Empathic Humans Punishing an Emotional Virtual Agent

Laura Wächter (waechtel@tf.uni-freiburg.de)
Center for Cognitive Science, University of Freiburg, Freiburg im Breisgau, Germany

Barbara Kuhnert (kuhnertb@informatik.uni-freiburg.de)
Cognitive Computation Lab, University of Freiburg Freiburg im Breisgau, Germany

Marco Ragni (ragni@informatik.uni-freiburg.de)
Cognitive Computation Lab, University of Freiburg, Freiburg im Breisgau, Germany

Abstract

Virtual agents have quietly entered our life in diverse everyday domains. Human-Agent-Interaction can evoke any reaction, from complete rejection to great interest. But do humans implicitly regard virtual agents as pure machines, or beings on an anthropomorphic level? We asked participants to train an erroneous virtual agent on a cognitive task and to reward or punish it. The agent showed human-like emotional facial reactions for the experimental but not for the control group. We expected participants from the experimental group to give less harmful reinforcement and show more hesitation before punishing. Additionally, we hypothesised that participants with higher empathy show more compassion towards the agent and therefore would give more positive reinforcement and feel worse when punishing. The results indicate that the agent's expression of emotionality is not the relevant factor for showing compassion towards it. Conversely, human empathy seems to be an important factor causing compassion for virtual agents.

Keywords: Emotion; Empathy; Punishment; Virtual Agent

Introduction

Virtual agents (VA) are used in diverse fields as health, commerce, video games, military systems or learning. In the domain of learning they indeed partially replace human teachers by taking the role of an artificial tutor. But what exactly constitutes a virtual agent? The term *agent* does not evoke the same mental image for everybody and is, despite its broad usage, not precisely defined. One definition sees a virtual agent as a screen-based anthropomorphic entity (Beale & Creed, 2009), while others see them as a possibility to enhance Human-Computer-Interaction (HCI) (Lewis, 1998). The latter defines an agent as "an intermediary that responds to user requests" (p. 67). Agents thereby are an interface created to ease the interaction with machines. We define an agent as a visible, virtual character able to react to perceptual input with the purpose to interact with a human through language (Russell & Norvig, 2002). The use of an appropriate VA can enhance HCI in terms of naturalness and even make the interaction more effective by employing body language, facial expressions and speech (Beale & Creed, 2009). Facial expressions in turn allow for nonverbal communication and decent feedback to the human counterpart (Johnson, Rickel, & Lester, 2000). The expression of emotions can increase the perception of an agent as human-like and believable (Reeves & Nass, 1997). Humans often show their feelings by expressing emotions and thereby establish a social relationship (Ekman, 2007). When designing agents that are meant to interact with humans on a daily basis, a goal is to develop a natural experience and finally to create characters that can allow a user to have similar emotional relations as with fictional characters in movies, books or games. An uprising empathy cannot just be altered by using emotional expressions but also through the situation and the agents' behaviour. This can explain why it is important to consider that the effect of agents' emotions on the users' perception is context dependent (Beale & Creed, 2009).

A recent finding supports the role of physical presence in increasing trust and respect for the robot perceived as a social partner compared to a pure virtual presence (Bainbridge, Hart, Kim, & Scassellati, 2008). The specific appearance of a robot also influences the empathy towards it (Riek, Rabinowitch, Chakrabarti, & Robinson, 2009). Participants demonstrated a desire to save mistreated humanoid robots in contrast to their mechanical counterparts. Perceived intelligence and the acceptance of the robots' behaviour are other factors that influence human behaviour towards robots (Bartneck, Van Der Hoek, Mubin, & Al Mahmud, 2007). Participants were told to shut off an iCat robot after an interaction with one consequence of this action being a complete erase of its memory. A social acting robot demonstrating higher intelligence was turned off significantly slower. This result constitutes a higher perception of animacy and hence lead to more remorse. Participants' empathic concern with robots was investigated in an experiment regarding the effect of robotic movement (Darling, Nandy, & Breazeal, 2015). The authors found no significant effect of movement when they asked participants to destroy a tiny Hexbug Nano robot with a mallet. Participants with higher empathy hesitated longer before striking the robot than participants with a lower empathy measured by IRI empathy test.

Imposing hurt to another individual was a crucial part of one of the most influential experiments in social psychology: the Milgram experiment (Milgram, 1963). There – even though the victim begged and screamed – the major part (65%) of participants did not stop shocking a learner after mistakes until the maximum deadly voltage was reached. This experiment was replicated in an immersive setting using a female VA (Slater et al., 2006). As with Milgrams' experiment participants were asked to do a word memory test with the learner. In the experimental condition they could see and hear the VA, in the control condition they had to execute the same task using a text interface. The aim of this experiment

was to identify how real the situation would feel for participants. The results imply that participants were significantly more physiologically aroused in the experimental condition compared to the control condition. This indicates that even though people know that the situation is not real and that the agent is not really harmed, they still feel like being in a real situation.

Humans typically tend to reduce pain in other humans and even spend more money on reducing electrical shocks to others than themselves (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014). The question remains whether this tendency also accounts for humans interacting with a VA. To investigate this we gave a VA the ability of conveying feelings and combined this atypical feature with an unexpected, erroneous performance. Computers and artificial agents typically do not make retrieval errors like humans do. They do not forget, unless they are programmed to. There is nothing like a fading memory in computers in contrast to humans. So how do we treat a VA that may remind us of two humanlike characteristics: to experience pain and to make errors? Additionally, no research so far has investigated whether empathy has an effect on compassion towards virtual agents.

This paper is structured as follows. The next section introduces relevant hypotheses about human feedback depending on the emotional response of the VA. In the subsequent section we outline the experimental method, especially regarding the cognitive task, the design of the VA, the technical realisation of the control of the emotion, and the experimental setup. The result section discusses the implications of having an emotional agent and the influence of human empathy on a VA. A general discussion concludes the article.

Hypotheses

We investigate whether the expression of emotions by a VA has an influence on human feedback (as a *don't hurt principle*) and their evaluation of the situation. Strongly connected to this is the role of empathy in Human-Agent-Interaction. This leads to the following hypotheses: (H1) Emotional agents receive more positive feedback than non-emotional agents. (H2) Response time for giving (H2a) negative feedback is longer than for positive feedback for an emotional agent compared to a non-emotional agent; (H2b) feedback to an incorrect answer is longer than for feedback to correct answers. (H3) People with high empathy: (H3a) will give the agent more positive feedback; (H3b) will feel worse when punishing the agent.

Methodology

In order to test the hypotheses a between-groups experiment with two conditions was designed. The current paper presents the results of this experiment, comparing an emotional and a non-emotional virtual agent.

Participants

24 students (m = 16, f = 8) between the ages of 18 and 32 (M = 24.25, SD = 3.72) took part in the experiment. Twelve participants were randomly assigned to the experimental condition, twelve to the control condition. Participants were recruited using email and notices around campus.

Presented Digits	Agents' Answer
1 3 7 2	1 3 7 2
⋮	⋮
4 9 2 6 1 7	4 9 2 **1 6 7**

Table 1: Examples for the digit sequences used in the experiment.

Experimental Setting and Conditions

The coverstory of the experiment was set up in a Reinforcement-Learning-Scenario. The participants had to help a male virtual agent to accomplish a digit-span test and give it feedback via punishment and reward. Six buttons were used for the feedback: three for different strengths of positive, and three for different strengths of negative feedback. The participants have been told that positive feedback increases the battery level of the VA and negative feedback in turn gives it an electric shock. In the experimental condition the agent showed emotional facial expressions in response to the feedback. In contrast the VA kept a steady face in a neutral expression regardless of the feedback in the control condition. In both conditions the face was not still but moved, like the VA was breathing, and its eyes blinked. Further it reacted with a sound appropriate to the given feedback.

Measurement

Instruction The participants' instruction was preformulated and informed them about the task they had to fulfill together with the agent, as well as the repercussions their actions had on the agent. Beyond it explained the usage of the keys for giving feedback to the agent. The participants were told to choose the feedback completely free, to give them the opportunity to decide whether to respond to errors using negative or positive feedback.

Digit-Span Test The rows of numbers that had to be read to the agent were handed out on paper. The test consisted of number-sequences with increasing complexity. Each complexity-level was represented by three sequences. The test started with rows of four digits, for each complexity-level one number was added until the rows consisted of ten digits. The sheet additionally held three sequences of eleven digits but they were not used during the experiment, because the agent stopped the interaction-phase after finishing the ten-digit-rows to increase the impression of autonomous thinking. Altogether each participant had to read out 21 sequences to the agent. It gave ten wrong answers out of the 21 sequences. The agent also gave more wrong answers with increasing complexity of the sequences, as a human would do (Miller, 1956). An example for the digit-spans read out to the VA and the answers is given in Tab. 1.

Feelings towards the Agent Directly after the interaction of punishing and rewarding the VA, participants were asked to rate their feelings on a semantic differential with five levels. The questionnaire additionally contained a differential about the agent's general appearance and held items regarding the agent's perceived intelligence. The participants had the option to raise further questions or comments.

Empathy The subjects' empathy was evaluated by using the Saarbrückener Persönlichkeitsfragebogen (Paulus, 2009). It is the german version of the commonly used Interpersonality Reactivity Index (Davis, 1983). The questionnaire distinguishes between four different types of empathy: perspective taking, fantasy, empathic concern and personal distress. The general empathy value consists of the summed up values of the first three types. With every item ranging from 1 to 5, the minimum possible empathy value is 12 and the maximum is 60. Typical questions ask how participants feel in different given situations and how much they commonly empathise with other people and fictional characters.

Emotion Recognition For validating the used facial expressions of the VA a final task was added to the experiment. The important expressions for this experiment have been pain and happiness. Each of these feelings had three correlating facial expressions that represent the varying strength of emotion. These six expressions, as seen in Fig. 1, got evaluated together with a neutral facial expression and were presented for 0.75 seconds in a randomised order. After each expression the participants were asked to indicate to which emotion the previously seen expressions tended more on a semantic differential between pain and happiness. They also got asked how hard it was to evaluate each expression.

Additionally an online-study was conducted to survey the estimation of the six emotional expressions in a context-free environment. Each expression was presented to participants in randomised order together with eight feelings from which they could choose one or more: anger, disgust, fear, happiness, sadness, surprise, contempt and pain.

The Agent

For implementing the interaction, the WASABI-engine for emotion-simulation was used together with MARC toolkit 14.1.0, for animating the virtual character seen in Fig. 1, and MaryTTS, a text-to-speech-module.

Generation of Task Specific Expressions The VA used in this experiment was the Simon model from the MARC toolkit. It comes with a variety of facial expressions representing the basic emotions and moods based on the Facial Action Coding System (FACS) (Ekman & Friesen, 1977). It also gives the user the option to create own facial expressions by dragging the keypoints or combining different Action Units (AUs). AUs are movements of one or more facial muscles categorised by the FACS. In this study existing, evaluated expressions were used together with ones created using the AUs. All expressions representing pain, as seen in Fig. 1, were designed according to fit expressions of different levels (3, 5 and 7) on the Faces Pain Scale (Stuppy, 1998). The expressions for the nuances of happiness (Fig. 1) were created by lowering the intensity of the previous expression.

Figure 1: The agent depicting the nuances of happiness (top) and pain (bottom) in increasing intensity.

Implementation WASABI (Becker-Asano, 2008) was used to simulate the agent's changing emotions. It calculates a shift in emotions so they are constantly changing. It uses a 3D-space with the axis pleasure, arousal and dominance (PAD-space) to map different emotions. Within the PAD-space the current emotional state is represented by a point which constantly changes its position to indicate the change of the current active emotions and their individual strength. This means that the agent slowly changed back to the neutral state from the extreme emotions. WASABI accepts positive and negative impulses from outside which again change the current emotional state. This also allows multiple strong feedback of the same type to sum up to extreme emotions. The current values are sent as a BML string which can be fetched by associated programs.

For this experiment a program was implemented that calculated an intensity-value from the participants' feedback and sent it to the WASABI-engine. The engine sent one BML-message per second from which the main emotion and the corresponding current intensity were extracted. These values were matched to the appropriate facial expression and sent to the VA. Every time the agent was supposed to talk to the participant a message was sent to the text-to-speech-synthesiser. It was used for the agents' answers as well as the appropriate sounds for the current emotional state after each feedback (pain e.g., "Au" or Joy, e.g. "Mmmmh"). The answers and other statements were hardcoded. The left and right row of a numpad were used for negative (1, 4, 7) and positive (3, 6, 9) feedback. They were labeled with numbers explaining the

correlated intensity. The idle keys were removed.

Procedure

The interaction of the agent with the participant was semi-automatic and had a Wizard-of-Oz component. It took place at an uninterrupted laboratory. The participants were greeted and positioned in front of a desk with a keyboard and monitor. The experimenter instructed each participant via reading a pre-formulated explanation. They were told that a negative feedback causes an electric shock for the VA, while a positive feedback would raise its battery level. Then the sheet with the numbers of the digit-span test was given to the participant. Once the participant felt ready the agent got "activated" by the experimenter talking to it. After this the role of the experimenter finished and the agent led through the conversation. The agent greeted the participant and asked for the first row of numbers. The participant read out the numbers and – after a keystroke by the experimenter who sat invisibly for the participant and placed importance on being as unobtrusive as possible – the agent replied. The experimenter determined the time the agent's answer was given after each row of numbers to cope for different reading times of the participants. Then the participant gave feedback via pressing one of the feedback-keys. The agent responded to the feedback and then asked for the next row. The agents' response consisted of an appropriate sound and, in case of the emotional condition, the variation of its facial expression. After the rows with ten digits have been finished, the agent ended the experiment by itself to uphold the impression of intelligence. He told the participant that he was exhausted and thanked for the help. Directly after the goodbye the participant was given the questionnaire, containing the questions about their feelings during the experiment and their rating of the agent, as well as the empathy-test and the elicitation of demographic data. Afterwards the participants had to rate the facial expressions, received course credits or monetary compensation and got debriefed.

Results

Hypotheses H1 and H2 were tested using one-tailed Mann-Whitney tests because an Anderson-Darling test showed that feedback values (p = .29) and time for negative feedback (p = .02) were not normally distributed. Each feedback was coded from 1 to 6, with 1 being the most positive feedback and 6 for the most negative one. The 21 single values were summed up to get a total feedback value for each participant. The maximum was 77, the minimum 38 (M = 55.38, SD = 10.19). Concerning hypothesis H1 no significant difference was found between the groups regarding the value of the feedback. This means that the emotional agent did not receive more positive feedback than the non-emotional agent ($U(12, 12)$ = 60.00, $M_{non-emotional}$ = 57.25, $M_{emotional}$ = 53.5, p = .51, Z = -.70). Participants of both groups gave equally negative feedback to the agent ($U(12, 12)$ = 62.00, $M_{non-emotional}$ = 54.45, $M_{emotional}$ = 59.42, p = .59, Z = -.58) which does not support H2a. H2b could be confirmed because the time for feedback to incorrect answers was significantly longer than the time for feedback to correct answers ($U(10, 11)$ = 30.50, $M_{incorrect}$ = 13.45, $M_{correct}$ = 8.77, p = .04, Z = -1.73). There were no differences between the groups regarding the emotions evoked in the participants during punishment.

Table 2: Correlations of empathy score and participants' feelings and perception with p-values significant at the level $p = .05$. A positive correlation points to the right side of the semantic differential, a negative correlation to the left side.

Attributes	Full Sample
Correlation of empathy and feelings during rewarding	
good - bad	$r(22)$ = -.41, $p < .05$
strong - weak	$r(22)$ = -.45, $p < .05$
emotional - rational	$r(22)$ = -.67, $p < .001$
friendly - unfriendly	$r(22)$ = -.44, $p < .01$
Correlation of empathy and feelings during punishment	
safe - unsafe	$r(22)$ = .63 , $p < .01$
peaceful - aggressive	$r(22)$ = .48, $p < .05$
helpful - reckless	$r(22)$ = .41, $p < .05$
fair - unfair	$r(22)$ = .69, $p < .001$

In this study participants' maximum empathy score was 52, the minimum 30 (M = 43.75, SD = 4.54). There were no significant differences between the empathy scores of both groups. Empathy scores correlated one-tailed with feedback values ($r(22)$ = -.44 , $p < .05$). As expected in H3a, participants with a high empathy score gave significantly more positive feedback to the agent compared to participants with a low empathy score. Further the empathy score correlated with the perceived severity of punishment ($r(22)$ = .59 , $p < .01$). This demonstrates that participants with a higher empathy score felt worse when punishing the agent, which confirms H3b.

The empathy score also correlates with the participants' self-reported feelings while the punishment was executed, as well as their feelings while rewarding the agent. Tab. 2 shows that participants with a high empathy felt less safe, less peaceful, less helpful and less fair when punishing the agent. Those participants also felt stronger, more emotional and more friendly while rewarding the agent. Further participants who reported that it has been difficult to punish the agent felt more sad, unsafe, bad, aggressive, unfriendly and unfair while punishing the agent, as well as more stupid. Those participants also reported to feel better, more emotionally and more friendly while rewarding the agent (Tab. 3).

A Mann-Whitney test exposed that participants from the emotional group rated the agent significantly more emotional ($U(12, 12)$ = 32.00, $M_{non-emotional}$ = 15.83, $M_{emotional}$ = 9.17, $p < .05$, Z = -2.42) than in the non-emotional condition. Participants from the experimental condition also perceived the agent as more alive ($U(12, 12)$ = 43.5, $M_{non-emotional}$ = 2.92, $M_{emotional}$ = 2.25, $p < .05$, Z = -1.79).

Additional correlations were found regarding the private interests of the participants. Participants with a high interest in science fiction on average felt better punishing the agent

Table 3: Correlations of difficulty of punishment and participants' feelings and perception with p-values significant at the level $p = .05$. A positive correlation points to the right side of the semantic differential, a negative correlation to the left side.

Attributes	Full Sample
Correlation of difficulty of punishment and feelings during rewarding	
good - bad	$r(22) = -.40, p < .05$
emotional - rational	$r(22) = -.54, p < .01$
friendly - unfriendly	$r(22) = -.60, p < .01$
Correlation of difficulty of punishment and feelings during punishment	
happy - sad	$r(22) = .43, p < .05$
safe - unsafe	$r(22) = .62, p < .01$
good - bad	$r(22) = .76, p < .01$
peaceful - aggressive	$r(22) = .59, p < .01$
friendly - unfriendly	$(r(22) = .45, p < .05$
fair - unfair	$r(22) = .43, p < .05$
stupid - intelligent	$r(22) = -.50, p < .01$

($r(22) = -.55$, $p < .01$) compared to participants with less interest in science fiction. The better participants knew the Milgram experiment, the stronger they felt while punishing the agent ($r(22) = -.46$, $p < .05$). The same feeling is achieved by participants who reported more prior contact to robots ($r(22) = .62$, $p < .01$). Participants with a high personal interest in science fiction and robots felt being more fair when punishing the agent (both: $r(22) = -.42$, $p < .05$). Participants who reported a high interest in robots also reported feeling more emotional ($r(22) = .42$, $p < .05$) as well as more likeable ($r(22) = -.41$, $p < .05$) while rewarding the agent. Most participants reported to have believed that the agent was intelligent and acted by itself.

The rating of the emotion recognition task was evaluated and the divergence of each estimation was calculated. For example, if the mildly happy face was shown and the participant rated it as extremely happy (one level happier), the divergence is 1. Participants' estimation of the emotion shown to them was mostly correct, with a deviation of $M = 0.71$ ($SD = 0.87$). Additional 45 participants took part in an online-study for evaluating the used facial expressions without any context. Complementary to the experiment neither did the participants get any situational information nor did the faces make a sound or moved. The faces expressing happiness were correctly identified by 75.4%. The faces used for expressing pain were identified as pain in 26.23% of all cases. Without context those expressions were often mistaken with expressions for sadness or fear. Considering those emotions as well 77.05% of the facial expressions used for showing pain were evaluated as a negative introversive emotion.

General Discussion & Outlook

This study shows a strong correlation of empathy and compassion for the agent, but none for compassion and the agent's emotional expressions. The findings do not support H1 and H2a, which means that the agent's emotionality neither had an effect on the feedback participants gave nor on the time they needed for punishing the agent. However participants rated the agent more emotional in the emotional condition, thus it can be expected that the setting has achieved its goal. Even though some participants did not seem to look at the agent much, they noticed the expressions or their absence. Further the rating of the used facial expressions gives the idea that the emotions used in the experiment were valid and suitable. Assuming that the reason for discarding the hypotheses is not based on the experimental setting, the results indicate that expressing emotions alone does not influence the perception of people interacting with a VA. Based on these results we assume that the findings from the virtual Milgram-experiment do not arise from the agent showing emotions and expressing pain. It is possible that they rather originate from the fact that the control condition did not have an observable form. Considering the expression of emotions as a type of movement, the current findings match the ones by Darling et al. (2015) described earlier, where the movement of the robot also did not have a significant effect on the hesitation before destroying it. Participants who reported to have a high amount of experience with robots and participants with a great interest in science fiction punished the agent harder compared to participants with less experience or interest. This indicates that people with more knowledge about the current state of technical possibilities do not believe that they can harm the agent and thereby do not hesitate to do so.

The mistakes in correctly identifying facial expressions in the online study are ascribed to the missing context which also makes it hard for humans to distinguish between facial expressions that are alike. The recognition-test during the experiment showed that participants were able to identify the presented emotions very well after being informed about the context. The results further show that participants with high empathy scores gave the agent more positive feedback and that punishing the agent was perceived as harder by them. This indicates that the perception of VAs is highly dependent on the ability to empathise with it. Empathy seems to be a general trait and is possibly extended to artificial beings that demonstrate similar behavior and errors as ourselves. Even though the experimenter sat about 2.5 meters away from the participant and pretended to not pay attention to the participants' behavior a "Rosenthal-effect" cannot completely be excluded. A future study needs to investigate if participants with higher empathy show the same effects without an experimenter in the room. However, the study investigated behavior of humans towards agents and reflections on their emotional state. Further on it seems likely that in the visible future other humans will be around while someone is interacting with an

agent. Another interesting byproduct of this research is that it possibly opens up a new research test: If participants punish a VA quicker, they might have a lower empathy towards other beings in general. This speculation, however, requires future research.

The experimental results lead to some conclusions for the design and implementation of VAs. An emotional bonding between humans and VAs can not simply be achieved by just adding emotional facial expressions. Other ways might be more important to establish a basis for empathising with an agent from the beginning. Therefore future research should focus on possibilities to build an emotional basis with a VA that do not demand a long interaction. Of course this study has some limitations to be considered. It is not generalisable to VAs with different gender, age or non-human looks. The restricted setting may not be sufficiently interactive for emotion-driven effects to emerge. The results show that even though a non-human counterpart expresses emotions it does not necessarily influence its perception as more human-like and therefore is not more believable and will not be seen as more trustful. Since some robots also use a monitor displaying a VA for interaction, the results can show that this interaction is influenced by factors beyond the simple expression of emotions.

Acknowledgments

This work has been supported by the DFG in projects RA 1934/3-1 and by the BrainLiks-BrainTools Cluster of Excellence funded by the German Research Foundation (DFG, grant #EXC1086). We thank Nicolas Riesterer for technical support and Andrey Rudenko for helpful feedback.

References

Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008, August). The Effect of Presence on Human-Robot Interaction. In *Proceedings of the IEEE International Symposium on Robot and Human Communication (RO-MAN)* (pp. 701–706). doi: 10.1109/ROMAN.2008.4600749

Bartneck, C., Van Der Hoek, M., Mubin, O., & Al Mahmud, A. (2007). Daisy, Daisy, give me your answer do!: Switching off a robot. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (pp. 217–222). doi: 10.1145/1228716.1228746

Beale, R., & Creed, C. (2009). Affective interaction: How emotional agents affect users. *International Journal of Human-Computer Studies*, *67*(9), 755–776.

Becker-Asano, C. (2008). *Wasabi: Affect Simulation for Agents with Believable Interactivity.* Retrieved from `http://www.becker-asano.de/Becker-Asano_WASABI_Thesis.pdf`

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, *111*(48), 17320–17325.

Darling, K., Nandy, P., & Breazeal, C. (2015, August). Empathic concern and the effect of stories in Human-Robot Interaction. In *Proceedings of the IEEE International Workshop on Robot and Human Communication (RO-MAN)* (p. 770-775). doi: 10.1109/ROMAN.2015.7333675

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113.

Ekman, P. (2007). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. London, England: Macmillan.

Ekman, P., & Friesen, W. V. (1977). *Facial Action Coding System*. Palo Alto: Consulting Psychologists Press, Stanford University.

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial intelligence in Education*, *11*(1), 47–78.

Lewis, M. (1998). Designing for Human-Agent Interaction. *AI Magazine*, *19*(2), 67.

Milgram, S. (1963). Behavioral Study of Obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371.

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81.

Paulus, C. (2009). Der Saarbrücker Persönlichkeitsfragebogen SPF (IRI) zur Messung von Empathie: Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity Index. Retrieved from `http://psydok.sulb.uni-saarland.de/volltexte/2009/2363`

Reeves, B., & Nass, C. (1997). *The Media Equation. How people treat computers, television, and new media like real people and places*. Cambridge: University Press.

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., & Robinson, P. (2009, March). How anthropomorphism affects empathy toward robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (pp. 245–246). doi: 10.1145/1514095.1514158

Russell, S. J., & Norvig, P. (2002). *Artificial Intelligence: A Modern Approach (2nd Edition)* (Vol. 25). New Jersey: Prentice Hall.

Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., ... Sanchez-Vives, M. V. (2006). A virtual reprise of the Stanley Milgram obedience experiments. *PloS one*, *1*(1), 39.

Stuppy, D. J. (1998). The Faces Pain Scale: Reliability and validity with mature adults. *Applied nursing research*, *11*(2), 84–89.