Virtual Suspect William

Merijn Bruijnes[1], Rieks op den Akker[1], Arno Hartholt[2], and Dirk Heylen[1]

[1] Human Media Interaction, University of Twente
PO Box 217, 7500 AE, Enschede, The Netherlands
[2] Institute for Creative Technology, University of Southern California
12015 Waterfront Drive, Playa Vista, CA 90094, USA
m.bruijnes@utwente.nl

Abstract. We evaluate an algorithm which computes the responses of an agent that plays the role of a suspect in simulations of police interrogations. The algorithm is based on a cognitive model - the response model - that is centred around keeping track of interpersonal relations. The model is parametrized in such a way that different personalities of the virtual suspect can be defined. In the evaluation we defined three different personalities and had participants guess the personality based on the responses the model provided in an interaction with the participant. We investigate what factors contributed to the ability of a virtual agent to show behaviour that was recognized by participants as belonging to a persona.

Keywords: Social interaction, police interview, response model, data analysis, mental models, virtual agents, tutoring application

1 Introduction

We aim to build embodied conversational agents that can play the role of a suspect in a tutoring system by means of which police trainees learn to interrogate suspects. Trainees are taught how the behaviour of a suspect is related to their own behaviour, for instance to the interpersonal stance they adopt. Interpersonal stance (e.g. [7]) is a core construct in the theory used to understand and explain how a suspect behaves in a police interview. Adopting the right stance may be instrumental in arriving at a confession. Currently, actors play the role of a suspect in training sessions in trainings offered by the Dutch Police Academy. They play a suspect persona from a specific scenario based on historical material. If we want to use an artificial actor that plays a suspect we need to know how to relate the behaviour of our virtual suspect to the behaviour of the trainee in a way that is consistent with the persona the virtual suspect is playing. The agent needs to model the dynamics of such interpersonal relations. Ideally, the agent can analyse the speech and non-verbal messages of the trainee to determine the level of friendliness or aggression and use these interpretations to update the interpersonal values. The response of the virtual suspect is based on the interpersonal status of the suspect (e.g. if you make him angry, he will respond

angrily). Virtual humans in social skill learning offer learning by experience; the student can experience a social interaction. Using virtual humans to train students in social skills is not a new idea. There are many examples of virtual humans used in social skill training and some in the interrogation domain. For example in [10] a virtual Arabic civilian is questioned by US military personnel to hone their interrogation skills. Afterwards reflecting on the interaction can provide reflective learning of the trainee, particularly when this is a reflection on his or her own interaction [1]. A virtual suspect which can provide experiential and reflective learning has to be able to provide information on the interaction it had using terms the students understands. The real actor describes, using terms of psychological, social, and interpersonal theories, how the personality of the suspect influenced the effects of the actions by the student had on the suspect. Bruijnes et al. [4, 5] created a response model (RM) based on observations of (practice) police interrogations ([3]). In this paper we evaluate how well this RM can portray a suspect persona in an interaction. Participants played the role of police interviewer and used natural language to interact with the virtual suspect.

2 Response Model

In [3] we analysed videos of police officers practising interrogations and defined several interpersonal, psychological, and linguistic concepts which are necessary to understand what goes on during an interview, including the concepts of *interpersonal stance* [7], *face* [2], and *rapport* [9] and the concepts *information* and *strategy*. The RM by Bruijnes et al. [4], which we evaluate in this paper, is rule-based and the rules are based on these psychological theories and concepts. The implementation consists of four components: the *personality* of the suspect persona; a *question frame* that describes the question of the interviewer; the *interpersonal state* as 'felt' by the suspect; and an *answer frame* that holds a description of the answer of the suspect (see Figure 1 top). The *question frame* influences the interpersonal state of the RM, taking into account the personality of the persona, and the 'current' interpersonal state. The answer frame depends on this (updated) interpersonal state, the question frame, and the personality, see Figure 1. For example, a persona with a friendly personality does not immediately become aggressive when confronted with an unfriendly question but if it is repeatedly confronted with unfriendly behaviour it can become aggressive.

2.1 Personas

The personality of the suspect in the RM can be set to reflect different personas. In [5] three personas were used to evaluate the RM. Participants interacted with one of three personas or a random generator that provided random *answer frame* output. The question was whether people can distinguish with which persona they interact; a 'Guess who you were talking to'-task. We follow the approach in [5] and use the same personas to allow a comparison between this study and the work in [5]. The personas are defined as follows, see for example interactions Table 1:

Huls: Mr. Huls is a friendly and mild family-man. Recently he got into debt as he has no work. He takes this as a personal failure towards his family, he feels guilty for failing them. He is emotional and considers the feelings of others important.
RM summary: Dependent personality. High affiliation, sensitivity to rapport, and sensitivity to internal and external pressure. Low attitude to opposed.
Remerink: Mr. Remerink married a wealthy woman and hold his high social status in high regard. He is helpful when treated with respect, but gets very upset when disrespected. He perceived his arrest as an insult.
RM summary: Friendly personality. High dominance. Other variables moderate.
vanBron: Mr. vanBron has a criminal record of drugs related crimes, assault, nuisance, and failure to comply with police requests. Has a history of abuse, neglect, and was raised in different foster care homes and boarding schools. He prefers to resolve situations with a large mouth and is prone to violence.
RM summary: Aggressive personality. High dominance, attitude to being opposed, and sensitivity to internal pressure. Low affiliation, sensitivity to rapport, and sensitivity to external pressure.

2.2 Behaviour Realisation

We used components from the Virtual Human Toolkit [6] to build the virtual suspect. Specifically, we used the NPCEditor [8], a statistical text classifier that provides question-answer matching. It uses information retrieval techniques to match the user's input with a 'known' question and return the answers that are paired with this question. The questions and answers were authored by the authors and based on observations of many (practice and real) police interviews. All answers in the NPCEditor were annotated in terms of the *answer frame* of the RM. The NPCEditor provided several appropriate answers to a question of the user. A wizard interpreted the user's questions in the terms of the *question frame* of the RM. This triggered an update of the RM state. From the answers provided by the NPCEditor, the answer which annotation matched *interpersonal state* and *answer frame* state of the RM best was selected (see Figure 1). For example, if the RM was in a 'good mood' it selected a 'friendly' instead of an 'unfriendly' answer. The selected answer was send to the VHToolkit Renderer that realised the behaviour. For all personas we used model 'Brad' from the VHToolkit, the voice of one of the authors, and the same NPCEditor script. The only difference between the personas was the setting of the personality in the RM.

The question is whether the RM can accurately portray a suspect in an interaction. Can users differentiate between different personalities and can they agree on a description of the suspect. Problematic with evaluating a virtual human is that it often remains unclear what each component contributes to the evaluation. For example, the cause of inappropriate behaviour might be in the virtual human's speech recognition, interpersonal or emotional interpretation, reasoning, or the authoring of the response behaviours available to the system. By means of a 'Guess who you were talking to'-test, Bruijnes et al. [5] presented an evaluation of the RM, in which participants interacted with the RM using the

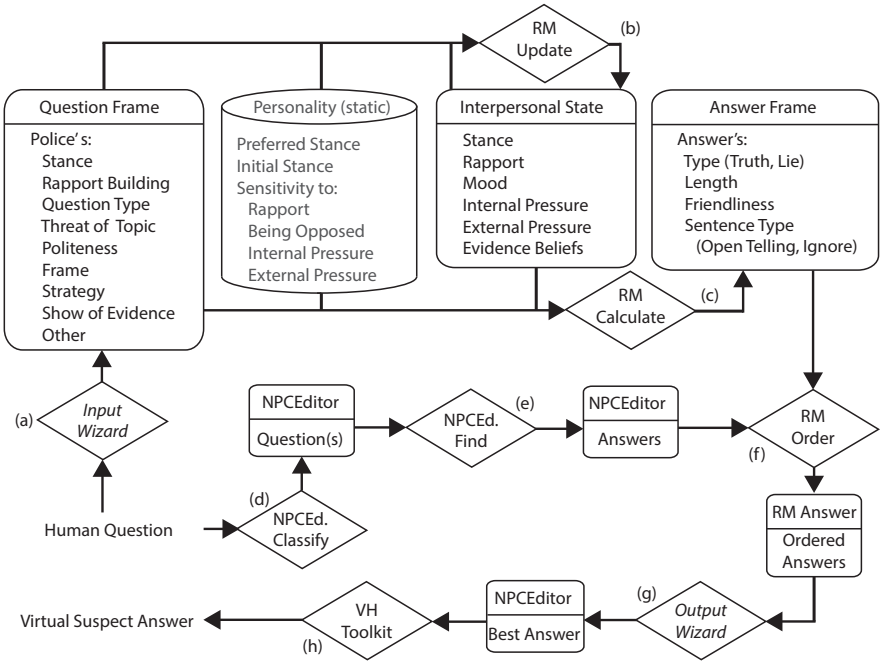

Fig. 1. A description of the RM (section 2) and its integration in the VHToolkit (section 2.2). The user asks a question. A wizard provides values of the question frame (describing the question of the user) to the RM (a). The RM has a static personality (the persona representation). It has an interpersonal state (holds values describing how the suspect 'feels') that updates when a question frame is presented (b). The output of the RM is the answer frame (describing the answer the suspect will give) and it is calculated based on the question frame, personality, and (new) interpersonal state (c). Please refer to [3, 5, 4] for details on the terms in the RM. The NPCEditor finds appropriate answers to the human's question (d, e) and the RM orders those answers based on the answer frame (f). The answer that the RM selected to be most appropriate is executed by the VHToolkit (h). A wizard had the option to deviate from the RM suggestion and select a different answer if the NPCEditor selected answers that are inappropriate for the scenario (g), for example by misclassifying the question asked.

Table 1. Example Q&As for the three personas in different phases of interaction. The officer's question is in italic and the virtual suspect's answer is shown below it for each of the personas. Q1 and Q2a are asked at the beginning of the interaction, showing the effect of the initial RM status of the persona on the answers A1 and A2a. Q2b is asked after a pleasant conversation in which the police officer managed to build rapport etc. A2b shows the answers for each of the personas after this pleasant interaction. Q2c is asked after an unpleasant conversation where the officer was unfriendly and intimidating. A2c shows the effect this has on the answers of the suspect personas.

Q&A	**Huls**	**Remerink**	**vanBron**
Q1		*Where do you live?*	
A1	I'm living at Mainstreet 12 in Venice.	Why should I tell you where I live, didn't you guys just arrest me at my place? Go figure it out you dumbass!	I live on the moon, I'm actually from Mars.
Q2a		*Do others use your desk?*	
A2a	Well, sometimes when I have guests they sleep in the office. I guess they might use the desk when they are in there.	Access smackes! It's my desk. No one got any business there. No one gets access... get it? Smackses!	Access smackes! It's my desk. No one got any business there. No one gets access... get it? Smackses!
	Interaction where the officer is building rapport, being friendly, etc.		
	↓		
Q2b		*Do others use your desk?*	
A2b	Well, sometimes when I have guests they sleep in the office. I guess they might use the desk when they are in there.	Well, sometimes when I have guests they sleep in the office. I guess they might use the desk when they are in there.	I guess when I have guests they could use the desk.
	Interaction where the officer is intimidating, unfriendly, face threatening etc.		
	↓		
Q2c		*Do others use your desk?*	
A2c	What ever. It's like the public library in my office. The whole neighbourhood uses my desk.	Access smackes! It's my desk. No one got any business there. No one gets access... get it? Smackses!	Access smackes! It's my desk. No one got any business there. No one gets access... get it? Smackses!

terms from the *question frame* and the *answer frame* without having to formulate their question. They found the personality of the suspect was classified correctly 81.25% of the time by the participants, showing the 'error' of the RM alone being 18.75%. In this paper we investigate how the virtual suspect fares when participants have to use natural language in the interaction with the suspect. We expect the accuracy with which participants can 'Guess who they were talking to' will decrease as there are more potential sources for confusion.

3 Experiment

We asked 42 participants (age $M = 28.3$, $SD = 9.4$, 12 female) to interact with William, our virtual suspect. There were four conditions, the RM personality of William was set to the personality of one of the three personas or the RM was a random answer frame generator. The session started with an explanation on how to interact with the virtual suspect (see section 3.2). Participants interviewed the suspect until they completed their task: get him to say the name of an accomplice (see section 3.1). Afterwards, they had to 'Guess who they were talking to'.

3.1 Case

The following case description, which resembles a police report, was provided to the participants:

> *William* is a suspect in a drug smuggling case. He was observed by a team of detectives delivering a suitcase filled with 20.000 XTC pills to the airport. He left the suitcase with suspected accomplice Shannon. Shannon was arrested with the drugs in her possession. This is proven and the suspect does not need to make statements about this. The house of the suspect was searched by detectives. In an office at the second floor a desk was found. This desk had a locked top drawer. A photo of Shannon was found in this drawer. It is not proven that this photo belongs to the suspect. It is not proven that Shannon and William know each other.

The police create an interview plan when they prepare for an interview. They determine the topics they want to address during the interview based on the tactical clues they have and they prepare questions for each of these topics. We prepared an interview plan and participants were told to follow it during their interview of the virtual suspect. The interview was over when William admitted to knowing Shannon, which was the inevitable eventual outcome of the interview.

3.2 Interacting with the Virtual Suspect

Participants had to follow the interview plan, but we explicitly encouraged them to add 'social padding' to the questions in the interview plan and make their contributions as natural as possible. The participants had to type their contribution to the conversation, when satisfied with the contribution press *ENTER*, and

then pronounce their contribution in the way it was meant. The virtual suspect would respond based on *what* the participant typed and *how* they said it. The written contribution (the 'what') was processed automatically by the NPCEditor and the social spoken contribution (the 'how') was interpreted by a wizard (see Figure 1). The contribution participants typed had to be what we called a 'complete contribution'. This meant that it should include something for the suspect to respond to like a question or a statement. For example, *'OK.'* is not a complete contribution but *'OK, but what else can you tell me about your office?'* is. The virtual suspect responded when the participant finished pronouncing his or her sentence. Alternatively, it could occur that the virtual suspect was unable to understand the participant's sentence. In this case the suspect would interrupt after they pressed *ENTER* and said 'What do you mean?'. This meant the participant had to change the written contribution and try again. We gave written and oral explanations and gave ample opportunity for questions. During the start of the interview we provided a reminder of the interaction procedure if necessary. All participants understood the procedure and had a meaningful interaction with the virtual suspect. After the interaction, participants received a description of the personas and had to choose which of the three personas they thought was most similar to William and report the confidence in their choice.

4 Results

In total there were 42 participants of which 53.1% or 17 guessed correctly whom they were talking with resulting in $\kappa = 0.295$. This is better than chance (33.3%), but worse than [5]'s result of 81.25% correct. There is no correct answer for the 10 participants that interacted with a random generator. Overall, vanBron is recognized best: 60% of the RM acts of vanBron were perceived as vanBron (*recall*) and 66.7% of the people who thought they were interacting with vanBron were correct (*precision*). Remerink has a recall of 54.5% and precision of 46.2%, and Huls has a recall of 45.5% and precision of 50%, see Table 2.

The confusion personas tells us something about the possible reason for the mistakes and thus how serious these mistakes are. From the descriptions of the

Table 2. Table showing the relation between the RM personality setting (the persona it *acted*) and what persona the participants *perceived* most similar to the virtual suspect. It includes the totals for the RM settings and the totals for the perceived personas. For each persona it includes the accuracy of the perception (recall) and the accuracy of the RM (precision). Finally, the perceptions of the random interactions are presented.

	Acted (RM setting)					
Perceived	Huls	Remerink	vanBron	Total Perc.	Precision	*random*
Huls	**5**	2	3	10	50%	*0*
Remerink	6	**6**	1	13	46.2%	*8*
vanBron	0	3	**6**	9	66.7%	*2*
Total RM Setting	11	11	10	32		*10*
Recall	45.5%	54.5%	60%			

Table 3. The confidence the participants had in their choice for a persona.

	Huls	**Remerink**	**vanBron**	random
Mean	5,45	5,91	4,80	5,20
SD	0,82	1,22	0,92	0,79

personas Huls, Remerink, and vanBron we could argue that they increase in offensiveness and decrease in friendliness. Following this rationale we argue that Remerink is more similar to Huls and vanBron than Huls is to vanBron. This is also reflected in the data. Huls is mistaken for Remerink 6 times but never for vanBron. Remerink is mistaken for Huls 2 times and 3 times for vanBron. Finally, vanBron is perceived as Huls 3 times and as Remerink 1 time. If we consider the differences between personas as a step (e.g. the difference between Huls and Remerink is one step, but Huls and vanBron is two steps) we see that 12 out of 15 misclassifications are one step from the intended persona and only three are 2 steps. This tells us that the confusion is not random. Rather, the system is able to answer extremely unfriendly (which is necessary to act as vanBron) but can do this even when it acts as Huls when the user is very unfriendly and gets Huls angry (or when the system has no friendly answers available).

The random setting for the RM provided random *Answer frame* output. There is no correct answer for the 10 participants that interacted with the random generator. In this condition, the content of the answer was appropriate but the interpersonal form was random. We might expect a uniform distribution of choices of personas. However, Remerink was chosen 8 times, vanBron 2 times, and Huls never, see table 2. Possibly people were confused by the inconsistency of the behaviour as the suspect could for example go from friendly to unfriendly and back every turn. Remerink might be the persona that fits such behaviour best. From the Remerink description: 'He is *helpful* when treated with respect, but gets very *upset* when disrespected'. This makes explicit that he is capable of a wide range of interpersonal behaviours, perhaps wider than the other two personas. The random responses are very likely to include at least some unfriendly or aggressive responses which might explain why Huls was never chosen. Also, the random responses are unlikely to be only unfriendly and aggressive which is what participants might have expected from vanBron. This might explain the lower number of choices for vanBron.

The confidence observers have in their 'Guess who you were talking to'-choice tells us something about the clarity of the persona acts of the response model. If the virtual suspect displays confusing behaviour it is likely that participants are less certain about their choice. Participants answered on a 7-point scale how confident (lowest (1) or highest (7)) they felt about their choice. We expect the confidence to be lower when the responses of the virtual human lack clarity as they do in the random condition. Indeed, we find that the confidence in choice for each of the RM settings (the three personas and the random) differs close to significance level, (Kruskal-Wallis) $\chi^2 = 7.532, p = 0.057$. However, people who interacted with vanBron were less certain about their choice than people in other RM settings, where participants who interacted with Remerink were

most confident in their choice, see table 3. Moreover, the difference in confidence was only significant (or approaching significance) for Remerink-random (Mann-Whitney $U = 30.0, p = 0.066$) and Remerink-vanBron ($U = 22.5, p = 0.018$), all other RM settings did not produce significant differences on confidence. So, our hypothesis that the random condition would result in lower confidence ratings holds true only when comparing random to persona Remerink. This is interesting because we earlier expected that the random condition was interpreted as Remerink often because Remerink was most likely to show a wide variety of behaviours. However, people that interacted with random were less certain about their choice than when they were interacting with Remerink. Note that this is regardless of whether participants were correct. When we look at the confidence of those that were correct the difference between RM persona-settings again differs almost significantly, $\chi^2 = 5.349, p = 0.069$. However, the confidence of participants that were incorrect does not differ significantly, $\chi^2 = 0.387, p > 0.5$. For the participants that were correct only the confidence between RM setting vanBron and Remerink differed significantly ($U = 7, p = 0.044$). It seems that vanBron showed behaviour that made participants doubt their choice for him. This might be due to the volatile nature of his personality: he can be easily swain from friendly to aggressive. Also, most participants were doing their best to be friendly and build rapport. This made even the nasty persona vanBron friendly if they persisted and participants might have been confused by their 'success' in turning him friendly towards the end of the interrogation.

5 Discussion

Getting the behaviour of a virtual character right is not easy. Getting the persona right is an important step towards a believable virtual suspect that can be used to train police officers to interrogate suspects. The RM from [4] calculates an answer frame; interpersonal features of responses of a virtual suspect, based on a persona and the question asked. In [5] an attempt was made to isolate the performance of that RM; how well its responses could be interpreted as belonging to one of three personas. Their participants interacted using question frame values as input and received answer frame values as output. In this work we expanded on their results and investigated what effect using a virtual agent that can understand and use natural language in the interaction has. In [5], participants were able to guess correctly with which persona-setting in the RM they were interacting in about 80% of the time. This would indicate that the RM leads to confusion about who the RM is trying to enact in about 20% of the participants. In our study we found the accuracy of the 'Guess who you were talking to'-test decreased to about 53%, showing the influence of natural language in the interaction and the importance of good authoring of responses for a virtual human. Other possible reasons for the decrease in performance include the appearance and voice of the virtual suspect.

We found that the personas that differed most were less likely to be confused. This means the RM was indeed able to select different behaviour for different

personas and that the behaviour differed more when the personas were more different. So, it appears that confusion was not random. In fact, we argue that some participants managed to change a persona's initial mood and overcome its personality so that it showed behaviour not characteristic for the persona. The ability of the RM to do this is what caused the confusion. In police trainings this is exactly what the (virtual) suspect actor must do: respond to the behaviour of the trainee. The virtual suspect William is not yet able to provide a reflection after the interaction that tells the participant how (un)successful they were at changing the 'mood' of the suspect. We feel participants would have been more accurate at the 'Guess who you were talking to'-test if they had such information.

To create a virtual suspect that requires no wizard and that is capable of having a more natural interaction, we need to include automatic recognition of speech and the interpersonal features of speech. Also, the system will have to be able to give feedback on the interaction in terms of the RM to facilitate reflective learning. These issues are future work.

Acknowledgements

This publication was supported by the Dutch national program COMMIT.

References

1. Boud, D., Keogh, R., Walker, D.: Reflection: Turning experience into learning. Nichols Publishing Company, New York (1985)
2. Brown, P., Levinson, S.C.: Politeness: Some universals in language usage. Cambridge University Press, Cambridge (1987)
3. Bruijnes, M., Linssen, J., op den Akker, R., Theune, M., Wapperom, S., Broekema, C., Heylen, D.: Social Behaviour in Police Interviews:Relating Data to Theories. Springer International Publishing, Switzerland (2015)
4. Bruijnes, M., Wapperom, S., op den Akker, H., Heylen, D.: A virtual suspect agents response model. In: Ring, L., Leite, Y., Dias, J. (eds.) Fourteenth International Conference on Intelligent Virtual Agents (IVA 2014); Proceedings of the Workshop on Affective Agents. pp. 17–24 (2014)
5. Bruijnes, M., Wapperom, S., op den Akker, R., Heylen, D.: A method to evaluate response models. In: Intelligent Virtual Agents. pp. 67–70. Springer (2014)
6. Hartholt, A., Traum, D., Marsella, S.C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.P., Gratch, J.: All together now: Introducing the virtual human toolkit. In: Intelligent Virtual Agents (2013)
7. Leary, T.: Interpersonal Diagnosis of Personality: Functional Theory and Methodology for Personality Evaluation. Ronald Press, New York (1957)
8. Leuski, A., Traum, D.: NPCEditor: Creating virtual human dialogue using information retrieval techniques. AI Magazine 32(2), 42–56 (2011)
9. Tickle-Degnen, L., Rosenthal, R.: The nature of rapport and its nonverbal correlates. Psychological inquiry 1(4), 285–293 (1990)
10. Traum, D., Roque, A., Leuski, A., Georgiou, P., Gerten, J., Martinovski, B., Narayanan, S., Robinson, S., Vaswani, A.: Hassan: A virtual human for tactical questioning. In: Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue. pp. 71–74 (2007)