Catch me if you can: exploring lying agents in social settings

Matthias Rehm, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Rehm, Matthias, and Elisabeth André. 2005. "Catch me if you can: exploring lying agents in social settings." In *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems, The Netherlands, July 25 - 29, 2005*, edited by M. Pechoucek, D. Steiner, and S. Thompson, 937–44. New York, NY: ACM. https://doi.org/10.1145/1082473.1082615.

Catch Me If You Can — Exploring Lying Agents in Social Settings

Matthias Rehm
Multimedia Concepts and Applications
Faculty of Applied Computer Science
University of Augsburg, Germany
rehm@informatik.uni-augsburg.de

Elisabeth André
Multimedia Concepts and Applications
Faculty of Applied Computer Science
University of Augsburg, Germany
andre@informatik.uni-augsburg.de

ABSTRACT

Embodied conversational agents become more and more realistic concerning their conversational and their nonverbal behaviors. But if the information conveyed nonverbally exhibits clues that are not consistent with the verbal part of an agent's action, how will the user react to such a discrepancy? Masking ones real emotions with a smile is a naturally occuring example of such a discrepancy. But such masks are often deficient and thus subtle clues of lying and deceiving manifest themselves in facial expressions. The questions is how users will react to these clues if they are conveyed by an agent. Will they render an application unattractive or on the contrary more human-like? In this paper, we examine such facial clues to deception and present the results of two empirical studies: i.) lies in monologues by a talking head presenting movies, ii.) lies in an interactive game of dice.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Artificial, augmented, and virtual realities; Evaluation/methodology; H.5.2 [**User Interfaces**]: Evaluation/methodology

General Terms

Design, Human Factors

Keywords

Embodied Conversational Agents, Non-verbal Behavior, Engagement, Deception

1. INTRODUCTION

In human-human communication, emotions are the number-one topic that people lie about, and studies show that up to 30% of social interactions longer than 10 minutes contain such deceptions [8]. Usually, social lies are employed to protect the face of others or the relationship to others. A typical example are excuses, such as "I would love to join you, but ...". Even though a significant amount of work has been devoted to the development of synthetic agents that emulate aspects of social interactions between humans, the simulation of social lies and deception are nearly non-existing topics.

McKenzie and colleagues [17] discuss the potential benefits of deceptive agents as training partners that help the user to recognize malicious intent, but do not present any implementation. Castelfranchi and Poggi [5] developed a theory of deception in communication which has grounded prototyping of a deception modeling tool in which both the deceiver and the receiver of the message are modeled [3, 6]. The issue of deception has also been addressed in the area of conversational systems [15] and in the area of multi-agent systems where different strategies of deception and their effects on the interaction of agents are explored [4, 28].

Besides work on cognitive models for deceptive agents, various attempts have been made to create synthetic agents that deliberately oppress or express a certain emotion. Pelachaud and colleagues [7] as well as Prendinger and colleagues [21] developed agents that are able to control their emotional displays if the social situation requires it. For instance, if the social distance between an agent and its conversational partner is high, Prendinger's agent would not show anger to the full extent. The virtual tutor COSMO [16] intentionally portrays emotions with the goal of motivating students and thus increasing the learning effect.

All these approaches start from the assumption that the agent is able to perfectly hide emotions if the social or pedagogical situation requires it. However, humans are not always capable of completely concealing their true emotions. For instance, masking smiles cannot entirely override the muscular program of the original emotion because not every facial muscle can be consciously controlled. As a consequence, such a mask will always include segments of one's felt emotion. The question arises of how to handle situations in which the agent decides to display an emotion which is in conflict with its internal appraisal processes. In some situations, it might be desirable to employ agents that perfectly convey "wrong" emotions with the aim to convince the interlocutor. Consider a sales agent on the web that has to advertise a product of minor quality. If it does not succeed in concealing its negative attitude towards the product, a decrease of the sales might be the consequence. On the other hand, agents in social settings may come across as little believable or cold if they are always able to per-

fectly hide their true emotions. In addition, the display of mixed emotions may even lead to a positive response from the interlocutor. For instance, students may feel sympathy towards a virtual teacher that desperately tries to hide its negative emotions caused by the students' bad performance. Last but not least, the emulation of deceptive behaviors may enrich our interactions with synthetic agents - especially in game-like environments.

In this paper, we will focus on synthetic agents that may express emotions that are in conflict with their appraisal processes. Unlike earlier work, we will model situations in which the agent fails to entirely conceal its "felt" emotions. We will also investigate the influence of such a behavior on the relationship between agent and user.

Earlier studies examine whether or not the presence of a synthetic agent contributes to the building of trust and how this process may be supported by the agent's conversational behaviors [26, 19, 1]. None of these works focuses, however, on the impact of subtle emotional expressions on the users attitude towards the agent. In this paper, we will present the results of two experiments we conducted in order to find out whether users notice a difference between agents that fake emotions perfectly and agents that reveal their "true" emotions by deceptive clues. Furthermore, we will investigate whether humans are able to correctly interpret such subtle clues of deception. For instance, it might be the case that users notice the agent's deceptive behaviours, but attribute them to a bad design of the agent or a malfunction of the system. Thirdly, we will analyze in how far the conversational setting influences the user's sensitivity towards deceptive clues. In particular, we will compare monologues in which the user just observes the agent with multi-player game scenarios in which the user actively interacts with the agent and other users.

2. A FIRST MODEL

The objective of our work is to develop an agent whose behaviors may reflect potential conflicts between "felt" and deliberately expressed emotions. As a first step, we concentrate on facial expressions of deception which have been profoundly researched in the psychological literature. According to Ekman [9], there are at least four ways in which facial expressions may vary if they accompany lies and deceptions: micro-expressions, masks, timing, and asymmetry.

1. *Micro-expressions:* A false emotion is displayed, but the felt emotion is unconsciously expressed for the fraction of a second. The detection of such micro-expressions is possible for a trained observer.

2. *Masks:* The felt emotion (e.g., disgust) is masked by a non- corresponding facial expression, in general by a smile. Because we are not able to control all of our facial muscles, such a masking smile is in some way deficient. Thus, it reveals at least in part the original emotion.

3. *Timing:* Facial expressions accompanying true emotions do not last for a very long time. Thus, the longer an expression lasts the more likely it is that it is accompanying a lie. A special case seems to be surprise, where elongated on- and offset times are a good indicator of a false emotion.

4. *Asymmetry:* Voluntarily created facial expressions like they occur during lying and deceiving tend to be displayed in an asymmetrical way, i.e., there is more activity on one side of the face than on the other.

To model the non-verbal behavior, we employ the Greta[1] agent system developed by Catherine Pelachaud and colleagues [20, 13]. This agent is compliant with the MPEG-4 standard which allows to control the facial expressions and body gestures by so-called facial animation parameters (FAPs) and body animation parameters (BAPs). Due to technical reasons, we had to limit our evaluations to masks and asymmetry. A more recent version of the Greta agent also enables the specification of micro-expressions and exact timing of expressions, and we will extend our model by these features in the near future.

Since it was not possible to get the original video material from Ekman due to legal reasons, we reconstructed the animations for the deceiving condition out of pictures from the original studies and based on textual descriptions of the facial clues found in [9], [10], and [11]. We concentrated on masking smiles for disgust, sadness, anger, and fear. Different masks are deficient in several aspects. For instance, we considered masks where the eyebrows are still frowning in anger, but the mouth displays a normal smile as well as masks where the frown is not very articulated and there is only a small smile. Different degrees of masking are combined with different degrees of asymmetry of the facial displays resulting in 32 possible facial expressions.

3. LIES IN MONOLOGUES

The objective of the first experiment was to measure the impact of deceptive clues on the user's subjective impression of the agent. On the basis of related studies, we expected more negative ratings for the deceptive than for the non-deceptive agent. Studies by Swerts and colleagues [25] indicate that humans are to a certain extent able to distinguish speakers with low confidence from speakers with high confidence based on audio-visual cues. We assumed that we would obtain similar results for presentations with a talking head. Nass and colleagues [19] observed that subjects perceive inconsistent agents less positively than consistent agents. Since our deceptive agents have to handle potential conflicts between felt emotions and emotions to be conveyed, they may end up with behaviors that appear as inconsistent and thus cause a similar effect as the inconsistent agents in the experiment by Nass and colleagues.

To explore this hypothesis, we devised an experimental setting in which different versions of the Greta agent presented movie reviews to the users. The comments of the agents were always positive implying that the agents liked the movies. As independent variable, we defined the absence (N) or presence of deceptive clues (L) which was manipulated within subjects. In condition N, the facial expressions corresponded to the positive comments. In condition L, facial clues indicated that the agent might have a different opinion on the movies than it actually verbalized.

In order to control for assumptions concerning preferred movie genres for Greta, a pretest was done with 10 students and two of our colleagues. The subjects were shown a

[1]We are grateful to Catherine Pelachaud, Maurizio Mancini, and Björn Hartmann for supporting our work with the Greta Agent.

Figure 1: A natural vs. a disgust-lie smile

Characteristic features	Agent	Result
Reliable	L	0.07
	N	0.29
Trustworthy	L	0.14
	N	0.32
Convincing	L	0.18
	N	0.39
Credible	L	0.14
	N	0.43
Certain	L	0.11
	N	0.32

Table 1: Results of the questionnaire

small presentation by Greta about the local weather over the last two weeks. This topic was chosen due to its small talk character [1]. Afterwards, subjects had to indicate which out of ten different movie genres Greta might prefer. There was no clearly identifiable genre the subjects associated with Greta's preferences. For our first evaluation, we decided to restrict ourselves to five out of the ten genres: horror, adventure, animation, comedy, and action.

Two movies were chosen for each genre that were at least among the top five box office hits in Germany to ensure an even distribution of knowledge about the single movies. After that, two versions for the presentation of each movie by the Greta agent were produced. In condition (N), only facial expressions from the original Greta library were employed. In condition (L), the utterances were accompanied by expressions we modeled to realize masking smiles (see Section 2). For example, in condition (N), the agent would say "I really enjoyed the happy end" accompanied by a true smile, i.e. express the emotion joy. In condition (L), the same utterance would be accompanied by a facial expression that masks disgust by a smile. Thus, the "true" emotion (disgust) leaks out in the facial expression of the agent. Figure 1 shows an example. Some features of disgust are present in the disgust-lie smile. The upper lip is raised, the eyelids are closed to a certain degree and the eyebrows moved inward and down. These features are less articulated than in real disgust, and they are blended with the natural smile. Moreover, there is asymmetry in the smile with more smiling action on the right side of the face. From the possible 32 facial expressions we created (see Section 2), only 7 were used frequently (2-8 times) during the presentations to control the relevant features more firmly. For both conditions, (neutral) synthetic speech was used.

In order to enable unique references to the single agents in the questionnaire, the different versions were represented by different hair colors: a blond and a brunette agent. To prevent unwanted biases by the hair color of the agents, two groups of students were tested. The first group consisted of 11 students. In this group, the brunette agent represented the non-deceiving (N), the blond one the deceiving condition (L). The second group consisted of 18 students. Here, the deceiving condition was represented by the brunette, the non-deceiving condition by the blond agent. At the beginning of the experiment, the subjects were given an introduction by both agents to get acquainted with the mimics and the synthesized voice of the agents.

After that, ten movies were presented in a row where the order in which the agents appeared was varied. Both agents presented one movie from each of the five genres. Subjects were told that we were interested in testing different synchronization methods for speech and facial displays and requested to pay special attention to the agents' mimics. In order to make sure that the subjects listened to the agents' reviews carefully, they were furthermore told that we wanted to investigate the effect of the different animation methods on memory and that they would be asked questions concerning the content of the presentations afterwards.

After the presentations, the subjects had to answer a questionnaire about the content of the presentation and the quality of the animations (synchronization of media, appropriateness of mimics and voice quality). The questionnaire also contained questions about characteristic features of the agents (trustworthy, convincing, sympathetic, credible, certain, dependable, reliable, competent, professionally competent). The subjects were asked to indicate for which of the agents a certain feature fits better (blond or brunette aka deceiving/non-deceiving). They could also mark a "no difference between agents"-box. In case they preferred one of the agents they had the option of giving reasons for this preference. If the subjects preferred an agent at all in respect to the tested features, they named the non-deceiving agent. The results for the crucial characteristic features are given in Table 1. Due to the high number of subjects with no preference, the results are only weakly significant, i.e., at a confidence level of $p < 0.1$.

Obviously, the non-deceiving agent is perceived as being more reliable, trustable, convincing, credible, and more certain about what it said. We interpret this finding as evidence that the subjects were effected by the facial clues of deceit shown by the deceptive agent in condition (L), other things being equal. The deceptive agent's "real" emotions about the movies leak through her facial expressions and make the propositional content of the agent's utterances, i.e., the praise of the presented movie, appear less likely. We did not observe any significant differences in the quality ratings of the animations for the two agent versions. Furthermore, the subjects could not name the reasons for their uneasiness with the deceptive agent.

The question arises of why we did not observe a much higher preference for the non-deceptive agent in terms of percentages. These results are in line with findings on facial clues to deception [9]. Because such clues are only subtle, some training is required to recognize them for certain during an interaction. Thus, it could not be expected that the

subjects would definitely detect the relevant features especially since they did not have any information on the true purpose of the experiment. Furthermore, as Krahmer and colleagues [14] notice inconsistencies between nonverbal cues might be less offensive than inconsistencies between verbal and nonverbal cues. We would also like to emphasize that we focused deliberately on the simulation of subtle signals of deception even though a synthetic character like Greta would of course also allow for more extreme facial expressions. Finally, our clues of deception were restricted to asymmetry and masking so far.

4. LIES IN INTERACTIVE GAME SCENARIOS

The first experiment has shown that subjects obviously rate an agent more negatively if it uses deceptive clues. The question arises, however, of whether the user would be able to correctly identify the cases in which the agent was telling the truth and in which cases it was lying. Secondly, we were wondering whether a deceptive agent would also be rated more negatively than a sincere agent in a situation in which lies are socially desirable.

To shed light on this question, we devised an interactive scenario called GAMBLE where two users play a simple game of dice (also known as Mexicali) with the agent. To win the game it is indispensable to lie to the other players and to catch them lying to you. The traditional (not computer-based) version of the game is played with two dice that are shaken in a cup. Let's assume player 1 casts the dice. He inspects the dice without permitting the other players to have a look. The cast is interpreted in the following way: the higher digit always represents the first part of the cast. Thus, a 5 and a 2 correspond to a 52. Two equal digits (11, ..., 66) have a higher value than the other casts, the highest cast is a 21. Player 1 has to announce his cast with the constraint that he has to say a higher number than the previous player. For instance, if he casts a 52, but the previous player already announced a 61, player 1 has to say at least 62. Now player 2 has to decide whether to believe the other player's claim. In this case, he has to cast next. Otherwise, the dice are shown and if player 1 has lied he has lost this round and has to start a new one. For the experiment, each player was equipped with a PDA which replaced the cup with the cubes in the original game.

As in the first evaluation, two conditions were tested: (i) an agent that does not show clues to deception in its facial expressions (N) and (ii) an agent that does exhibit these clues (L). The same facial expressions as before were used. One frequent comment in the first evaluation was the bad voice quality of the agent. To make the game more entertaining, we dubbed the animations with a real human voice. Moreover, a full body agent was used and a number of emblematic german gestures were modelled relying on the descriptions in the Berlin dictionary of everyday gestures ("Berliner Lexikon der Alltagsgesten", [2]).

In order to make sure that the users paid sufficient attention to the agent's behaviour (and did not just concentrate on the PDA display or the other user), they were told that the agent might not be able to conceal her emotions perfectly, but left it open how deceptive behaviours might be detected. Consequently, the subjects had no idea which channel of expression to concentrate on. To incite the subjects to find out whether the agent was lying or not, they were promised a reward of 5 Euro in case they won against the agent.

Figure 2: The setting

	Features	Condition	Result
1.	Caught	L	0.72
		N	0.73
2.	Falsely accused	L	0.54
		N	0.54

Table 2: Results of the GAMBLE evaluation

In this study, 24 students from computer science and communication studies participated. They were divided into 12 teams from which 6 teams were randomly assigned to the L-, the other 6 to the N-condition. Each team played two rounds of 12 minutes, participants changed position after the first round. Thus, each participant came to play after the agent and had to judge the agent's announcements (see Fig. 2). We videotaped the interactions and we logged the game progress for the analysis.

The log-files were used to calculate how often the agent was caught lying and how often it was falsely accused by the human players. Table 2 gives the results of this calculation. There was nearly no difference between the two conditions. The agent was caught in 72% of the cases in the L- and in 73% in the N-condition. It was falsely accused of lying 54% in the L- as well as in the N-condition. Furthermore, there were no significant differences in the user's subjective rating of the two agent versions. We did not expect the users to negatively rate the agents' deceptive behaviors in particular. After all, deceptive behaviors are a major element of the game. Nevertheless, we considered the possibility that the deceptive agent would get a less positive rating due to the inconsistencies in its behaviors as in the first experiment. However, the experiment led to clear and non-ambiguous results that were not expected in this clarity from the previous evaluation. Obviously the clues of deception that the agent exhibits in the L-condition are either disregarded, not interpreted or unintelligible and have no effect at all on the overall game progress. In the following, we will discuss sev-

look at	overall	acc.
yes	111	60
no	62	30
yes/no	14	4
no/yes	15	6

Table 3: Looking behavior of human players

eral reasons why we might have got different results in the first and the second evaluation.

Reason 1: Users decide without looking at the agent. In some demo runs we discovered that users tended to look at their PDAs instead of at the agent. To remedy this, the information displayed on the PDA was reduced to a minimum. Especially the announcements of the players are no longer displayed on the PDA making it necessary to attend to the other players and the agent. But perhaps users still continued deciding if the agent was lying without looking at the agent at all. To verify this hypothesis, we analysed the videotapes for the L-condition as to whether people looked at the agent during the agent's announcement of its casts. Table 3 provides the results. Leaving the cases aside in which subjects first looked at the agent and then away or vice versa, the agent was looked at in two thirds of the cases and only disregarded in one third of the cases. We also counted the number of times that the agent was accused (acc.) by the subjects giving us the same numbers, i.e., in two thirds of the cases the agent was looked at, in one third it was disregarded. Thus, users tend to look at the agent during its announcements and they tend to look at the agent before accusing it of a false announcement. Consequently, this hypothesis does not explain the outcome of our experiment.

Reason 2: Users decide based on voice quality. In our first evaluation, subjects gave us the feedback that the synthesized voice is hard to understand, sounds boring and does not go very well with the presentation of the movies. To make the interaction more natural and to enhance the entertaining factor of the GAMBLE system, we decided to dub the animations of the Greta agent with a human voice this time. Concerning the ease of interaction, this choice was a success because only one subject mentioned the voice quality as a problem in the GAMBLE system. Different to our first evaluation, the signals coming from the audio channel were more natural now than the signals coming from the visual channel which might have induced the subjects to over-interpret the agents' voice. During the game subjects occasionally commented on the voice quality by saying e.g., "That sounded like a lie."[2] or "Now she sounded shy."[3] and decided accordingly if the agent was lying or not. Thus, it might well be the case that people heavily relied on the verbal channel of communication for detecting deceptive behaviors. Unfortunately, we have only scarce evidence for this effect because users did not frequently spontaneously explain the rationale behind their decisions during the game.

Reason 3: Users decide based on what is said. To make the interaction in GAMBLE as natural as possible, a sample corpus of interactions between human players was utilised to create the agent's comments and announcements. Four types of agent actions were realized in this first version of GAMBLE: (i) reactions to the announcement of the previous player, such as "Are you crazy?"[4], (ii) reactions to the agent's own cast, such as "What shall I do now?"[5], (iii) announcements of the cast, such as "I have 62", and (iv) reactions to having won or lost, such as "You have lost". Of these types only (iii) is interesting for our analysis. In the corpus of the real players, announcements were found, such as "I have 24, sorry 42" or "I have 62, ehh 63". The first one is attributable to a misreading of the cast because in GAMBLE, the higher digit always comes first, thus a 2 and a 4 is always a 42. The second one is more crucial. Here, the repaired speech is likely to make the next player feel insecure as to whether this was just a slip of the tongue or an indication that the announcement is false. We made the observation that users usually interpreted speech repairs and hesitations as a sign of lying which was also reflected by their comments during the game. Subjects occasionally indicated their disbelief and irritation, but also their surprise that the agent was so cunning. This effect occurred of course in both conditions. Nevertheless, the rich verbal channel might have caused the subjects to more or less ignore the visual channel.

Reason 4: Users decide on objective criteria like probability of casts. It is very unlikely that a player lies about a low cast like 32 if it is sufficient for him. On the other hand, announcing a 55 makes it very likely that this is a lie because the probability to score 55 or above is $\frac{1}{9}$ which is roughly 11%. Table 4 gives the probabilities for the possible casts. Usually, it makes sense to start from the assumption that the probability to lie increases equally with the decrease in the probability of the cast. There are exceptions from this rule with experienced liars or at the beginning of a game where it is unlikely that a player starts with a lie. Taking the need to lie into account, the log-files were re-examined according to how often the agent was caught lying or was falsely accused of lying in regard to the probability of its announced cast. Table 5 presents the results. In the case of casts with a probability above 50% (31-54) there was no difference at all (1a, 0% vs. 0%). In fact, the agent very rarely lied and was not caught at all in either condition. In case of the casts above 50% (62-21) a slight tendency towards the L-condition can be seen (1b, 84% vs. 77%). Assuming that in the case of casts below 22% people tend to just disbelief their previous player, we looked into the differences between the L- and N-condition in the range of 44% to 22% (1c, 62-65). In this case, subjects in the L-condition were better lie catchers (70% vs. 53%). Next, we analysed how often users falsely accused the agent of lying, i.e., how often they didn't believe the agent's claim even though the agent told the truth. The results for this analysis (2a-c) are comparable and consistent with the analysis on how often the accusation was right. Looking into the cases where the probability is above (2a, 100% vs. 100%) or below 50% (2b, 51% vs. 50%), there is no difference at all. As mentioned above, the agent very rarely lied in the case of casts with a probability above

[2]Das klingt wie eine Lüge
[3]Jetzt hörte sie sich aber schüchtern an
[4]Du spinnst wohl?
[5]Was mach ich jetzt nur

Thr.	Prob.	Thr.	Prob.	Thr.	Prob.
31	1 =100%	53	.	11	4/18=22%
32	17/18	54	.	22	3.5/18
41	.	61	9/18=50%	33	3/18
42	.	62	8/18=44%	44	.
43	.	63	.	55	.
51	.	64	.	66	.
52	12/18	65	.	21	1/18

Table 4: Probabilities of the possible casts

	Features	Condition	Result
1.	a.) Caught (>50%)	L	0.0
		N	0.0
	b.) Caught (<50%)	L	0.84
		N	0.77
	c.) Caught (>22%, <50%)	L	0.7
		N	0.53
2.	a.) Falsely accused (>50%)	L	1.0
		N	1.0
	b.) Falsely accused (<50%)	L	0.51
		N	0.5
	c.) Falsely accused (>22%, <50%)	L	0.52
		N	0.69

Table 5: Results regarding probability of cast

50%, thus an accusation in this case is likely to be wrong and indeed all accusations in this value range are wrong regardless of the condition. Last we looked into the area we identified above as crucial, i.e., where the probability of the cast is between 44% and 22%. Again, a difference was found in this case. The agent is less often falsely accused in the L-condition (2c, 52% vs 69%). Although the differences are well explained by the game's logic, they are not statistically significant which is mostly attributable to the small number of lies encountered. In the L-condition, the agent lied 22 times, in the N-condition 19 times, i.e, 1.8 times per subject in the L-, 1.6 times in the N-condition. Given that subjects really decided rationally according to the probability of the casts, meaning that only in the crucial area below 50% they were interested in other features than this objective one, they had not much opportunity to learn how to interpret the agent's facial expressions. To verify this hypothesis a long-time study would be necessary, giving the participants the opportunity to get acquainted with the behavior of the agent more thoroughly.

Reason 5: Users are too engaged in the interaction to pay much attention to facial expressions. There is one big difference between the first and the second evaluation study. In the first evaluation, the subjects had to watch and listen to presentations given by the agents passively without any kind of interaction or disturbance. They could devote their full concentration to the agents and their verbal and non-verbal behaviors. In the GAMBLE setting, the subjects become an active part interacting with another human player and with the agent. Thus, the subjects' attention cannot be fully directed towards the agent. Instead, they have to evaluate the announcements made by the other players, they have to think about their next move based on this evaluation, and they have to operate the PDA interface which is simple, but nevertheless new to them. Standing face to face to the agent instead of just observing the head projected to the classroom wall leaves the subjects with a smaller facial display to interpret. Although we tried to compensate for this by zooming towards the agent when it announces its cast (see Fig. 2), the resulting size of the head does not nearly match the size in the face-only condition. The setting itself is more immersive because instead of seeing the agent's head projected to the wall while sitting in the classroom, the agent shares the interaction space with the human players (see Fig. 2). The full body agent is projected on the wall between the two human players thus creating a triangular face-to-face arrangement. Consequently, subjects are more involved into the interaction which leaves less capacity to concentrate on observing the agent's behavior. This immersion of the human players manifests itself on different levels: (i) The subjects reacted directly to the agent and its comments. If the agent e.g., said "I wanna see that!"[6] it happened that subjects showed their PDA to the agent or said something like "Yeah just take a look"[7]. (ii) When examining the subjects attentive behaviors, we observed that the subjects frequently looked at the agent (see above). According to Sidner and colleagues [24], this behavior may be interpreted as a sign of engagement with the interaction partner. The more often users look at the agent the more engaged they seem to be in the interaction. Up to now, we have analyzed the looking behaviors of the users during the agent's announcements. (iii) At the end of the experiment every subject had to fill in a short questionnaire to rate the interaction experience. Out of the 24 subjects nearly everybody thought it was enjoyable (24), funny (22), and interesting (23). Nobody rated it as boring, but only seven subjects found it amazing. A number of subjects also said it was drab (9) or monotonous (10). This seems to contrast with the overall rating of funny and enjoyable, but can be explained when looking at the comments provided by the participants. The small amount of utterances and their repetition were the most frequents comments (12) on drawbacks of the system which makes the agent appear a little bit drab or monotonous. Because the animations were dubbed beforehand, there was only a limited number of them available at runtime leading to repetitions during longer interactions.

Being more engaged in the interaction and thus having less capacity to interpret the facial expressions of the agent corresponds to Ekman's findings that people tend to disregard such facial clues to deceit in everyday life [9]. Ekman and colleagues [11] observed that subjects were generally not better than chance distinguishing honest from deceptive faces. Those that did better than chance pointed to facial clues as their decision aid. Thus, despite their visibility those clues were often not used.

5. CURRENT AND FUTURE WORK

The GAMBLE system was presented as a testbed to investigate the effects that deceiving ECAs have on the human-ECA interaction. Apart from user reactions to deceiving ECAs, GAMBLE also allows us to study the following aspects of human-agent interactions:

- Affective interactions: Prendinger et al. [22] have

[6]Das will ich sehn!
[7]Na dann schau mal

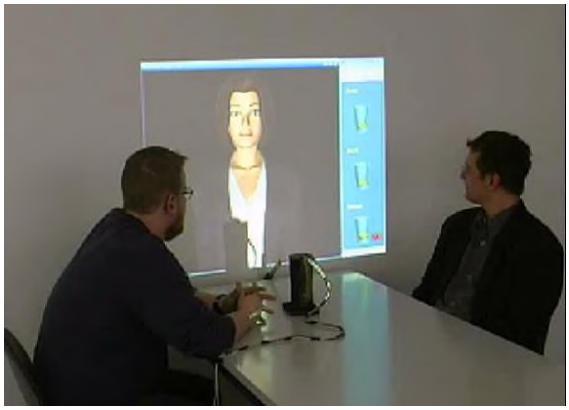

Figure 3: Improving the setting

shown how the display of emotional cues in an agent influences the user's appraisal of a situation. In GAMBLE, highly emotional situations are created, e.g. when the agent blames the user for deceit or when the user detects such an attempt and the agent has to react to it. Measuring the user's affective states by means of physiological sensors, we will investigate how different expressive behaviors of the agent exert an influence on these states.

- Engagement in multiparty interactions: According to Reeves and Nass [23] users tend to regard their social rules even in interacting with computers. And indeed, a number of studies of face to face interaction between an user and an ECA have confirmed that this tendency exists (e.g., [18]). Less is known about the effects that arise if multiple users interact with a single agent. In GAMBLE, the user's attention is divided between another human player and the agent. Thus, engaging the user in the interaction with the agent becomes less predictable.

The first evaluation which we presented in this article showed on the one hand that users were drawn into the game and engaged in the interaction with the agent. On the other hand, the comments in the questionnaires indicated some shortcomings of the system.

Although the setting was engaging to the users, standing in front of the wall to which the agent is projected felt a little bit awkward to the subjects. Now, the players sit at a table, the agent is projected to the wall at the head of the table, creating a much more natural atmosphere for such a game of dice (s. Fig. 3). The dubbed voice rendered the agent more engaging but had two crucial drawbacks. (i) As mentioned above, subjects seemed to pay too much attention to the voice quality, and (ii) the number of the agent's announcements was limited. At the moment, we are experimenting with different speech synthesizers which will make a context and situation dependent generation of dialogue acts possible. A thorough analysis of the users looking behavior over the whole time of the interaction revealed a conspicuous interest in the handheld devices that served as interfaces to the system. The gaze of the users was attracted by the PDA over 50% of the time. To get rid of the PDAs, a haptic interface was developed: the camcup which is a camera-mounted cup of dice that allows the users to just cast the dice in a natural way. The camera image enables the game server to keep track of the game process. The players announcements, i.e., result and belief statements, are captured by a speaker independent speech recognition system [12].

Little is know about the interaction dynamics, if an agent is confronted with more than one user. The video material collected during the first evaluation study is now employed to inform the design of a model of gaze behavior in such a multiparty scenario because apart from a study by Vertegaal et al. [27] no information is available on multiparty gaze behaviors. With an active gaze behavior of the agent, it will e.g. become possible to investigate gaze as another clue to deception in an agent.

6. CONCLUSION

In this paper, we presented an approach to the expression of emotions which considers the conflict between emotions an agent actually feels and emotions that it wishes to convey deliberately. Unlike earlier work, we did not start from the assumption that the agent is always able to conceal its "true" emotions perfectly, but simulated the occurrence of deceptive clues based on Ekman's studies of human facial displays. In addition, we presented two experiments we conducted in order to find out how deceptive clues are subjectively perceived by a human user and to what extent users are able to correctly interpret them.

Our first study indicates that even subtle expressions of deception may have an unfavorable impact on the user's perception of the agent - especially in situations where the user is expected to devote her full attention to the agent. Although people reacted to facial clues of deceit when they had the opportunity to carefully watch and compare different instances of agents, they were not able to name the reasons for these reactions (see Sec. 3). A designer of an interface agent should take such effects into account in order to prevent that unintended clues are conveyed by accident.

The results of the first study could, however, not be confirmed for the second scenario in which the experimental conditions were much less controlled. We have discussed a number of reasons why the users might have responded differently in the second experiment, such as the probability of the lies, the overestimation of other channels of expression and the distraction of the players by the game. In a more natural and engaging face-to-face situation, subjects tend to disregard deceptive clues which seems to be a natural phenomenon. Even in a domain where it is crucial to catch the other interaction partners lying, other communication features seem to be more important in the decision making process.

The second experiment also showed that it is hard to identify the clues users actually rely on. Clearly, we could have requested the user beforehand to pay special attention to the face as in the first experiment. The purpose of this experiment was, however, to investigate the impact of deceptive clues in a natural scenario in which the user may freely interact without being forced to concentrate on a specific channel (which would have also affected the entertaining value of the game). Obviously, people's expectations about an agent's abilities heavily influences their interpretation of the agent's behavior. The second experiment indicates that people tend to over interpret signals coming from the most sophisticated channel (even if they are obviously hard coded).

7. REFERENCES

[1] Timothy Bickmore and Justine Cassell. Relational agents: a model and implementation of building user trust. In *CHI '01*, pages 396–403. ACM Press, 2001.

[2] BLAG. Berliner Lexikon der Alltagsgesten. *http://www.ims.uni-stuttgart.de/projekte/nite/BLAG/*, last visited: 09.12.2004.

[3] Valeria Carofiglio, Fiorella de Rosis, and Cristiano Castelfranchi. Ascribing and weighting beliefs in deceptive information exchanges. In M. Bauer, P. J. Gmytrasiewicz, B. A. Juliano, R. S. Renner, and C. J. K. Tan, editors, *Computational Science — ICCS 2001*, pages 222–224, Berlin, 2001. Springer.

[4] C. Castelfranchi, R. Falcone, and F. de Rosis. Deceiving in golem: how to strategically pilfer help. In C. Castelfranchi, R. Falcone, B. S. Firozabadi, and Y. H. Tan, editors, *Autonomous Agents 98: Working notes of the workshop on Deception, Fraud, and Trust in Agent Societies*, 1998.

[5] Cristiano Castelfranchi and Isabella Poggi. Lying as pretending to give information. In Herman Parret, editor, *Pretending to communicate*, pages 276–291. de Gruyter, Berlin, New York, 1993.

[6] Fiorella de Rosis, Cristiano Castelfranchi, Valeria Carofiglio, and R. Grassano. Can computer deliberately deceive? a simulation tool and its application to turing's imitation game. *Computational Intelligence Journal*, 19(3):235–263, 2003.

[7] Fiorella de Rosis, Catherine Pelachaud, Isabella Poggi, Valeria Carofiglio, and Berardina De Carolis. From Greta's mind to her gace: modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, 59:81–118, 2003.

[8] Bella M. DePaulo, Deborah A. Kashy, Susan E. Kirkendol, Melissa M. Wyer, and Jennifer A. Epstein. Lying in Everyday Live. *Journal of Personality and Social Psychology*, 70(5):979–995, 1996.

[9] Paul Ekman. *Telling Lies — Clues to Deceit in the Marketplace, Politics, and Marriage*. Norton and Co. Ltd., New York, 3rd edition, 1992.

[10] Paul Ekman and Wallace V. Friesen. Felt, False, and Miserable Smiles. *Journal of Nonverbal Behavior*, 6(4):238–254, 1982.

[11] Paul Ekman, Wallace V. Friesen, and Maureen OSullivan. Smiles When Lying. *Journal of Personality and Social Psychology*, 54(3):414–420, 1988.

[12] G. A. Fink. Developing HMM-based recognizers with Esmeralda. In V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Lecture notes in artificial intelligence*, pages 229–234. Springer, Berlin, Heidelberg, 1999.

[13] B. Hartmann, M. Mancini, and C. Pelachaud. Formational parameters and adaptive prototype instantiation for mpeg-4 compliant gesture synthesis. In *CASA 2002*, pages 111–119, 2002.

[14] E. Krahmer, S. van Buuren, Zs. Ruttkay, and W. Wesselink. Audiovisual cues to personality: An experimental approach. In *Proc. of the AAMAS workshop on embodied agents as individuals*, Melbourne, Australia, 2003.

[15] M. Lee and Y. Wilks. Eliminating deception and mistaked belief to infer conversational implicature. In *IJCAI Workshop on "Cooperation, Collaboration and Conflict in Dialogue Systems"*, 1997.

[16] J. C. Lester, S. G. Towns, C. B. Callaway, J. L. Voerman, and P. J. FitzGerald. Deictic and emotive communication in animated pedagogical agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*. MIT Press, 2000.

[17] Frederic McKenzie, Mark Scerbo, Jean Catanzaro, and Mark Phillips. Noververbal indicators of malicious intent: affective components of interrogative virtual reality training. *International Journal of Human-Computer Studies*, 59:237–244, 2003.

[18] Yukiko I. Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. Towards a Model of Face-to-face Grounding. In *Proc. of the Association for Comp. Linguistics*, Sapporo, Japan, July 1–12 2003.

[19] Clifford Nass, Katherine . Isbister, and Eun-Ju Lee. Truth is beauty: researching embodied conversational agents. In J. Cassell, S. Prevost, J. Sullivan, and E. Churchill, editors, *Embodied conversational agents*, pages 374–402. MIT Press, 2000.

[20] C. Pelachaud and I. Poggi. Subtleties of facial expressions in embodied agents. *Journal of Visualization and Computer Animation*, 31:301–312, 2002.

[21] Helmut Prendinger and Mitsuru Ishizuka. Social Role Awareness in Animated Agents. In *Proceedings of Agents '01, Montreal, Canada*, pages 270–277, 2001.

[22] Helmut Prendinger, Sonja Mayer, Junichiro Mori, and Mitsuru Ishizuka. Persona effect revisited. using bio-signals to measure and reflect the impact of character-based interfaces. In *Intelligent Virtual Agents*, pages 283–291. Springer, Berlin, 2003.

[23] Byron Reeves and Clifford Nass. *The Media Equation — How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, Cambridge, 1996.

[24] Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. Where to look: a study of human-robot engagement. In *Proceedings of the 9th international conference on Intelligent user interface*, pages 78–84, 2004.

[25] M. Swerts, E. Krahmer, P. Barkhuysen, and L. van de Laar. Audiovisual cues to uncertainty. In *Proc. of the ISCA Workshop on Error Handling in Spoken Dialogue Systems*, Chateau-D'Oex, Switzerland, 2003.

[26] Susanne van Mulken, Elisabeth André, and Jochen Müller. An empirical study on the trustworthiness of life-like interface agents. In *Proceedings of the HCI International '99*, pages 152–156. Lawrence Erlbaum Associates, Inc., 1999.

[27] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. Eye Gaze Patterns in Conversations: There is More to Conversational Agents Than Meets the Eyes. In *Proceedings of SIGCHI 2001*, Seattle, WA, 2001.

[28] D. Ward and H. Hexmoor. Deception as a means for power among collaborative agents. In *Int. WS on Collaborative Agents: Autonomous Agents for Collaborative Environments*, pages 61–66, 2003.