribution emotions" stem from the appraisal of an action as praiseworthy or blameworthy. In this Section, we describe how we model the appraisal of actions along the praiseworthiness dimension. Praiseworthiness and blameworthiness are defined on the basis of the compliance with the values of the appraising agents (here, equated to the notion of "standards" in OCC model)

3.1 Standards as Values

In previous work by [8], value–sensitive agents are modeled as BDI agents, augmented with the notion of value. An agent features a set of values, arranged into a subjective 'scale of values' [44]. Each value is associated with a set of conditions: when one or more conditions hold in the state of the world, the agent's value is put at stake. The value–sensitive agent monitors the state of the world for values at stake. When the agent realizes that some value is at stake, it modifies its commitment accordingly, by forming a goal (value–dependent goal) that contributes to re-establishing the value (or the values) at stake. Notice that, according to this model, the monitoring of values is carried out not only on what the agent believes to be the current state of the world, but also on the agent's expectations about the outcomes of the events and of the other agents' actions.

Here, we define the *praiseworthiness* of an action on the basis of the goal that motivates the action itself, mapping the notion of agent's "standards" on the notion of "values". So, in our model, an action is praiseworthy if it is motivated by a value–dependent goal and the value the goal depends on is acknowledged as such by the appraising agent. In other words, we assume that the praiseworthiness is not an intrinsic property of the action, but resides in the motivations that determine the agent's intention to execute it, i.e., the commitment to a goal. By doing so, we anchor into subjective values the interpersonal dimension of the so–called attribution emotions. The role of values is relevant not only for the appraisal of an agent's own actions, but also for the appraisal of other agents' behavior. If the appraising agent is the agent of the appraised action, this equates to saying that the agent considers her/his own action as praiseworthy only if she/he has formed the intention to execute the action in response to a value at stake. If the appraising agent differs from the agent of the appraised action, this equates to saying that the agent evaluates other agent's action according to her/his own values, praising that action only if she/he can ascribe to the other agent the value–dependent goal to re–establish a value at stake (that she/he shares with the other agent).

Conversely, the *blameworthiness* of an action is defined, in our model, on the basis of the effects it brings about in the state of the world. If an action puts at stake a value of the appraising agent, it is considered blameworthy, independently of the motivation of the action's agent to execute it. [2]

Clearly, for an agent to consider her/his own action as blameworthy, there must be a goal/value conflict inner to the agent. In case a conflict between the two evaluations arises, the highest–ranked value for the appraising agent determines the praiseworthiness judgment.

An important implication of the model by [8] is that, as a consequence of the fact that values are ordered into subjective scales, agents sharing the same set of values may react differently to the same situation, due to to different orderings of their values. In our mapping of standards onto values, however, this is not necessarily true: an agent may consider another agent's action praiseworthy even if the value at stake that determines the other agent's commitment does not have the highest priority – in the current state of the world – for the appraising agent.

In the simplest case, an agent just observes the actions of another agent and considers them as praiseworthy or blameworthy on the sole basis of their compliance with her/his own system of values. In a more complex model, the appraising agent tries to recognize the reason (goals, values) of the other agent for behaving in that way, so that putting a value deliberately at stake deserves blame, while the unwanted side effect of an action does not result in blame. More complex models could try to reconstruct the value system of the other agent, in order to ascertain (and discuss) the reasons for a possible misalignment with the value system of the apprais-

[2]Notice that, according to this model, an action can not be simultaneously appraised as praiseworthy and blameworthy if it implies a conflict between values. However, following [44], such conflict can be solved thanks to the scales of values.

	EXECUTED BY SELF		EXECUTED BY OTHERS	
praiseworhty action	PRIDE	+ JOY	ADMIRATION	+ JOY
		GRATIFICATION		GRATITUDE
blameworthy action	SELF-REPROACH	+ DISTRESS	REPROACH	+ DISTRESS
		REMORSE		ANGER

Figure 1: Attribution emotions in the OCC model.

ing agent. In all cases, however, the appraising agent should form the intention to re-establish the value at stake.

Although we focus on attribution emotions, we also describe the appraisal of events as desirable or undesirable, a dimension that determines the "well being" emotions in OCC model. Modeling well being emotions is necessary to encompass "compound emotions", i.e., emotions that arise when the same situation is appraised at the same time as an action and a event. Differently from attribution emotions, that are directed towards actions, well–being emotions are directed towards events, i.e., the intentionality of the appraised process is not relevant. Following an established line in emotion modeling [35], we define the desirability of an event with respect to its consequences for the agent's goals. If an event brings about a state of affairs in which a goal of the appraising agent is satisfied, the event is desirable, undesirable if its effects are in conflict with the satisfaction of a goal of appraising agent.

Agents' values not only determine the moral appraisal of self and others' actions, but also affect empathic emotions. In our model, empathic emotions compete with attribution emotions: an agent feels "happy-for" (or "sorry-for") another agent only if the other agent's goal does not put her/his values at stake.

In the following, we briefly describe how the model we propose can be encoded on the top of the Drammar ontology of story and character, with the goal of testing the model on paradigmatic situations in well known stories.

3.2 Drammar ontology

The purpose of the Drammar ontology is to encode the behavior of characters appearing in stories in a semantic format, so as to support the reuse of this knowledge in agent-based applications [7]. The ontology borrows the definition of character from the BDI agent model, with integrations aimed at representing emotional states and moral values. The advantage of using the BDI model is twofold: on the one side, it sets the conditions for the interoperability with most agent-based systems; on the other side, it provides a good metaphor for the mechanism of identification with characters postulated by contemporary aesthetics [13]. Characters, the primary medium for the audience identification [6, 17], are expected to be rational agents by the audience [37] and must manifest an intentional behavior to acquire believability.

Aimed at the annotation of stories, the Drammar ontology assumes that a story can be segmented in a sequence of units: a unit is enacted by certain characters, who perform actions in it, and/or contains certain naturally occurring events. As a result of these actions and events (collectively named incidents), the unit brings the world state from an initial state to a final state. In a situation calculus perspective [26], a unit can be seen as an operator characterized by preconditions and effects, that bridges the story world from a state in which the preconditions hold to one in which the effects hold.

The top level of the Drammar ontology consists of four disjoint classes: `Unit`, `Dynamics`, `Entity` and `Relation`. A story is segmented into units (`Unit` class); units feature entities (`Entity` class), i.e., agents (i.e., characters) and objects, involved in actions and events (the unit incidents). The `Dynamics` class models the advancement of drama as a sequence of states interconnected by incidents. Finally, the `Relation` class describes the properties of drama entities in a certain unit, such as the agents' goals and the conflicts among them.

A Unit is `enactedBy` some Agents and contains (`contains Event`) some incidents (`UnitIncident`). The `UnitIncident` class (inspired by the Time Indexed Situation and the Time Indexed Participation patterns defined in [14, 15]) connects the occurrence of an event – namely, an agent's action or a naturally occurring event – with the entities (agents and objects) which participate in it (`incidentFeatures`) and the process (action or event) which constitutes the incident (`featuresProcess`), setting the incident into the time extent provided by a unit. Similarly to the `UnitIncident` class, the `StoryState` class connects the occurrence of a state (`State` class, i.e., a state of affairs or a mental state) with the entities (agents and objects, `featuresAgentInState` and `featuresObject` respectively) which participate to the state, and sets this event in relation to a unit (`hasPrecondition` and `hasEffect`). The linguistic description of the incident, then, is attached to the `ProcessSchema` (or `StateSchema`) class (not represented in the figure), which in turn is connected to the entities which play a role in the incident through the `Role` class.

Agents' motivations and emotional states are modeled by the `MentalState` class, further subdivided in `Belief`, `Goal`, `Emotion` and `Value`. Since all these properties are dynamic (i.e., unit–dependent), they are not directly connected with the `Agent` class. For example, an agent may form a goal and maintain it along several units, but the goal may be active only in a subset of these units. Indeed, the connection between `Agent` and its properties (including mental states, like goals) is mediated by the `AgentInUnit` and `AgentInState` classes. Both classes are subsumed by the `Relation` top–level class.

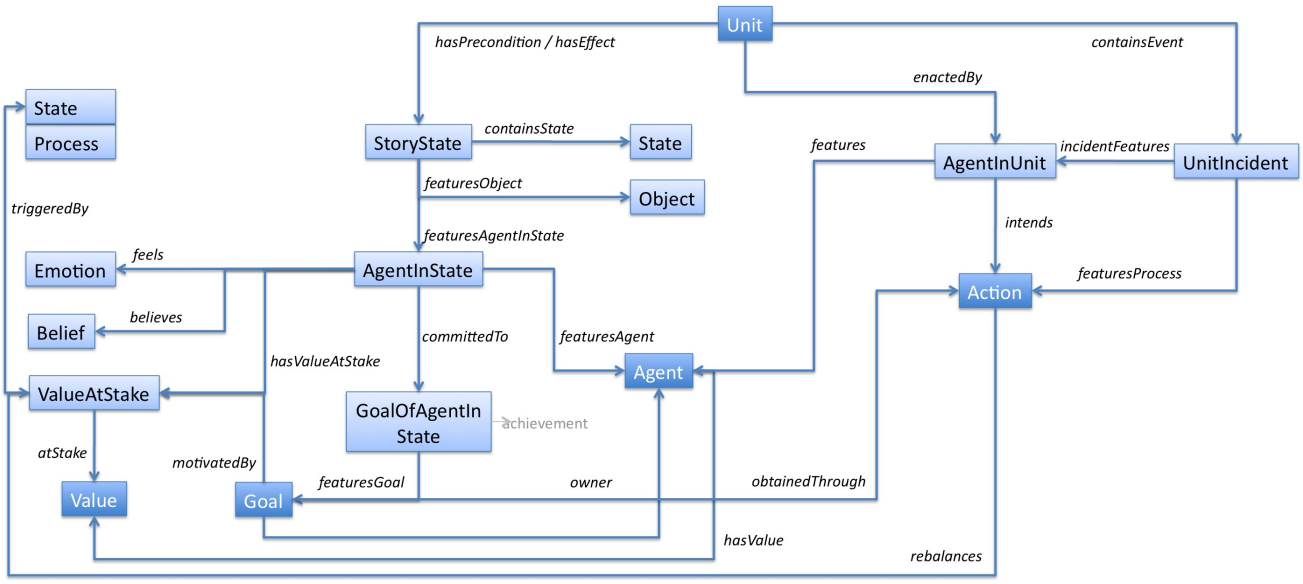


Figure 2: The representation of the preconditions (and effects) and incidents of a story unit in Drammar.

The `AgentInUnit` class represents the participation of an agent to a certain unit, where it displays specific features (like specific mental states, qualities, and so on) that are unit-specific and cannot be attached to the definition of the agent at the level of the entire story (the "character bible" in scriptwriting terms [39]). Notice that the model does not model the distinction between desires and goals, assuming that desires are high–level goals attached to the agent's definition, as part of the character's bible.

The representation of values and their relations with an agent's goals assumes the schema described in [8]. Values are attached to the `Agent` class through the `hasValue` property. An agent's value can be put at stake by the occurrence of a certain process or state (`triggeredBy` property). In response to a value at stake, an agent formulates a value-related goal (see the `motivatedBy` property that connects a `Goal` with a `Value`); a value is not at stake anymore when re–established by an action or event (`rebalances` property).

4. RULES FOR MORAL APPRAISAL

As described in the previous section, the Drammar ontology describes story units in terms of a triple composed of the story world preceding the unit, the unit incidents, and the story world following the units, treating units as operators in which the actions and events (i.e., the story incidents) bring the state of the story world from a certain configuration to another. So, we leverage this structure to model how the emotional state of an agent changes as an effect of the occurrence of the story incidents.

For praiseworthiness, the preconditions of the unit are relevant, since they contain the representation of the appraising agent's values at stake and of the motivations of the agent of the appraised action. For blameworthiness, the effects of the unit are relevant, since they represent the values of the appraising agent that became at stake after the unit incidents (as long as they consist of actions). In both cases, the agent's emotions are established in the effects of the unit as a consequence of the appraisal process. In the following, we describe the SWRL rules [20] for the activation of emotions in agents. In story annotation, using SWRL rules allows comparing manually assigned emotions with the predictions of the model, with the twofold advantage of validating the model onto real stories and supporting the work of human annotators.

According to OCC, the **Pride** emotion type belongs to the "attribution emotions" and is generated by the appraisal of an action of the agent her/himself: for this emotional state to arise, the appraised action must be considered praiseworthy. In our model, praiseworthiness corresponds to bringing to balance a value at stake. The activation of this emotion depends upon:

1. an agent's *value at stake* – in the preconditions of the unit (*balance* property set to false);
2. an agent's *goal* to bring the value back to balance – in the preconditions of the unit;
3. the appraised *action*, executed by the agent in the unit as a consequence of her/his commitment to the value–dependent goal;

Notice that this emotion only depends on the agent's intention to bring the value at stake back to balance, but it doesn't require the intended action to succeed and/or to re–establish the value at stake[3].

When the attribution is directed towards the actions of another agent, the emotional state of **Admiration** is generated in the appraising agent:

[3]In our view, the focus is on the intention of trying. We do not consider other factors related to the context, such as the ability to do our job or our responsability, that may affect an emotion of pride.

1. the appraising agent's *value at stake* – in the preconditions of the unit (*balanced* property set to false);
2. the appraising agent's *goal* to bring the value at stake back to balance – in the preconditions of the unit;
3. the appraising agent's *belief* that the agent of the appraised action has the goal to re-establish the value at stake – in the preconditions of the unit;
4. the appraised *action*, executed by the other agent in a unit incident as a consequence of her/his commitment to the value–dependent goal.

Again, this emotion only depends on the agent's belief that the agent of the appraised action has the intention to bring the value at stake back to balance, and not on its actual achievement. Notice that, in order for the emotion to be triggered, the appraising agent must share with the agent of the appraised action not only the value at stake but also the goal to remove it. Since this assumption may be difficultly implemented in practical scenarios, a weaker form of this rule can be hypothesized, where the Admiration emotion is triggered by the mere observation of the effect of another's agent action on the state of a value as stake. According to this weaker rule, Admiration depends upon conditions 1 and 2 of the previous definition, plus:

1. the appraised *action*, that brings the value at stake back to its balance (balanced property set to false);

In our model, blameworthiness corresponds to putting a value at stake. When an action is appraised as blameworthy and the focus is on the agency of self, according to OCC model, the emotions of self-reproach can arise. The activation of **Self-reproach** emotion depends on:

1. an agent's *value* – not at stake in the preconditions of the unit (balanced property set to true);
2. the appraised *action*, executed by the agent in the unit, putting the value at stake – in the effects of the unit;
3. the agent's *goal* to re-establish the value at stake;

When the focus is on the agency of others, the emotion generated is Reproach. The activation of **Reproach** emotion depends on:

1. an agent's value not at stake – in the preconditions of the unit (balanced property set to true);
2. the appraised *action*, executed by another agent in the unit, putting the value at stake – in the effects of the unit
3. an agent's *goal* to bring the value back to balance – in the effects of the unit

When the same incident is appraised at the same time along the praiseworthiness dimension and the desirability dimension, compound emotions are generated. For example, the combination of Joy and Pride gives the Gratification emotion, i.e., the agent's is at the same time proud of having executed a praiseworthy action and joyful for the desirability of the effects of this action with respect to some other goal.

The rule for the **Joy** emotion depends on the following elements:

1. an agent's (unachieved) *goal* is the precondition of the unit;
2. the *achievement of the goal* in the unit's effects;
3. a *process* (no matter if it is an action or an event) occurred in a unit's incident, which has determined the goal achievement.

Notice that, since we assume that each time an agent's value is put at stake, she/he forms a goal to remove it from being at stake, the joy emotion does not apply to value–dependent goals (unless the same event happens to achieve a standard goal and a value–dependent goal at the same time).

5. EXAMPLE

To test the generation of emotions according to the SWRL rules we rely on the well-known Thirty-Six Dramatic Situations described by the French writer and dramatist Georges Polti [34]. Polti's list of situations is the most famous example of drama classification and has risen the interest of film industry since its publication [32]. Polti takes as domain a corpus of 1012 important plays – from the Classical Greek tragedies to dramas of his contemporary authors (Ibsen as well as Conan Doyle).

Our example is taken from the legend of Don Juan (classified in the Fifth Situation by Polti) which has inspired many authors including Molière, Da Ponte and Byron. Don Juan is a libertine who takes great pleasure in seducing women. One day, in a graveyard, he encounters the Commandant Don Gonzalo, the father's ghost of one of the girls he seduced. Don Juan invites the ghost (a statue) to dine at his house, with the goal of deriding him.

The Commandant accepts and in turn invites Don Juan to dine with him at the graveyard. The Commandant, whose moral value is "sobriety", wants to kill Don Juan to avenge his daughter. So, when Don Juan goes to the graveyard, the Commandant asks him to shake his hand. When Don Juan extends his arm, the statue grabs hold of his hand and drags him away to Hell, thus executing the punishment he deserves for his libertine life.

For our example, we model the climactic story segment in which the Commandant kills Don Juan. In Drammar (see Figure 3), we model this segment of the story as a Unit (named `Punishment`); this unit is enacted by two agents, the `Commandant` and `DonJuan`. The participation of these agents in the unit is bridged by two instances of the `AgentInUnit` class, `Commandant_in_Unit` and `DonJuan_in_Unit` (the latter is not in the figure for space reasons).

Before the unit occurs, in the state which constitutes the precondition of the unit (`hasPrecondition`), the Commandant (through the `AgentInState` class) is committed to the goal of killing Don Juan (see the `committedTo` property linking the `AgentInState` instance (`Commandant_in_Antecedent`) with the `avenging_daughter` instance of `Goal`, so he `intends` to execute the `killing` action. In the unit precondition, the "sobriety" value of the `Commandant` is at stake (due to the libertinage of Don Juan): in order to restore this value, the `Commandant` has formed the goal of `punishing_DonJuan`. Both goals, `punishing_DonJuan` and `avenging_daughter`, are obtained through the killing action in the unit. In the unit effects, Don Juan is dead and the value of "sobriety" is not at stake anymore.

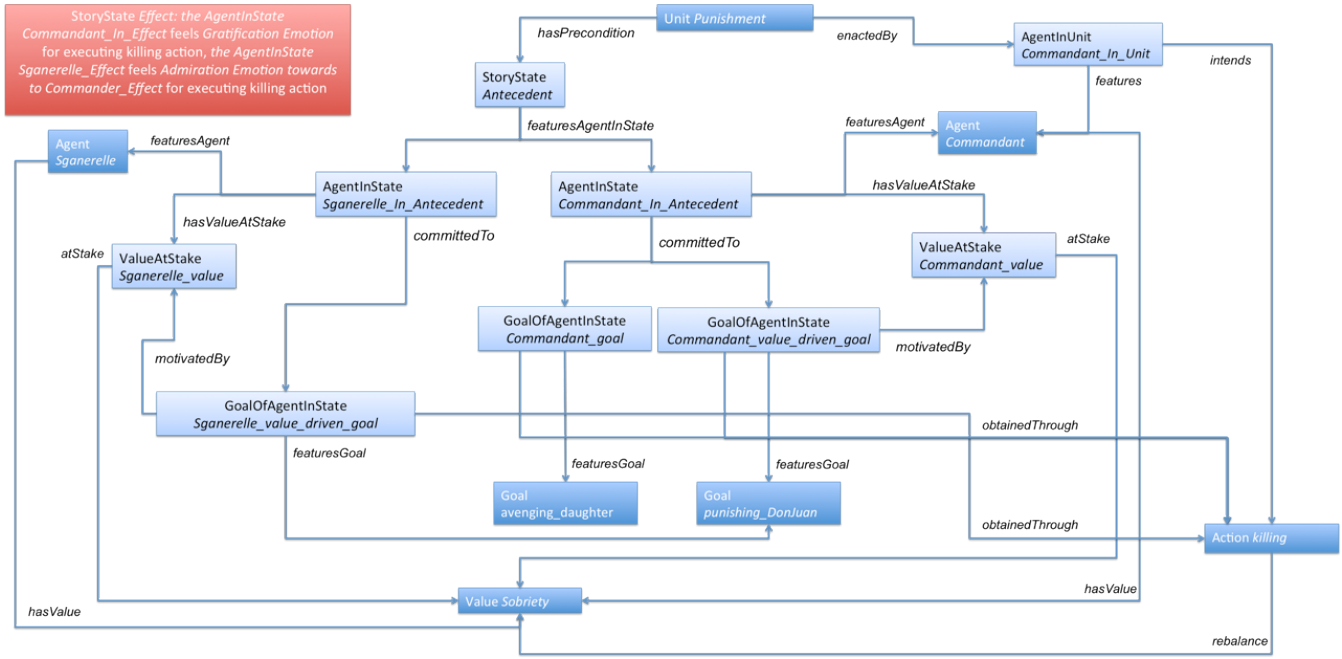


Figure 3: The punishment scene from the legend of Don Juan, encoded in Drammar.

By applying the SWRL rules defined on the Drammar ontology to this representation (Section 4), the rule for *Gratification* fires. Since *Gratification* is a compound emotion, it encompasses both the conditions for *Pride* and for *Joy*. The *Pride* rule fires because the Commandant's own action (killing Don Juan) is appraised as praiseworthy, since it is motivated (see unit preconditions) by the value–related goal of restoring a Commandant's value at stake. The *Joy* rule is applied because the Commandant has also achieved, through the same action, his goal of avenging his daughter.

In this example, we also consider also the emotional state of Don Juan's servant, Sganerelle. According to the legend of Don Juan, Sganerelle shares the value of sobriety with the Commandant. So, according to SWRL rules, Sganerelle *admires* the action performed by another agent, the Commander because he appraises the action of killing Don Juan as praiseworthy, according to his values. For Sganerelle, the appraisal is based on his belief that Don Juan wants to re-establish a value at stake he shares.

6. CONCLUSIONS

In this paper, we described a model of the appraisal of the moral dimension of emotions. Relying on the notion of value, we proposed a general model of value–based appraisal of actions, an intrinsically interpersonal dimension in emotion generation. Through the model we propose, a range of emotional states can be elicited, depending on the agent's subjective and shared values, and the generation of empathic emotions can be traded-off against the moral dimension of appraisal.

We implemented our model as a set of SWRL rules on the top of an ontology, previously developed for the annotation of story and characters, and tested it on a narrative situation. This methodology is based on the assumption that stories provide a valid testbed for emotional agents, especially when the moral dimension is concerned.

The model we propose can be applied to the annotation of stories, both for deriving the emotional states of the characters from the story description, and for validating the characters' emotions hand-coded by human annotators. Annotated stories, then, provide inspiration to the design and evaluation of believable artificial characters.

However, the model has implications also for the design of agents in general, since it contributes to establishing a connection between values and emotions. Values, in fact, are relevant not only for the realm of interactive narrative and drama, but also for multi–agent applications – where they can be employed to model the shared values of a society and the interpersonal conflicts among individuals – and for interactive systems. In particular, in interactive systems, they can be employed to make virtual agents reverberate with the user's values, opening the way to a more comprehensive model of empathy.

7. REFERENCES

[1] J. Bates, A.B. Loyall, and W.S. Reilly. An architecture for action, emotion, and social behaviour. *Artificial Social Systems, LNAI 830*, 1994.

[2] Christian Becker-Asano. *WASABI: Affect Simulation for Agents with Believable Interactivity*. PhD thesis, Faculty of Technology, University of Bielefeld, 2008. IOS Press (DISKI 319).

[3] Michael Bratman. *Intentions, Plans, and Practical Reason*. Harvard University Press, 1987.

[4] J. Bruner. The narrative construction of reality. *Critical Inquiry*, 18(1):1–21, 1991.

[5] Joseph Campbell. *The Hero with a Thousand Faces*. Princeton University Press, Princeton, 1949.

[6] Noel Carroll. *Beyond Esthetics: Philosophical Essays*. Cambridge University Press, New York: Cambridge, 2001.

[7] M. Cataldi, R. Damiano, V. Lombardo, A. Pizzo, and D. Sergi. Integrating commonsense knowledge into the semantic annotation of narrative media objects. *AI* IA 2011: Artificial Intelligence Around Man and Beyond*, pages 312–323, 2011.

[8] R. Damiano and V. Lombardo. An Architecture for Directing Value-Driven Artificial Characters. *Agents for Games and Simulations II: Trends in Techniques, Concepts and Design*, pages 76–90, 2010.

[9] Denis Diderot. *Paradoxe sur le comédien*. Sautelet, 1830.

[10] L. Egri. *The Art of Dramatic Writing*. Simon and Schuster, New York, 1946.

[11] Paul Ekman. *Basic emotions*, chapter 3, pages 45–60. John Wiley & Sons Ltd, New York, 1999.

[12] Clark D. Elliott. *The affective reasoner: a process model of emotions in a multi-agent system*. PhD thesis, Northwestern University, Evanston, IL, USA, 1992.

[13] Susan L. Feagin. On Noel Carrol on narrative closure. *Philosophical Studies*, (135):17–25, 2007.

[14] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with dolce. In *Proc. EKAW 2002*, Siguenza (SP), 2002.

[15] A. Gangemi and V. Presutti. Ontology design patterns. *Handbook on Ontologies*, pages 221–243, 2009.

[16] Patrick Gebhard. Alma: a layered model of affect. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, AAMAS '05, pages 29–36, New York, NY, USA, 2005. ACM.

[17] Alessandro Giovannelli. *In Sympathy with Narrative Character*, volume I of *Journal of Aesthetics and Art Criticism*, pages 83–95. Wiley Blackwell, 2009.

[18] J. Gratch and S. Marsella. A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4):269–306, 2004.

[19] J. Gratch and S. Marsella. The architectural role of emotion in cognitive systems. *Integrated models of cognition systems*, page 230, 2007.

[20] I. Horrocks, P.F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean, et al. Swrl: A semantic web rule language combining owl and ruleml. *W3C Member submission*, 21:79, 2004.

[21] Oliver P. John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A. Pervin and Oliver P. John, editors, *Handbook of Personality: Theory and Research*, pages 102–138. Guilford Press, New York, second edition, 1999.

[22] Richard S. Lazarus. *Emotion and Adaptation*. Oxford University Press, USA, August 1991.

[23] R.S. Lazarus. *Emotion and adaptation*. Oxford University Press, USA, 1991.

[24] S. Marsella and J. Gratch. EMA: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90, March 2009.

[25] J.C. Martin, R. Niewiadomski, L. Devillers, S. Buisine, and C. Pelachaud. Multimodal complex emotions: Gesture expressivity and blended facial expressions. *International Journal of Humanoid Robotics*, 3(3):269–292, 2006.

[26] A. McCarthy. Mental situation calculus. *TARK: Theoretical Aspects of Reasoning about Knowledge*, 1986.

[27] R. McKee. *Story*. Harper Collins, New York, 1997.

[28] Marvin Minsky. *The society of mind*. Simon & Schuster, Inc., New York, NY, USA, 1986.

[29] K. Oatley and P. N. Johnson-Laird. Towards a cognitive theory of emotions. *Cognition & Emotion*, 1(1):29–50, 1987.

[30] A. Ortony, G. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambrigde University Press, 1988.

[31] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, July 1988.

[32] Frederick Palmer. *Photoplay Plot Encyclopedia*. Palmer Photoplay Corporation, Los Angeles, California, 1920.

[33] F. Peinado, M. Cavazza, and D. Pizzi. Revisiting Character-based Affective Storytelling under a Narrative BDI Framework. In *Proc. of ICIDIS08*, Erfurt, Germany, 2008.

[34] G. Polti. *Les trente-six situations dramatiques*. Mercure de France, Paris, 1895.

[35] W. Scott Reilly and Joseph Bates. Building emotional agents, 1992.

[36] James A. Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145–172, January 2003.

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

[38] K. R. Scherer. The role of culture in emotion-antecedent appraisal. *Journal of Personality and Social Psychology*, 73:902–922, 1997.

[39] L. Seger. *Creating Unforgettable Characters*. Henry Holt and Company, New York, 1990.

[40] M. Si, S.C. Marsella, and D.V. Pynadath. Modeling appraisal in theory of mind reasoning. *Autonomous Agents and Multi-Agent Systems*, 20(1):14–31, 2010.

[41] Mei Si, Stacy C. Marsella, and David V. Pynadath. THESPIAN: An architecture for interactive pedagogical drama. In *Proceeding of the 2005 conference on Artificial Intelligence in Education*, pages 595–602, Amsterdam, The Netherlands, The Netherlands, 2005. IOS Press.

[42] Greg M. Smith. *Film Structure and the Emotion System*. Cambridge University Press, Cambridge, 2003.

[43] K. T. Strongman. *The psychology of emotion*. J. Wiley London, New York„ 1973.

[44] B. van Fraassen. Values and the heart's command. *Journal of Philosophy*, 70(1):5–19, 1973.

[45] R. B. Zajonc. Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35(2):151–175, February 1980.

When agents meet: empathy, moral circle, ritual, and culture

Nick Degens,
Gert Jan Hofstede,
John Mc Breen,
Adrie Beulens
Wageningen University
6706KN, Wageningen,
the Netherlands
{nick.degens, gertjan.hofstede, john.mcbreen, adrie.beulens}@wur.nl

Samuel Mascarenhas,
Nuno Ferreira,
Ana Paiva
Instituto Superior Técnico,
Technical University of Lisbon
INESC-ID, TagusPark
2780-990 Porto Salvo,
Portugal
{samuel.mascarenhas, nuno.ferreira}@gaips.inesc-id.pt, ana.paiva@inesc-id.pt

Frank Dignum
Utrecht University
3508TB, Utrecht,
the Netherlands
f.p.m.dignum@uu.nl

ABSTRACT

Creating agents that are capable of emulating the same kind of socio-cultural dynamics found in human interaction remains one of the hardest challenges of artificial intelligence. This problem becomes particularly important when considering embodied agents that are meant to interact with humans in a believable and empathic manner.

We propose a list of basic requirements for these agents to be capable of such behaviour and we introduce a model of the social world intended for implementation in affective agent architectures. In our framework culture alters agents' social relationships rather than directly determining actions, allowing for a deeper representation of empathy.

Categories and Subject Descriptors

H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces—*Theory and models*; J.4 [Social and Behavioural Sciences]: Sociology

General Terms

Design, Human Factors

Keywords

Group dynamics, culture, virtual environment, virtual agents, modelling social interaction

Appears in: *Proceedings of the Workshop on Emotional and Empathic Agents, in the 11th International Conference On Autonomous Agents and Multiagent Systems (AAMAS2012),* June, 4-8, 2012, Valencia, Spain.

1. INTRODUCTION

> *Horatio and his girlfriend Nadia are two agents sitting in a bar. They've been together as a couple for a while now. When they order a drink, a lady bartender agent walks by and Horatio starts to talk with her. After a few minutes, Nadia stands up, walks away, and shouts over her shoulder: "It's always the same with you!"*

Based on the information above, humans would have almost no difficulty trying to describe what Nadia must have been feeling. This is because we are able to make assumptions about the social relationship between the boy and the girl.

However, for an agent to be able to make the same assumptions, it needs to have clearly operationalized parameters of the social world. What is the relationship between the boy and the girl? Why does the boy talk to the other girl for a few minutes? Why does the girl stand up and walk away? These are instances of what we call socio-cultural dynamics: given any social situation, depending on the participants' personalia and cultures, how does the situation unfold?

Besides being able to make assumptions about the social world, there is also the issue of making social judgements; what is right and what is wrong. Changing a few simple elements of this scenario could change our perception of right and wrong, and this is something that an empathic agent should be able to do as well.

These judgements would become even more complicated when you take culture into account. What if talking to the other girl was an acceptable thing to do where you are from? What if it didn't mean that you might be romantically interested in them? It adds an extra level of complexity to the already quite challenging level of social behaviour. As basis of the article we take the stance: All people are moral, but culture modifies that morality.

In this paper we aim to identity and take the first steps to create a conceptual model for social behaviour in virtual agents. There are no theoretical bounds to the level of social complexity that we want to represent in our model. However, the model should be as simple as possible, but still rich enough to allow for short emergent interactions between agents with different cultural configurations. Through these simple interactions, people will be able to see the effect of culture on behaviour.

To establish the minimal modelling requirements, we will use a story of two agents meeting each other on the street. They don't know each other and one of them needs a favour from the other. This short, and simple, setup allows us to identify important requirements for empathic agents. Since this paper only focuses on the creation of a conceptual model for social behaviour, many questions related to the implementation of these requirements will be left unanswered.

The paper is organized in the following manner. We will start by describing related work on cultural agents. The following section will focus on the notion of *rituals*, a construct through which behaviour gains social meaning for a group of agents that have shared attention. After that we focus on different interpretations of these actions by having different *moral circles* active in the mind of an agent based on the ritual. In the last part of the paper we will look at how culture can modify these rituals and moral circles to create culturally-varying behaviour in agents.

2. RELATED WORK

The increasing need for embodied agents to interact in a social and empathic manner has lead researchers to address different aspects of social interaction. Particularly related to the work presented in this paper is the Synthetic Group Dynamics (SGD) model, proposed by Prada and Paiva [1], as it aims to create believable interactions in social groups formed by autonomous agents. In order to achieve this, agents build social relations of power and interpersonal attraction with each other. They also have the notion of belonging to a group in which they are regarded as more or less important, according to their status and/or level of expertise.

Similar to the SGD model, our proposed model also places a strong emphasis on embedding group dynamics and social relationships in the agent's mind. However, differently from SGD, we also address the relationship between culture and the dynamics of groups.

When designing social agents, culture has often been overlooked despite its huge influence on human behaviour. Without taking culture into account, we argue that the social richness of agent-based simulations becomes significantly limited. For instance, it becomes difficult for agents to empathise with users from different cultures, if they lack the ability to interpret actions from different cultural perspectives. Moreover, modelling culture has been an essential endeavour when considering agent-based applications for intercultural training such as ORIENT [2], ELECT BiLAT [3], or TLTS [4].

Research on cultural agents is steadily rising. So far, several systems have focused on the adaptation of directly observable features of conversational behaviour to specific cultures. For instance, the work of Jan et al. [5] addresses differences in proxemics, gaze and speech overlap between the North American, Mexican and Arabic cultures. Similarly, the work of Endrass et al. [6] addresses the integration of non-verbal behaviour and communication management aspects, considering differences between the German and Japanese cultures.

While the aforementioned models focus on modelling the effects of culture on communication aspects, the research presented in this paper addresses another important facet of culture. Namely, how it influences decision making and behaviour selection.

In the model proposed in Mascarenhas et al [7], two of Hofstede's dimensions of culture, individualism and power distance, are directly used to influence the agent's decision making and appraisal processes. However, this is done only at the individual level without considering important elements from the social context such as group membership and relational variables.

Another agent model where culture affects decision making is the model proposed by Solomon et al. [8] which concerns the definition of specific cultural norms. The model allows defining links between specific actions (e.g. show-picture-of-wife) and one or more cultural norms (e.g. respectful-of-modesty). An association link can either be positive in the case where the action promotes the norm or negative in the opposite case. One drawback of this model is that it requires a great deal of manual configuration as it tries to associate culture directly to individual actions.

One step towards generating culturally appropriate behaviour within an agent model was taken by Mc Breen et al. [9] who propose the concept of *meta-norms* to operationalize culture. These use the Hofstede Dimensions of Culture to explain how you can create a set of generic rules that give agents a propensity to behave in a certain way in certain relational contexts.

In our proposed model, we argue that actions are often selected not because of their instrumental effects but because they are an important symbolic step of an on-going ritual, thus making rituals an essential part of social interaction.

The idea that rituals are important to model cultural differences in embodied agents was also explored in Mascarenhas et al [10], where a computational model of rituals was implemented and integrated into an affective agent architecture, developed by Dias and Paiva [11]. One limitation of their proposed model is that it assumes that agents have a shared knowledge of rituals. This assumption is not true when

considering scenarios where agents from different cultures may meet as exemplified in this paper.

3. MODELLING CULTURAL AGENTS

3.1 The Structure of a Ritual

> *Horatio is in a city he doesn't know, and is trying to find his hotel. After walking around for a while, he is unsure in which direction to continue and decides that it would be best to ask somebody on the street for more information. At that moment, Claudius, who is on his way to work, is walking in the opposite direction of Horatio. Horatio decides to draw the attention of Claudius…*

Some actions may be purely instrumental, e.g. picking up an object that has fallen on the floor. However, in a social world, such actions usually have a symbolic effect as well. For instance, what objects would you pick off the floor, in which places, and with which people present? To create an empathic agent, they need to be able to understand the social effect of these actions.

These symbolic elements of actions have some effect on the relationship between yourself and others. However, such an action will only take effect if the other is paying attention; if not, the social meaning of the action might be lost on him.

The first requirement for our model of social behaviour is:

- Groups of agents should be able to have a degree of shared attention and purpose within a certain environment.

This requirement closely matches the definition of a ritual, found in Rothenbuhler [12]. He states that rituals range from the ceremonial and memorable to the mundane and transient. In fact, any group of people (in our sense of the word, as a collection of people gathered in one place) that has a degree of shared attention, can be said to be engaged in a ritual.

Rituals help mediate changes in social order and are thus an essential element of social behaviour. As Hofstede et al. [13] say in their work, rituals are: "Collective activities that are technically superfluous to reach desired ends but that, within a culture, are considered socially essential."

> *…In Horatio's mind there is a certain structure to asking a favour of a stranger. First you would politely greet him, and after exchanging pleasantries you would then proceed to ask him for help. Doing so would make the stranger feel obliged to help you…*

In a further operationalization of the ritual, Hofstede [14] explains that a ritual consists of three elements: a beginning, a body, and an end.

The *beginning* is characterized by an initiating move and a response. This initiating response carries the social meaning of the ritual. The response can be classified as running along two dimensions: direction (going along or opposing) and strength of the response (ranging from low to high). Depending on the response, a ritual is either initiated or aborted; if the purpose of the ritual is clear to both parties and agreed upon, they proceed to the *body* of the ritual.

Within the *body* of the ritual, the actual social change is put into actions. Depending on the type of change, the participants of the ritual must act in an appropriate manner.

The last stage of the ritual would be the *end*, in which the social change is reinforced in an appropriate manner and the ritual is brought to its conclusion.

3.2 Different Interpretations

> *On his way to work, Claudius sees a stranger walk up to him with an uncertain look on his face. This kind of behaviour is typical of people who need directions and have need of somebody to help them on their way…*

Not all behaviour will be interpreted in the same way. This issue might be particularly true for people from different cultures, but even within the same culture there is no guarantee that you 'speak' the same language.

In the example above, Claudius recognizes that when Horatio walks up to him in a certain way, it means that he needs a favour. Now if someone would do that at night in a shady part of town, it might mean that they want to steal your valuables.

Different interpretations don't just depend on the environment that you're in, but also on the people that you interact with. In our example, Claudius and Horatio don't know each other. But what if they had been old friends? Would Horatio still have walked up to Claudius in the same manner and, if so, would it have meant the same thing?

The second requirement for our model is:

- The same action needs to have different interpretations for different people in different environments.

Within our models we choose to have rituals as events that have an impact on the social world. In our model we represent this social world through the use of *moral circles*, which can be created or changed by rituals. Moral circles are a pragmatic concept that we can use to define relational variables and social order in groups of people.

A first, informal definition is as follows. A moral circle is comprised of three elements: the people to whom it applies, their mutual perceptions of social attributes, and the social norms that regulate their behaviour.

Why use the concept of a moral circle? To begin with, it is generic. Hofstede et al. [13] use it as a general indication of a human unit of social agency, ranging from a few people to all of humanity, taking inspiration from evolutionary biologist David Sloan Wilson, who describes humans as a 'eusocial'

species, i.e. one in which the group has supplanted the individual as the main level of evolution.

Now, while in most eusocial species it is rather simple to determine the unit of evolution – it would be the colony of bees, for instance – this is not so in humans. Yet the assumption is that we have a biological propensity, including moral sentiments, to act as group members. In other words, acting for the survival and prosperity of our Moral Circles is in our nature. It is this propensity that is the main justification for our concept of moral circle – which we shall often abbreviate 'MC' from this point onwards.

> *... Claudius wonders if he has time to help this stranger. In an hour he has an important deadline at work and he still has some things left to prepare. So he is left with a choice: he can either stop for a few seconds and talk to the stranger or he can ignore the stranger and carry on to work...*

Each context shapes its own MC typology, which depends on who is involved and what MCs they perceive to be relevant to the situation. A person can belong to many different MCs at the same time. While these MCs will affect the actions of any one person at any time, one MC is usually more salient than others. For instance, in most cultures, leaving work duties to marry or bury a family member would be allowable, or even endorsed. The priority between events is itself symbolic of a prioritisation among MCs.

MCs come in different types. They can range from the default MC of "all people who count as people", to which strangers may or may not belong, to long-lasting organised groups, such as families or ethnic communities or companies, to the relatively informal, such as groups of acquaintances.

A more formal MC has both more specific social norms (rules of appropriate behaviour) and a strong inertia in membership; whether you're in or out is usually being determined by clear attributes e.g. employment or club membership. Membership changes in more formal MCs are usually mediated by formal rituals, often denoting a change in status.

More informal MCs can be, for example, groups of specific friends (some you might know from your studies, others from your sports club). These more informal MCs still develop guides to appropriate behaviour. Membership of such an informal MC is often not as clearly defined as in more formal MCs. The relevant social norms for an informal MC will not be stated in any text and can evolve more freely through an emergent consensual process, than is usual in formal MCs.

A particularly difficult social issue is how to behave when more than one MC could be relevant. Culture can help determine the relative salience of these MCs.

This leads to the third requirement for our model:

- There needs to be some mechanism that helps determine the salience of Moral Circles based on the ritual that the agent is participating in.

3.3 Who They Are to You and What Effect That Has

> *...Horatio walks up to Claudius and recognizes that he's dealing with an older man who is wearing a very formal suit. The old man is looking at his watch and Horatio realizes that the older man is probably in a hurry...*

There are different relational primitives that can be present between members within a MC. Imagine that the stranger on the street is older than you are? How would that influence your behaviour? What if they were younger, would you treat them differently? Normally speaking we talk about *hierarchical* status in the sense of formal roles, such as a boss in the work environment. But it could even be an elderly gentleman, who might have higher status due to his age.

Status helps to establish dominance, which is used to establish the pecking order within a group. Many difficulties between individuals arise because there are differences in perceived status (You're not in charge, I am!). To avoid such conflicts, formal MCs usually have formal roles with explicit rights and obligations, which can range from that of the managing director of a multinational company to the most junior trainee.

In the example above, Horatio is able to make an assumption about the status of Claudius because of two factors: his age and the suit he is wearing. Note that Horatio might be wrong in his appreciation of these attributes; these symbols might mean something different to Claudius than they do to Horatio.

The fourth requirement:

- Agents must be able to infer the status of characters, either through public variables, or through observation and interpretation of symbols.

3.4 The Agent's Social World

At this point it becomes necessary to specify in some more detail the social world in which our agents live (see figure 1 on the next page).

In our simulations, some variables are taken for granted and will not change throughout a session in which a group of agents interact. This includes the uppermost level in the figure, the components of which will be described in more detail below. Other components that may or may not change can be found in the middle level. The bottom level shows the elements that make up the visible part of the agent interaction.

An important aspect of the figure is the realization that when there is no data available from the middle level, an agent will fall back on their top level attributes. This might be the case

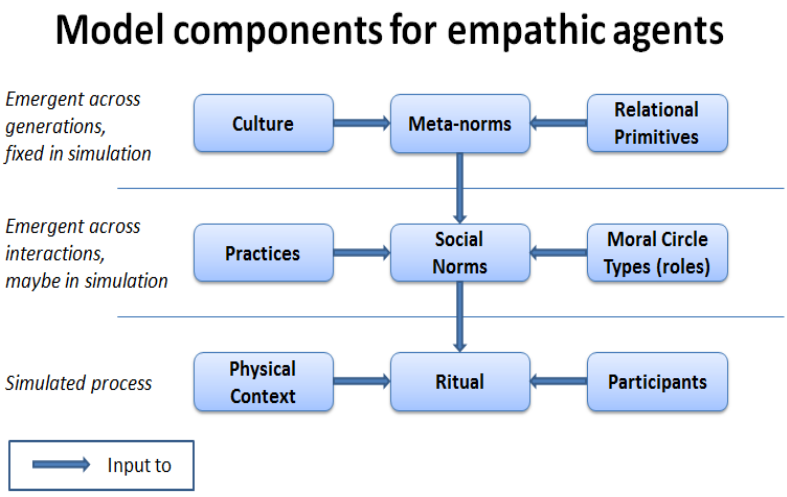


Figure 1. From culture to actions: model components for empathic agents

when an agent is put into a new and 'strange' environment, where they have no specific rules for behaviour. We shall now first look at the middle level.

3.5 What Is Right and What Is Wrong?

3.5.1 Social

... Claudius has no idea where the hotel is that Horatio is looking for. In his eyes, a young man like Horatio should be better prepared in planning his trip. Claudius tells the man that he has no idea where the hotel is, wishes him good luck, says he has to go, and rushes to work. If he had more time, he would have helped Horatio more…

How does one behave within a ritual? To answer this question, we need to look at social norms. These norms can be considered the practices of a group and while they reflect underlying value structures, they are not determined by them. They evolve to be accepted by the larger part of a society, or a segment of that society, as a short-term guide to proper moral behaviour.

Both the interpretation of the moral quality of behaviour and the translation of intentions into actions, are mediated by the current social norms. These social norms are very malleable; a population can come to believe that drink-driving or smoking indoors in the presence of non-smokers are normatively wrong, in a relatively short period of time.

However, the underlying value structure and MC dynamics will not have altered significantly, if at all. The detailed functioning of MCs in practice reflects the underlying cultural values, as culture moulds the social norms of a society. Social norms are one of the tools for interpreting the moral quality of the actions of others. They also indicate what behaviours are allowed (and effective) for translating social intentions into actions.

In our example Claudius is judging Horatio for his behaviour: Horatio should have been more prepared. As a result, Claudius believes that it is more important for him to carry on to work, instead of helping this youth, who should have been better prepared.

- For all MCs, rituals and contexts that are simulated, social norms should be present and tied to MCs and rituals.

3.5.2 Cultural

Horatio is left confused: Where he is from, people usually help strangers, even if you are in a hurry. He decides to carry on and continues on his journey…

In their work, Mc Breen et al. [9] propose the concept of meta-norms to operationalize culture. They use the Hofstede Dimensions of Culture to explain how you can create a set of generic rules that help determine agent behaviour.

Meta-norms as defined by Mc Breen et al. model agents' propensity to behave in a certain way in certain relational contexts. In contrast to the shorter-term guides to behaviour, social norms (middle level of figure 1), meta-norms are longer-term guides to social behaviour (upper level in figure 1). They are about the fundamentals of social life and they are shared within any society that has the same culture. They deal with the basic question of how people should behave with respect to each other depending on who they are. They are close to the values of a culture, in the Hofstede sense of 'cultural programming of the mind', shared tendencies to perceive the social world, and act in it, in certain ways.

In our example Horatio has a different way of determining the importance of MCs from Claudius. For Horatio is it unthinkable that you would leave a stranger needing help on the street to go to work. This shows one way how culture would influence the behaviour of agents.

Within our model, culture will influence two elements: the social structure of moral circles and their social norms (SNs is what follows). The culturally modifiable parameters are the weight of MC primitives, the salience of MCs and the salience of SNs (see Table 1). The most salient MC and the most salient SNs can be established using this operationalization of meta-norms, e.g. "duties of work prevails over social duties towards strangers", or "what my boss wants of me is more important than what anybody else wants of me". There should be room to add culture as a weighting and salience mechanisms for MCs and SNs.

Table 1. Parameters that can be modified by culture

Culturally Modifiable Parameters
Weighting of MC primitives
Salience of MCs
Salience of Social Norms

3.6 Reputation

Where Horatio is from, you can always rely on getting help from strangers.

In Horatio's culture there is a salient meta-norm about helping the needy, whatever the context. Living up to meta-norms and social norms play a paramount role in determining reputation. This is a measure of how well a person lives up to their MC derived obligations and their respect for the rights of other MC members. It can be named 'standing', a variable that could be binary or scalar. An agent can be 'in good standing' versus 'in bad standing' with its fellows [15]. Reputation is essential for agents that can recognise each other and act empathically based on previous interactions.

Within our model we want to represent moral behaviour. This means that two important elements need to be present within our model: actions have to be judged as to whether they are moral or not and members of the moral circle need a perceived level of morality (with unknown people these will be primarily based on meta-norms and on perceived attributes). These are the concepts that will be instantiated as Moral Circle Reputation (MCR) within our model.

Each MC has certain rights and obligations conferred on its members, depending on their roles in the MC. So if a member of a MC does something that goes against expectations based on an understanding of these rights and obligations, it has an effect on their perceived MCR. Each member of the MC has a perception of the MCR of other known members and of their own. So you might think less of yourself if you have done something wrong and others might also think less of you. This decrease can, depending on the level of MCR change, be attenuated by an appropriate atonement.

...Horatio is in town to attend an academic conference. The next day he encounters Claudius there as a senior member of the host university. He wonders whether he should speak to Claudius or not, as his first impression was unfavourable, but maybe that's just how people behave here...

To be able to model these kinds of interactions within empathic agents, it is important that agents are able to keep some form of relational bookkeeping. This leads to the following requirement:

- Some memory of previous interactions is necessary to represent believable behaviour in agents. This memory will concern other agents' personal information and MC memberships, including status and reputation. It will be shaped by the agents' social norms and meta-norms.

3.7 The Effect of Culture

...Horatio needs to request something from his hosts. He speaks to Claudius, who remembers him and asks if he found the hotel without too much difficulty. Horatio replies that he was helped by a shopkeeper shortly after approaching Claudius. Claudius then deals with Horatio's request efficiently and in a very friendly manner...

Horatio feels that there is a contrast in the behaviour of Claudius in both situations. He wonders what the underlying reason is for that contrast. Is it due to his status as a guest at the conference?

Every culture, through the different modifications it brings to the content and salience of MCs and social norms, will cause agents to behave differently and to judge the behaviour of others differently as well.

How can we begin to represent these varying behaviours and judgements in agent architectures? We propose to do this using the Hofstede dimensional model of culture [13].

3.8 Operationalizing Culture

We give an example of modifying the behaviour of agents based on their cultural background by linking the weighting of MC primitives to the Hofstede Dimensions of Culture.

3.8.1 Hierarchy: Large Power Distance Versus Small Power Distance

The importance given by agents to *status* depends on the dimension of Power Distance, which deals with how hierarchy is perceived in a culture.

This is the extent to which the less powerful members of a society expect and accept that power and rights are distributed unequally. Large PDI splits up the MC into status levels MCs that are not permeable and depend on position in society. Agents in cultures of large power distance will respond

differently to others depending on how they perceive their MCS relative to their own. Status differences will be effective barriers to communication; particularly to volitional behaviour travelling upwards.

Horatio will feel that the behaviour of Claudius was appropriate if he comes from a Large Power Distance culture. Indeed, if Horatio was from a very Large Power Distance culture he would never have approached Claudius in the first place. The fact that he did so implies that he is from a Small Power Distance culture.

3.8.2 Aggression and Gender: Masculinity Versus Femininity

The importance given to *reputation* depends on the cultural dimension of Masculinity.

This dimension is about assertive dominance and emotional gender roles. It contrasts a strong-handed, competitive orientation in 'masculine' cultures, in which people in general do not assume others to be trustworthy, men are supposed to be tough, and women subservient and tender; versus a consensus-seeking and care-taking orientation for both women and men in 'feminine' cultures. For our MC primitives in masculine cultures, moral circle reputation will be very unequally divided across the MC, with a tendency to blame the weak and admire the strong. MCR will be more evenly distributed in feminine cultures and will not change so radically with poor behaviour.

In our example Horatio would tend to judge Claudius harshly for not helping him, just as Claudius would judge Horatio harshly for being ill-prepared. In a feminine culture both would be more forgiving of the apparent faults of the other and would expect this same forgiveness of others for their own mistakes.

4. CONCLUSION

The series of requirements that we have presented during the interaction between Horatio and Claudius represent elements that are important to consider when designing empathic virtual agents. Taking these requirements as a starting point, we have discussed elements of our model that will help show realistic social behaviour that can be modified by culture.

Through rituals, in which a set of agents have shared attention in a certain environment, agents are able to act appropriately by applying the relevant moral circles and their social norms. This selection mechanism allows for different interpretations in different contexts.

Culture can then be applied in two ways: through meta-norms and culturally modifiable parameters. In the absence of appropriate moral circles, and the social norms that apply to that moral circle, meta-norms provide guidance. These meta-norms will be particularly relevant for intercultural training, as one generally has difficulties recognizing moral circles and its relational primitives in 'foreign' surroundings. Culture also has an effect on behaviour through the weighting of social norms and moral circles. This structure allows us to have culture influence social relationships rather than act directly on behaviour.

We believe that this paper makes some necessary steps to make virtual agents more empathic. In future work we aim to put the concepts presented in this paper into an affective agent architecture to create believable culturally-varying behaviour in agents for educational purposes. The translation of the concepts presented in this paper to operationalized elements of an affective agent architecture will allow us to discover flaws and additional modelling requirements for empathic agents.

5. ACKNOWLEDGEMENTS

This work was partially supported by European Community (EC) and is currently funded by the ECUTE (ICT-5-4.2 257666) and SEMIRA projects. SEMIRA is partially funded by the Portuguese Fundacao para a Ciencia e a Tecnologia (FCT), (ERA-Compl/0002/2009). The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC or the FCT, which are not responsible for any use that might be made of data appearing therein.

6. REFERENCES

[1] Prada. R., and Paiva. A., 2006. Believable groups of synthetic characters. In Proceedings of the 4th International Joint Conference on Autonomous Agents and Multi Agent Systems. (Utrecht, The Netherlands, 2005)

[2] Aylett, R., Paiva, A., Vannini, N., Enz, S., Andre, E., and Hall, L. 2009. But that was in another country: agents and intercultural empathy. In Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (Budapest, Hungary, 2009). IFAMAAS/ACM DL.

[3] Hill, R.W., Belanich, J., Lane, H.C., and Core, M. 2006. Pedagogically structured game-based training: Development of the elect bilat simulation. In Proceedings of the 25th Army Science Conference (Florida, U.S.A., 2006).

[4] Johnson, W.L., Vilhjalmsson, H.H., and Marsella, S. 2005. Serious games for language learning: How much game, how much A.I.? In Chee-Kit Looi, Gordon I. McCalla, Bert Bredeweg, and Joost Breuker, editors, AIED, volume 125 of Frontiers in Artificial Intelligence and Applications, 306–313. IOS Press

[5] Jan, D., Herrera, D., Martinovsky, B., Novick, D., and Traum D. 2007. A computational model of culture-specific conversational behaviour. In Intelligent Virtual Agents (Paris, France, 2007). 45–56.

[6] Endrass B., Rehm, M., Lipi, A., Nakano, Y., and André, E. 2011. Culture-related differences in aspects of behavior for virtual characters across Germany and Japan. In Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (Taipei, Taiwan, 2011). 441-448

[7] Mascarenhas, S., Dias, J., Afonso, N., Enz, S., Paiva, A. 2009. Using rituals to express cultural differences in synthetic characters. In Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (Budapest, Hungary, 2009). IFAMAAS/ACM DL

[8] Solomon, S., van Lent, M., Core, M., Carpenter, P., and Rosenberg, M. 2008. A language for modeling cultural norms, biases and stereotypes for human behavior models. In Proceedings of the 18th International Conference on Behaviour Representation in Modeling and Simulation (Rhode Island, U.S.A., 2009)

[9] Mc Breen, J., Di Tosto, G., Dignum, F., Hofstede, G.J. 2011. Linking norms and culture. In Proceedings of the 2nd International Conference on Culture and Computing (Kyoto, Japan, 2011)

[10] Mascarenhas, S., Dias, J., Prada, R., and Paiva, A. 2010. A dimensional model for cultural behaviour in virtual agents. International Journal of Applied Artificial Intelligence: Special Issue on Virtual Agents, 2010.

[11] Dias, J., Paiva, A. 2005. Feeling and reasoning: a computational model for emotional agents. In Proceedings of the 12th Portuguese Conference on Artificial Intelligence, EPIA, Springer, 127–140.

[12] Rothenbuhler, E. W. 1998. Ritual communication: From everyday conversation to mediated ceremony. Thousand Oaks, CA: Sage

[13] Hofstede, G. H., Hofstede, G.J., Minkov, M. 2010. Cultures and Organizations: Software of the Mind (3rd edition), McGraw-Hill.

[14] Hofstede, G.J. 2011. Modelling rituals for Homo biologicus. In Proceedings of the 7th European Conference on Social Simulation Association (ESSA, Montpelier, France, 2011)

[15] Nowak, M. A. & Sigmund, K. 2005. Evolution of indirect reciprocity. Nature v437, 1291-1298