Strategic Intentions based on an Affective Model and a simple Theory of Mind

Hazaël Jones[1] Nicolas Sabouret[2] Atef Ben Youssef[2]

[1] UMR ITAP, SupAgro, Montpellier
[2] LIMSI-CNRS, UPR 3251, Orsay

Résumé

Cet article présente un modèle informatique pour raisonner sur les émotions de l'interlocuteur, en utilisant un paradigme de théorie de l'esprit (Theory of Mind, ou ToM, en anglais). Le système manipule des représentations sur des croyances à propos des émotions, des préférences et des buts de l'interlocuteur. Notre modèle affectif est conçu dans le contexte de la simulation d'entretien d'embauche mais il n'est pas lié à un ensemble d'affects spécifique. Il s'appuie sur des règles simples pour sélectionner les types de question lors de l'entretien en fonction de la personnalité de l'agent. Nous l'avons implémenté en utilisant une représentation de type OCC des émotions et un modèle dimensionnel PAD pour les humeurs.

Mots Clef

Théorie de l'Esprit, intentions stratégiques, modèles affectifs, entretiens d'embauche.

Abstract

This paper presents a computational model for reasoning about affects of the interlocutor, using a Theory of Mind (ToM) paradigm: the system manipulates representations of beliefs about the interlocutor's affects, preferences and goals. Our affective model is designed for the context of job interview simulation, but it does not depend on a specific set of affects. It relies on simple rules for selecting topics depending on the virtual agent's personality. We have implemented it using an OCC-based representation of emotions and a PAD model for moods.

Keywords

Theory of Mind, Strategic intentions, Affective model, Job interview.

1 Introduction

In order to build a credible interaction between a human and a virtual character, affective computing [Picard, 1995] proposes to simulate human affects in virtual agents, making them more realistic and engaging for interactions. In this context, one main challenge for Artificial Intelligence researchers is to make the virtual character adapt its behaviour to the perceived user's affective state, which will lead to a more natural and credible interaction for the user. To this purpose, we claim that virtual characters must not only use reactive behaviour in answer to a wide range of affects (emotions, moods, social attitudes...) such as in [Marsella et al., 2004, Kriegel and Aylett, 2008, Jones and Sabouret, 2013, Schröder, 2010]. It must also use *strategic intentions* about the human it interacts with. Strategic intentions can be seen as long term goals [Haddadi, 1996] for an agent. Indeed, in an interaction, people have intentions about the goal of a conversation, such as obtaining an information, finding an agreement, changing the interlocutor's point of view or having a fun and relaxing conversation. This paper proposes to analyse these strategic intentions and to use them in the reasoning model of an affective agent. To this purpose, we define a general model that can be adapted to different context. In our work, we apply this general model to a specific case of a formal interaction: job interviews in which the goal of the recruiter is to obtain concrete information about the interlocutor's social and technical skills, so as to select the best candidate.

Our general model is based on logical rules and is inspired by the theory of mind [Baron-Cohen, 1995] paradigm. Based on affect perception from Social Signal Interpretation (SSI), our virtual agent's model derives beliefs about the interlocutor's self-estimation in the interview (job skills, importance of the salary, etc). These informations are confronted to the agent's goals so as to select the next course of actions in the interaction (in our case, to conduct the job interview).

This paper is organized as follows. Section 2 makes a brief state of the art on theory of mind and shows how this has been used in the context of a virtual agent's reasonner. Section 3 briefly presents the job interview context and its specific features. Section 4 presents our architecture. In Section 5, we present in details our affective model that integrates the theory of mind for reasoning about affects in the context of interactions. The rules of this general model are illustrated on examples from the job interview context. The last section concludes on the model and its application to the job interview situation.

2 Related work on Theory of Mind

Theory of mind [Leslie, 1994, Baron-Cohen, 1995], or *ToM*, is the ability to attribute mental states (beliefs, intentions, desires, affects, ...) to others. The literature reports

numerous ToM studies and implementations in agent-agent interaction [Bosse et al., 2007, Dastani and Lorini, 2012]. In our work, we want to model the reasoning process of an agent that reason about the reasoning process of a human (the applicant). This particular configuration raises additional difficulties and leads to a original model for our representation of the ToM. For example, in [Pynadath and Marsella, 2005], an agent has beliefs about others in a subjective way. Agent A has belief about agent B following the real structure of agent B beliefs. However, in our work, since agent B is the human applicant, we do not have any information about its belief structure. We must guess them from the outputs of the affect recognition module.

Nevertheless, this model of influence and belief change [Pynadath and Marsella, 2005] is based on work in psychology: the authors use influence psychological factors in their simulation framework: consistency, self-interest, speaker's self-interest and trust (or affinity). We believe that similar high-level reasoning structures must be proposed in reasoning models, to complement low-level reasoning on perceptions such as what is done in [Scassellati, 2002, Peters, 2006]. These papers focus on the perception aspects of ToM such as the desire of engagement, and are tailored for signal interpretation, not for the cognitive model of the virtual agent.

Several other applications have been studied with a Theory of Mind approach. For instance, [Bosse et al., 2007] proposes a reasoner for task avoidance, the agent can change its behaviour in order to alter the other agent's desires, intentions and *in fine*, actions to occur. This work has been extended in a more generic version [Bosse et al., 2011] that proposes a two-level BDI agent model: the first level is the agent's reasoner and the second one computes the ToM. Following a different approach, [Dastani and Lorini, 2012] also propose a model based on modal logics that extend the BDI paradigm. Each agent has a set of actions and a set of formulas that represent the agent's mental state. A formula has a degree of desirability for the agent and a degree of plausibility. The use of modal logic allows researchers to model the recruiter beliefs, desires and intentions, but it seems difficult to represent a real humans' mental states based only on perceptions.

This is the reason why we propose a model based on general rules that takes as inputs recognised affects from the interlocutor and strategic intentions for the virtual agents, and combines them in the ToM-based affective model. The goal of our model is to represent the reasoning process of an agent that reason about the reasoning process of a human. Our model will be applied and illustrated in the context of the TARDIS project[1] that considers a job interview simulation as an interaction.

[1]TARDIS stands for Training young Adult's Regulation of emotions and Development of social Interaction Skills. url: `http://www.tardis-project.eu/`

3 Job interview context

In job interviews, the recruiter needs to reason about the potential behaviour of the applicant in front of him. This evaluation is done by selecting different questions in order to provoke particular reactions on the interviewee [Rynes and Connerley, 1993]. For example, to test the applicant's capability to manage his or her stress, the recruiter can be voluntarily aggressive during the interview. Our goal in the TARDIS project is to model that kind of recruiter strategic intentions. To this purpose, we propose in the next section a formalism to represent these strategies and to reason about the applicant's state of mind, based on perceived social and emotional signals.

It is important to note that job interview simulation is an interesting situation for studying multimodal affective interaction with a virtual agent. The literature shows that expressed emotions and non-verbal behaviour play a key role in job interview: 1) it is used by the recruiter for evaluating the candidate [Sieverding, 2009] and 2) it influences the applicant's behaviour [Sieverding, 2009]. In addition, [Gatewood et al., 2010] shows that the recruiter uses different questions to assess the applicant's work performance, individual quality and specific regarding the job.

In our work, we take into account these three aspects. First, the applicant's non-verbal performance is evaluated by comparing expected social cues to detected ones, using the SSI system. Second, we compute an affective reaction (see section 5.1) for the recruiter (that should influence the candidate's behaviour). Third, we select the next topic of interest for the recruiter that tries to evaluate the applicant's competencies (see section 5.2). We propose several strategies for a recruiter (in section 5.6), from provocative to helpful, which change this topic selection (and its affective reaction) so as to influence the applicant's behaviour.

4 TARDIS general architecture

Figure 1 shows our global architecture. The TARDIS architecture considers four main components:

- The SSI component provides the affective model with information about the applicant's affects and social attitudes that are detected by the system.
- The Interview Scenario component tells the virtual recruiter the expectation in terms of emotions and attitudes, depending on the interview progress. In TARDIS, the agent has no understanding of the applicant's actual answers to the questions. It follows a scenario, that can be influenced by the recruiter perception and internal states and focus on the affective recognition and adaptation.
- The Animation component is responsible for expressing the virtual recruiter's affective state through its behaviour and expressions.
- The virtual recruiter component which is composed of two modules:

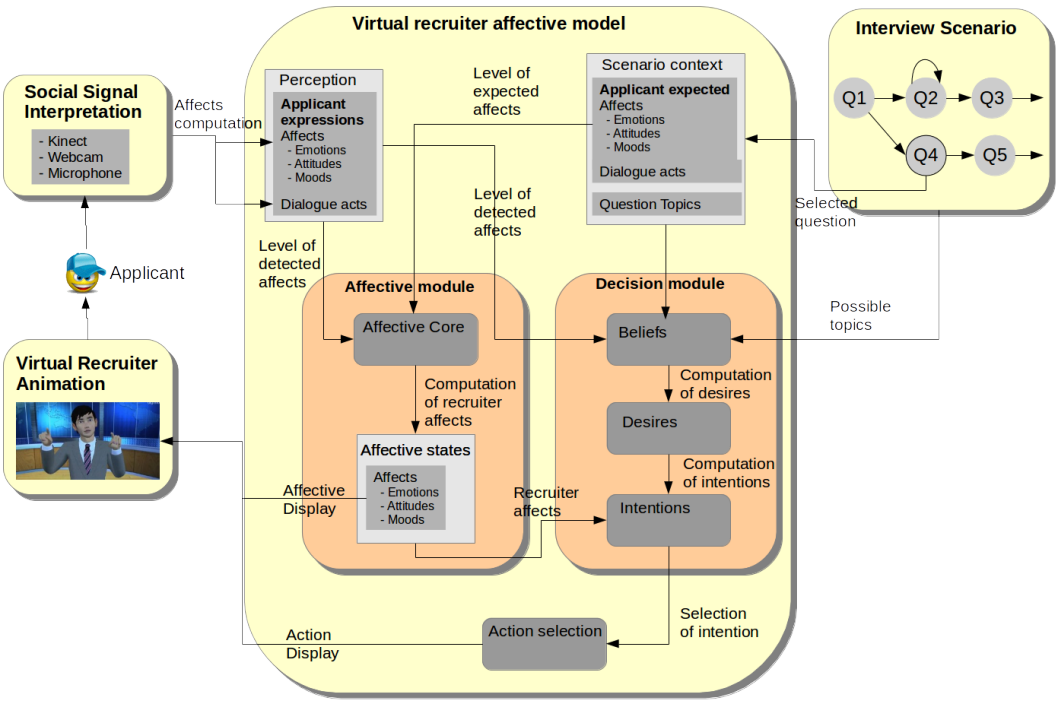

Figure 1: Global architecture for a recruiter in a job interview - Affective and Decision modules

- The Affective module, which is detailed in [Jones and Sabouret, 2012]. It provides a reactive model based on expectations from the recruiter and SSI of the applicant's affects expression [Jones and Sabouret, 2013]. It allows a computation of the recruiter emotions, moods and social attitudes.
- The Decision module, that is the focus of this document. The goal of this module is to build a theory of mind for a cognitive agent in the context of a job interview. Our agent (the recruiter) will deduce intentions of the applicant considering its answers (based on SSI) in a particular context (the question that has just been asked by the recruiter). This model will also influence new questions.

5 A ToM-based model for a cognitive virtual agent

In this section, we will present our model, and then shows its application in the TARDIS project.

5.1 General model for theory of mind

Our main objective is to draw beliefs about the interlocutor's mental states, preferences and understanding of the situation in the course of human-agent interaction, based on the user's reaction in terms of non-verbal behaviour (and social signal interpretation). In human-agent interaction, the agent has to select a new question at each turn-taking. To choose each new question, it is interesting that the agent has an idea of interlocutor's mental state regarding the past questions. This can be used in a wide spectrum of domains in which a human interacts with a virtual avatar in an interview simulation, such as teaching, training, ... The common aspect of these simulations is the use of questions by the avatar. Our model considers the use of question in order to manage the context of the answers of the person in interaction with the simulation.
To summarize, our theory of mind model has three main properties:

- It is about a real person who interacts with the system,
- It is centred on the person's preferences, expectations and interest for the job,
- It uses the context of questions and the affective behaviour to analyse user responses.

Our ToM however does not include any memory, other that the immediate response of the person and the previous ToM. We do not represent the knowledge that the interlocutor might have acquired during the interaction: we focus on its (estimated) preferences and expectations.

5.2 Context Management

With a view to manage the context, labels are given to the questions/sentences of the virtual character in order to interpret the answer/reaction of the human in term of beliefs on some topics. A list of topics can be done for each specific application. The set of topics set_{topic} contains N topics: $\{topic_1, topic_2, \ldots, topic_N\}$. Each subject is applica-

tion dependent and based on the domain of the simulation. A question is concerned by 0 to n topics.

List of topics. In order to manage the context, some labels are given to the questions of the recruiter in order to interpret the answer of the applicant in term of beliefs on certain subject. Here is the set of topics set_{topic} that can be tagged for a job interview:

- $topic_{applicant}$: questions about the applicant (general questions),
- $topic_{job}$: questions about the job,
- $topic_{salary}$: questions about the salary,
- $topic_{hours}$: questions about the working hours,
- $topic_{skill}$: questions about the competencies, the skills of the applicant regarding the job,
- $topic_{socialSkill}$: questions about the general social skills of the applicant,

A question is concerned by 0 to n topics. For example, the question "In what position will you like to work in our enterprise?" can be tagged by two different topics: $topic_{skill}$ and $topic_{job}$ because it tells about the applicant skills (the position he thinks he can apply for in this job) and its knowledge of the job (organisation of the enterprise).

5.3 Beliefs build

In order to build beliefs about the human who interacts with the system, we consider the questions/sentences that were just expressed by the virtual agent (identified by labels about topics) and the quality of the answer of the human from an affective point of view (which is obtained by Social Signal Interpretation, or SSI). Based on that, the agent will update its beliefs about the human on a particular subject. We denote the beliefs of the agent about the human $B_{Human}(topic_i)$ for i in $\{1, \ldots, N\}$.
According to the topic(s) raised by the question/remark of the agent, beliefs will be updated. In pursuance of building the beliefs of the human, we consider its answer (perceived via SSI) and decide if the answer is rather positive, negative or neutral. In order to determine if the global answer is positive or not, we use a performance index that compares the expected social cues (such as smile, large gestures, body movement, directed gaze...) with the detected ones. Expectations can be expressed as positive (signals that should be detected) or negative (behaviours that the user should avoid). This method in presented in [Jones et al., 2014]. The performance index pi is valued in the interval $[0, 1]$.
Based on this value and the topic tags of the question/remarks just done by the agent, the beliefs can be computed. Updates of each belief are done with the following formula for each topic:

$$B_{Human}(topic_i) \leftarrow B_{Human}(topic_i) + \alpha \times pi$$

with $\alpha \in [0, 1]$ a value that can be altered if we want the recruiter beliefs about the human to evolve quickly ($\alpha = 1$) or not (α near of 0). It can rely on the personality of the agent. An impulsive agent has an α near of 1 and a moderate one near of 0.
For instance, after a question about the job, with $\alpha = 1$ (impulsive recruiter) and an actual belief value $B_{Human}(job) = 0.5$, $B_{Human}(job)$ will become -0.3 for an $AverageAns$ of -0.8 making the recruiter beliefs about the applicant change sign in one answer. A moderate recruiter ($\alpha = 0.2$) will obtain a belief $B_{Human}(job) = 0.34$ which will change the dynamic of our simulation. An impulsive recruiter will cause strong dynamics and a moderate one smooth ones, which is the expected behaviour.

List of beliefs. The list of beliefs we consider in the context of job interviews is the following:

- self-confidence $B_{Human}(young)$,
- knowledge about the job $B_{Human}(job)$,
- importance of the salary for the applicant $B_{Human}(salary)$,
- importance of scheduling and working hours for the applicant $B_{Human}(hours)$,
- qualities of job skills $B_{Human}(skills)$,
- qualities of social skills $B_{Human}(socialSkills)$,

According to the topic raised by the question of the recruiter, beliefs can mean different things. For instance $B(young)$ is the belief of the applicant in himself. For a belief equal to 1, the applicant is very confident, and for -1, he has an important lack of self-esteem. Here the value quantifies the confidence of the applicant. If we look at $B(salary)$, the signification is different, it is about the importance that the applicant put in the salary when applying for this job.

5.4 Desires and goals

The desires are used to define the strategic intentions of the agent. We organize our desires in two categories: the high-level ones and the more specific ones. The high level intentions are directly linked to social attitudes. Attitudes can be initialized with personality and can evolve during the simulation but with a dynamics slower than the emotional one which is quite reactive. For more detail about the computation of social attitudes, refer to [Jones and Sabouret, 2012].
The high-level intentions are about the general intentions of the agent for the interaction, the specific ones are about specific beliefs about the human that interest the agent during the interaction.
The high-level desires are denoted: $D(Attitude)$. The specific desires are denoted: $D(B_{Human}(topic_i))$ because specific desires in an interaction are about beliefs of the human on a particular topic. For instance $D(B_{John}(football)$ is the desire of the agent to know if John has knowledge in the $football$ topic.

List of desires and goals. For a job-interview simulation, the recruiter will have a limited set of high-level intentions (provocative, pugnacious, friendly and helpful) and only one of them will be triggered in the same time.
The specific intentions are about subjects that the recruiter want to favour during the interview. The level of each subject will be adapted in function of the high-level intentions and will also consider the beliefs about the applicant. Actually, these specific goals for the recruiter are about the knowledge of applicant's beliefs on certain subjects. For instance, a question about the job will be associated to the goal $G(B_{Human}(job))$ because this question will give more information to the recruiter about the belief $B_{Human}(job)$.

5.5 Dynamics of goals

The high-level desires evolve in function of the social attitude of the agent. Social attitudes used can be defined on Leary circumplex [Leary, 1996]. According to the application, some attitudes will be relevant and some not. As shown by Leary, attitudes can be separated in two categories, the positive ones (friendly, cooperative, extroverted, ...) and the negative ones (hostile, critical, ...).
Based on these two kind of attitudes, we define algorithm 1 in order to update the desires of the agent.

Algorithm 1 Desires computation

if $(Attitude \in set(attitude_{-}))$ **then**
  for $B_{Human}(topic) \in set_{topic}$ **do**
    if $(AverageAns < 0)$ **then**
      $D(topic) \leftarrow D(topic) + \alpha \times |AverageAns|$
    else
      $D(topic) \leftarrow D(topic) - \alpha \times |AverageAns|$
if $(Attitude \in set(attitude_{+}))$ **then**
  for $B_{Human}(topic) \in set_{topic}$ **do**
    if $(AverageAns < 0)$ **then**
      $D(topic) \leftarrow D(topic) - \alpha \times |AverageAns|$
    else
      $D(topic) \leftarrow D(topic) + \alpha \times |AverageAns|$

This algorithm works as follows: if the agent has a negative attitude, he intends to select topics with a negative answer for the human. On the contrary, if the agent has a positive attitude, its desires are about topics with a positive answer from the human.

5.6 Goal selection

Several strategies can be defined for the selection of one desire in the list of possible desires. The most natural one is to select the desire with the maximum value in the available desires. At one moment of the dialogue, every possibilities (topics) cannot be approached in order to conserve the logical sequence of the conversation (a scenario for instance).

Goal selection based on recruiter's high level intentions. The high level goals can be defined directly in the scenario or be computed on the personality of the recruiter. We define 4 main strategies (2 for the positive attitudes and 2 for the negative attitudes). A recruiter with positive attitudes will have positive desires on topics where the applicant has positive average answers. On the contrary, a recruiter with negative attitude will have positive desires on topics where the applicant has negative average answers. Here are some strategies that we use for the virtual recruiter:

- Provocative recruiter: the recruiter will have a negative attitude and will always select the worst topic for the user (the one with the maximum Desire for the negative agent).
  $Intention = max(B_{Human}(subject))$
- Pugnacious recruiter: the recruiter will have a negative attitude but will randomly select one of the worst topic but not always the same.
  $Intention = random(max_n(B_{Human}(subject)))$
  with max_n, the n worst subjects.
- Friendly recruiter: the recruiter will have a positive attitude and will randomly select one of the best topic for the user but not always the same.
  $Intention = random(max_n(B_{Human}(subject)))$.
- Helpful recruiter: the recruiter will have a positive attitude and will always select the best topic (the one with the maximum Desire for the positive agent).
  $Intention = max(B_{Human}(subject))$

These different strategies lead to different goals. At one moment of the interaction, only some subjects can be approached according to the possibilities of the scenario. The recruiter will select the maximum in the available choices.

6 Conclusion

In this article, we propose a theory of mind model for an affective virtual agent. The theory of mind is about a real person in interaction with the system. It is centred on the interpretation of the affective states perceived through Social Signal Interpretation. By building beliefs about the person in interaction with the simulation, we allow an interaction by understanding the subjects where the person is confident or not. Then, according to the virtual agent high level intentions, new questions will be selected in a coherent and credible way regarding the personality of the agent.
This work is actually in the process of integration in the TARDIS platform. After integration, it will be evaluated in order to confirm that our model proposes a credible virtual recruiter for a job interview scenario. The theory of mind should provide coherent actions of the recruiter according to the reactions of the applicant and the personality of the recruiter. This aspect can be evaluated through literature and thanks to the applicants that will interact with the system. One current limit of our model is that it requires manual annotation of the scenario. The definition of an

automated annotation process, based on the utterance's semantic and contextual information, would greatly increase the scalability of the model.

Acknowledgment

This research is funded by the European Union Information Society and Media Seventh Framework Programme FP7-ICT-2011-7 under grant agreement 288578.

References

[Baron-Cohen, 1995] Baron-Cohen, S. (1995). *Mindblindness*. MIT Press, Cambridge, Massachusetts.

[Bosse et al., 2007] Bosse, T., Memon, Z. A., and Treur, J. (2007). A two-level BDI-agent model for theory of mind and its use in social manipulation. In *In AISB 2007 Workshop on Mindful Environments*, pages 335–342.

[Bosse et al., 2011] Bosse, T., Memon, Z. A., and Treur, J. (2011). A recursive BDI agent model for theory of mind and its applications. *Applied Artificial Intelligence*, 25(1):1–44.

[Dastani and Lorini, 2012] Dastani, M. and Lorini, E. (2012). A logic of emotions: from appraisal to coping. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '12, pages 1133–1140, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[Gatewood et al., 2010] Gatewood, R. D., Feild, H. S., and Barrick, M. (2010). *Human resource selection*. South-Western Pub.

[Haddadi, 1996] Haddadi, A. (1996). *Communication and Cooperation in Agent Systems - A Pragmatic Theory*. Springer.

[Jones and Sabouret, 2012] Jones, H. and Sabouret, N. (2012). An affective model for a virtual recruiter in a job interview context. In *4th International Conference on Games and Virtual Worlds for Serious Applications, Genoa, Italy, 29/10/12-31/10/12*, page in press. VS-GAMES'12.

[Jones and Sabouret, 2013] Jones, H. and Sabouret, N. (2013). TARDIS - A simulation platform with an affective virtual recruiter for job interviews. In *IDGEI (Intelligent Digital Games for Empowerment and Inclusion)*.

[Jones et al., 2014] Jones, H., Sabouret, N., Damian, I., Baur, T., André, E., Porayska-Pomsta, K., and Rizzo, P. (2014). Interpreting social cues to generate credible affective reactions of virtual job interviewers. *IDGEI (Intelligent Digital Games for Empowerment and Inclusion)*.

[Kriegel and Aylett, 2008] Kriegel, M. and Aylett, R. (2008). Emergent narrative as a novel framework for massively collaborative authoring. In *Intelligent Virtual Agents*, pages 73–80. Springer.

[Leary, 1996] Leary, T. (1996). Interpersonal circumplex. *Journal of Personality Assessment*, 66(2):301–307.

[Leslie, 1994] Leslie, A. M. (1994). ToMM, ToBy, and agency: Core architecture and domain specificity. In Hirschfeld, L. A. and Gelman, S. A., editors, *Mapping the mind Domain specificity in cognition and culture*, chapter 5, pages 119–148. Cambridge University Press.

[Marsella et al., 2004] Marsella, S. C., Pynadath, D. V., and Read, S. J. (2004). PsychSim: Agent-based modeling of social interactions and influence. In Munro, P., editor, *Proceedings of the International Conference on Cognitive Modeling*, volume 36, pages 243–248. Citeseer.

[Peters, 2006] Peters, C. (2006). A perceptually-based theory of mind for agent interaction initiation. *International Journal of Humanoid Robotics*.

[Picard, 1995] Picard, R. W. (1995). Affective Computing. *Emotion*, TR 221(321):97–97.

[Pynadath and Marsella, 2005] Pynadath, D. V. and Marsella, S. C. (2005). PsychSim : Modeling Theory of Mind with Decision-Theoretic Agents. *Information Sciences*, 19(1):1181–1186.

[Rynes and Connerley, 1993] Rynes, S. and Connerley, M. (1993). Applicant reactions to alternative selection procedures. *Journal of Business and Psychology*, 7(3):261–277.

[Scassellati, 2002] Scassellati, B. (2002). Theory of Mind for a Humanoid Robot. *Autonomous Robots*, 12(1999):13–24.

[Schröder, 2010] Schröder, M. (2010). The SEMAINE API: towards a standards-based framework for building emotion-oriented systems. *Advances in human-computer interaction*, 2010:2.

[Sieverding, 2009] Sieverding, M. (2009). 'Be Cool!': Emotional costs of hiding feelings in a job interview. *International Journal of Selection and Assessment*, 17(4).