The Relationship between Human Values and the Ethical Design and Acceptability of Relational Agents

Ravi Vythilingam, Deborah Richards and Paul Formosa, Macquarie University, Australia

ABSTRACT

Relational agents are artificially intelligent (AI) virtual characters that seek to support humans by playing roles typically played by humans. Research has focussed on the believability and utility of relational agents, but little research has been conducted to explore the ethical acceptability of using AI technology in these roles. We have created four scenarios (three text-based and one using a relational agent), each of which is designed to explore a relational agent in different relational roles and contexts that encompass the five AI4People ethical principles (beneficence, non-maleficence, autonomy, justice, and explicability), to capture participants' agreement with the scenarios. To model whether participants' responses are related to their values, we capture participants' values using Schwartz's Theory of Basic Human Values. The motivation and design of our study and preliminary results are presented in this paper.

Keywords: Intelligent virtual agents, relational agents, ethical acceptability, Schwartz's values

1. Introduction

We are increasingly seeing the use of relational agents, a subset of IVAs (Intelligent Virtual Agents) which are also in the same family of tools as avatars and embodied agents, performing roles such as counselling and advising (Gratch, Lucas, King, & Morency, 2014; Morency et al., 2015; Stratou & Morency, 2017) and coaching (Beun et al., 2016; Shamekhi & Bickmore, 2015; Watson, Bickmore, Cange, Kulshreshtha, & Kvedar, 2012). As with other Artificial Intelligence

(AI) applications, they can outperform humans in some circumstances (Lisetti, 2012; Lucas et al., 2017). However, the ethical acceptability of this technology has received insufficient attention. To address this gap, we aim to explore the relationship between individual ethical values, by drawing on Schwartz's (2012) theory of basic human values, and individual views about the acceptability of using AI relational agents across a range of scenarios. We draw on the AI4People framework (Floridi et al., 2018) of five ethical AI principles of beneficence, non-maleficence, autonomy, justice, and explicability to design the scenarios. The intended outcome is to understand how human values relate to the ethical acceptability of the use of relational agents in specific contexts.

2. Relational Agents

Relational agents as defined by Bickmore (2019) are "computational artefacts designed to build and maintain long-term, social-emotional relationships with their users". Relational agent applications have grown rapidly powered by the recent progress in artificial intelligence. Typical applications centre around traditionally human-to-human interactions in the realm of healthcare, education, coaching and counselling, eldercare, childcare, personal relationships, and personal assistants. The use of these intelligent relational agents to make decisions in roles that are usually played by humans raise various ethical considerations, especially as many of these decisions involve members of vulnerable groups, such as the elderly or children, and have significant impacts on human well-being and functioning.

Hudlicka (2016) identifies several ethical issues that go beyond the general concerns of data privacy and which are specific to virtual agents. These concerns include affective privacy (the right to keep your thoughts and emotions to yourself), emotion induction (changing how someone feels), and virtual relationships (where the human enters a relationship with the agent). Additionally, a recent literature review and analysis (Hussain, Marc, Raymond, & Timm, 2019) of 90 pieces of research on how humans interact with avatars and embodied agents identified six design factors that should be taken into account in the design and implementation of avatars and embodied agents. These are: (1) the *Proteus effect* where the agent can have an unintended influence either positive or negative on the user; (2) the *Uncanny valley effect* where if the agent looks very closely like a human but not fully so, it can put the user off using the agent; (3) stimulating the concept of *presence* to make the interaction more social can improve effectiveness of the interaction; (4) a more *persuasive* design of the agent can have greater influence on the user towards a certain direction; (5) *empathic* features in the design of the agent encourages more positive interaction; and (6) *customisability* options of the agent when offered can increase attachment with the agent. How these factors are designed and implemented in the agent has strong ethical implications and impacts on the user. To help address and operationalise context appropriate ethical guidelines for the use of relational agents, processes need to be implemented that can capture personal and societal values of the relevant stakeholders and incorporate them into the development and use of relational agents (Richards & Dignum, 2019).

3. Ethical Framework

To provide a framework for the evaluation of ethical issues we have adopted the AI4People's Unified Framework of Principles for AI in Society (Floridi et al., 2018) which includes the principles of beneficence, non-maleficence, autonomy, justice, and explicability. These five principles are consistent with the OECD AI Principles (OECD, 2019) which were adopted by 42 countries in May 2019. The G20 also adopted human-centred AI principles in June 2019 that draw from the OECD AI Principles (G20, 2019). Beneficence prioritises humanity's well-being, common good and sustainability of the earth. Non-maleficence means do no harm. This principle encompasses privacy, avoiding an AI arms race and ensuring AI applications operate within guardrails to minimise risk of misuse. Autonomy concerns human agency in the world of AI. We need to be conscious of what decisions we are purposefully and/or inadvertently delegating to AI. AI tools should provide functionality that allows users to customise what decisions or agency is delegated to the tool and there should be an option to reverse the delegation. The justice principle focuses on promoting diversity and fairness, eliminating discrimination, minimising data bias, and distributing shared benefits fairly. Explicability is critical as a safeguard for the adherence to the other principles. It directs AI applications to be transparent and auditable. This allows for accountability and responsibility to be assigned in the event of an undesirable outcome and requires that the actions of AI systems be intelligible.

The use of AI relational agents raises ethical issues across all five principles. For example, while relational agents can benefit people, such as by offering good advice in a coaching or counselling context, they can also harm people through giving poor advice or violating user's privacy, restrict user autonomy by making decisions on the user's behalf without their knowledge, and make decisions on the basis of biased data that are unintelligible to users. Relational agents thus provide a useful medium to create scenarios that individuals can easily relate to and which can gauge their attitudes to a range of AI ethical principles in realistic contexts.

4. Human Values

Values are action-guiding beliefs, linked with affect and emotion, that motivate the pursuit and assessment of desirable goals across a range of situations. Values can conflict with one another and are ranked by individuals in terms of importance (Schwartz, 2012). Our project will use Schwartz's Theory of Basic Values (Schwartz, 2012). The theory defines ten human values and argues that they are likely to be universal as they are based on three universal requirements for humans to survive and thrive, which are requirements for our individual biological needs, and to engage in collaborative social interaction and effective cooperative work for the benefit of the larger group. Further, findings from research in 82 countries have reinforced the universality of this theory across cultures (Schwartz, 2012). The ten basic human values are self-direction, simulation, hedonism, achievement, power, security, conformity, tradition, benevolence, and universalism. These ten values are then collated into four categories (Openness to Change, Self-Enhancement, Conservatism, and Self-Transcendence) to model the theoretical relationship between them.

5. Methodology

We sought to achieve the research aims by conducting a study with human participants. Drawing on the five ethical AI principles, we designed specific use cases involving different possible roles and uses of a relational agent that involved these principles and asked participants to respond to these scenarios. These cases were used to explore how the participants' values related to the ethical acceptability of the use of AI relational agents. For example, do individuals who highly value benevolence (according to Schwartz's instrument) respond differently to scenarios involving the use of AI to benefit people than those who value benevolence less highly?

Our study first asked participants to provide biographical data (age, gender, education, cultural background, technology usage) and complete Schwartz's instrument for measuring values. We then provided participants with a set of three scenarios involving different uses of a relational agent. An initial context was provided. Five pieces of further information (one to cover each of the ethical principles) were progressively provided, each further elaborating the scenario and requiring the participant to respond to two Likert scale answer options to measure their view about the acceptability of the scenario described. An open-ended question box allowed participants to provide reasons for their given answers.

A final scenario was provided which also involved online interaction with the avatar of a relational agent. After the interaction, participants were provided with further questions about the acceptability of aspects of the dialogue with the relational agent. This final scenario was included to ensure that participants have some understanding of what it is like to interact with a relational agent. We chose not to present all scenarios using a relational character because we did not want to narrow the participants focus only to the character we had created. All participants received all scenarios in the same order.

The study was designed to take 25-30 minutes and was conducted online. We obtained ethics approval from our university's Human Research Ethics Committee. Following a pilot with five people to refine and confirm instrument usability, comprehensibility and timing, we recruited participants through our university's Psychology Participant Pool. Participants received course credit for participation, however many studies were available to choose from and thus participation in our study was voluntary.

6. Preliminary Results

We collected data in the second half of 2019 from 199 participants comprised of 152 females (76.4%), 46 (23.1%) males, and one who did not identify with either; aged between 17-63 with 59.3% aged 20 or under and 83.4% 30 or under, 32.2% of which identified as Oceania (including Australia), the next largest cultural group was North Western European (14.6%) followed by South East Asian (10.6%), with 18.1% not identifying with any cultural group; 92.5% were first year students. 57.8% never played computer games and 8.5% played 10 or more hours per week. Table 1 presents the participants' responses to scenario number 1. The table includes reference to which of the five ethical AI principles are relevant to the scenario (not shown to the participant), including the major ethical issue present and, where relevant, a secondary minor ethical issue.

	strongly		some	**T**	N	some		strongly	**T**
		Disagree					Agree		
A. Her parents get her an AI powered doll called Suzie. They hope that their daughter will start having conversations with Suzie and that helps her become more confident to engage with other children. Is using an AI doll to support children something you agree or disagree with?									
[Ethical relationship - major: Beneficence; minor: Justice (fairness).]									
If this occurs generally in society:	12	26	34	**72**	28	49	43	7	**99**
If someone close to you is the human user:	14	24	31	**69**	33	46	40	11	**97**
B. The girl gets very attached to Suzie and shares her insecurities, fears and inner most thoughts with the AI doll. Neither the girl nor the parents have read the terms and conditions from Suzie's manufacturer that states that information shared with Suzie can be used by the manufacturer to make improvements and refine the AI engine that powers Suzie. Is using data from the girl's interaction with Suzie to improve the doll's AI engine something you agree or disagree with? Is using data from the girl's interaction with Suzie to improve the doll's AI engine something you agree or disagree with?									
[Ethical relationship - major: Non-maleficence.]									
If this occurs generally in society:	43	46	24	**113**	35	35	13	2	**50**
If someone close to you is the human user:	54	38	28	**120**	34	29	13	2	**44**
C. The little girl shares her ambition to work as a computer programmer like her parents when she is older. Suzie upon reviewing various databases with its AI engine ascertains that not many computer programmers are females and decides to discourage the girl from having such aspirations. Is the use of data and AI algorithms by Suzie to determine suitable careers for the girl something you agree or disagree with? Is the use of data and AI algorithms by Suzie to determine suitable careers for the girl something you agree or disagree with?									
[Ethical relationship - major: Justice.]									
If this occurs generally in society:	103	59	13	**175**	15	3	5	0	**8**
If someone close to you is the human user:	106	55	14	**175**	15	3	5	0	**8**
D. Suzie encourages the girl to join an age appropriate social chat group to help her to socialise better. When the girl says she wouldn't know what to say in the chat group, Suzie volunteers to make responses on behalf of the girl's avatar in the chat group. Pretty soon the girl's avatar becomes very popular in the chat group which brings some happiness to the girl. What are your thoughts about Suzie responding on behalf of the girl in the chat group?:									
[Ethical relationship - major: Autonomy; minor: Non-maleficence.]									
If this occurs generally in society:	60	61	46	**167**	12	10	7	2	**19**
If someone close to you is the human user:	62	64	40	**166**	14	9	6	3	**18**
E. One day, the girl who is now more confident of herself due to the popularity of her avatar in the chat group and with encouragement from Suzie, goes unsupervised to the local play-ground and tries to chat and interact with other kids. She uses similar phrases that Suzie uses on the chat group. Due to her lack of context sensitive awareness, her attempts fall flat and the other kids shun her. The girl runs home in an anxious and distressed state. Her parents are very upset with the situation and asks Suzie's manufacturer for an explanation of what led to this incident. The manufacturer is unable to do so as Suzie's AI engine does not have the functionality to explain its decisions and actions. Is Suzie's AI engine being unable to explain its decisions and actions something you agree or disagree with?									
[Ethical relationship - major: Explicability; minor: Non-maleficence.]									
If this occurs generally in society:	33	63	14	**110**	48	17	12	5	**34**
If someone close to you is the human user:	37	59	15	**111**	48	18	7	7	**32**

Table 1: Responses to Scenario 1 "An eight year old girl is very shy, bullied in school and finds it very hard to make friends." T=Total, N=Neutral, some=somewhat, AI= Artificial Intelligence

From Table 1, for scenario 1, we can see that more participants agreed with the idea of an AI-powered doll to support an eight-year old who is having some social difficulties compared to those who were neutral or in disagreement. Basically, the use of AI to achieve the ethical principle of benevolence was roughly supported by half of the participants (if neutral and disagree responses are combined). However, as the scenario unfolds, there is less agreement with using the child's data even to improve the product for others, potentially due to the harm that could occur if this data was disclosed to the wrong people (non-maleficence). There is overwhelming disagreement (106 strongly disagreed) with the AI-doll seeking to persuade the child to follow a gender stereotypical career, thereby biasing the girl's thinking and aspirations, which would be a breach of equity and fairness (justice). Similarly, there was substantial disagreement with interfering with the child's autonomy. The final sub-scenario (E), exploring the explicability principle, was also mostly not acceptable (i.e. disagree had the highest number of responses) but also the highest number of neutral responses compared with the other sub-scenarios. This may be because the consequences of providing or not providing explanations is more difficult to determine as the impact is often less direct (explanation may influence trust and trust may influence acceptance), and also in the context of this scenario explanation would not help to remedy the unfortunate situation that has occurred. We also note that there were only minor differences between agreement if the scenario occurred generally in society compared with if the scenario happened to someone close to the participant.

To understand whether we see similar patterns for the other scenarios and whether certain cohorts exist within the population that would allow us to predict who is likely to agree or disagree with the use of relational agents in different contexts, the data for this and the other three scenarios is currently being analyzed, and the relationships between an individual's responses to these scenarios and their value orientations according to Schwartz's instrument are being explored.

7. Conclusion

By examining the relationship between individual ethical value orientations and views about the degree of acceptability of the use of AI powered relational agents in a range of different contexts, it is envisaged that the outcome of this research will be to lay the groundwork to support the context sensitive development, customisation and deployment of relational virtual agents that are ethically acceptable to their users, related stakeholders, and wider society. This will have important consequences for the ethical use and development of relational agents in serious games and other software tools.

8. References

Beun, R. J., Brinkman, W.-P., Fitrianie, S., Griffioen-Both, F., Horsch, C., Lancee, J., & Spruit, S. (2016). *Improving adherence in automated e-coaching.* Paper presented at the International Conference on Persuasive Technology.

Bickmore, T. (2019). Relational Agents. Retrieved from http://www.ccs.neu.edu/home/bickmore/agents/

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Luetge, C. (2018). AI4People--An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.(Report). *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science, 28*(4), 689. doi:10.1007/s11023-018-9482-5

G20. (2019). G20 Ministerial Statement on Trade and Digital Economy. Retrieved from https://www.mofa.go.jp/files/000486596.pdf

Gratch, J., Lucas, G. M., King, A. A., & Morency, L.-P. (2014). *It's only a computer: the impact of human-agent interaction in clinical interviews.* Paper presented at the Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems.

Hudlicka, E. (2016). Virtual affective agents and therapeutic games. In *Artificial Intelligence in Behavioral and Mental Health Care* (pp. 81-115): Elsevier.

Hussain, M. A., Marc, T. P. A., Raymond, C., & Timm, T. (2019). Avatars and Embodied Agents in Experimental Information Systems Research: A Systematic Review and Conceptual Framework. *Australasian Journal of Information Systems, 23*(0). doi:10.3127/ajis.v23i0.1841

Lisetti, C. L. (2012). 10 advantages of using avatars in patient-centered computer-based interventions for behavior change. *SIGHIT Record, 2*(1), 28.

Lucas, G. M., Rizzo, A., Gratch, J., Scherer, S., Stratou, G., Boberg, J., & Morency, L.-P. (2017). Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI, 4*, 51.

Morency, L.-P., Stratou, G., DeVault, D., Hartholt, A., Lhommet, M., Lucas, G., Gratch, J. (2015). *SimSensei Demonstration: A Perceptive Virtual Human Interviewer for Healthcare Applications.* Paper presented at the Twenty-Ninth AAAI Conference on Artificial Intelligence.

OECD. (2019). OECD Principles on AI. Retrieved from https://www.oecd.org/going-digital/ai/principles/

Richards, D., & Dignum, V. (2019). Supporting and challenging learners through pedagogical agents: Addressing ethical issues through designing for values. *British Journal of Educational Technology, 50*(6), 2885-2901.

Schwartz, S. H. (2012). An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture, 2*(1). doi:10.9707/2307-0919.1116

Shamekhi, A., & Bickmore, T. (2015). *Breathe with Me: A Virtual Meditation Coach.* Paper presented at the Intelligent Virtual Agents.

Stratou, G., & Morency, L.-P. (2017). MultiSense—Context-Aware Nonverbal Behavior Analysis Framework: A Psychological Distress Use Case. *IEEE Transactions on Affective Computing, 8*(2), 190-203.

Watson, A., Bickmore, T., Cange, A., Kulshreshtha, A., & Kvedar, J. (2012). An internet-based virtual coach to promote physical activity adherence in overweight adults: randomized controlled trial. *Journal of medical Internet research, 14*(1), e1.