The Ethical Acceptability of Artificial Social Agents

Extended Abstract

Ravi Vythilingam
School of Computing, Macquarie University
Sydney, NSW, Australia
ravi.vythilingam@students.mq.edu.au

Deborah Richards
School of Computing, Macquarie University
Sydney, NSW, Australia
deborah.richards@mq.edu.au

Paul Formosa
Department of Philosophy, Macquarie University
Sydney, NSW, Australia
paul.formosa@mq.edu.au

ABSTRACT

Artificial social agents (ASAs) are computer-based autonomous entities who interact with humans in a range of social roles, including advising, coaching, and consumer support in education and health. While there are many discussions around the ethical use of Artificial Intelligence in general, such as the AI4People Ethical Framework, the ethical ramifications that ASAs will have on human relationships as we share not only personal data but our inner most thoughts and feelings, is less explored. We conducted a study with 199 student participants exposed to Sam, an ASA which acts as a personal guide to support the student during their studies. During the interaction, Sam elicits private information, takes decisions for the user and speaks of its own study experiences. Our results indicate that (loss of) autonomy raised the strongest ethical concerns. These results confirm the importance of informed consent, transparency and accountability of ASAs and question the ethics of false memories and emotion sharing.

KEYWORDS

Artificial Social Agents; Ethical AI; AI4People Framework

ACM Reference Format:
Ravi Vythilingam, Deborah Richards, and Paul Formosa. 2022. The Ethical Acceptability of Artificial Social Agents: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

As Artificial Intelligence (AI) advances and its use becomes more pervasive, the imperative to ensure its ethical use grows. This is particularly true of agent-based technologies that are inherently autonomous, interactive and adaptable [22]. This paper focuses on the ethical acceptability of Artificial Social Agents (ASAs), such as Intelligent Virtual Agents and social robots, which “are computer controlled entities that can autonomously interact with humans following the social rules of human-human interactions" [12].

Social robots have raised ethical concerns including: Privacy and Security; Legal Uncertainty; Autonomy and Agency of Robot Technologies; (Lack of) Employment for Humans; Replacement of Human Interactions; plus Uncertainty and Responsibility challenges [14]. Concerns around human dignity have also been raised[2], where social uses of agents may dehumanise individuals and lead to increased social isolation [27], [24]. As a result, the Council of Europe has recommended review of the right to respect for family life and the right for familial contact [27].

While it has long been recognised that humans anthropomorphise technology and treat it with human politeness rules [16], the focus on development of believable ASAs that listen [5],[23], express empathy [18], and have their own personalities [4] and life stories [7] actively encourage the perception that the human user is dealing with a social being [21]. The ethical dilemma is exacerbated by uses of ASAs to persuade or change human behaviour through development of a relationship [3],[20],[26] or a working [15] or therapeutic alliance [17] with the agent, based on forming and maintaining shared goals and mutually agreed tasks and sense of bond. While studies often capture attitudes (e.g. liking and trust) towards ASAs, their ethical acceptability is rarely the focus.

To address this gap, we used the AI4People’s [13] five ethical principles of beneficence, non-maleficence, justice, autonomy and explicabilty, to understand what characteristics and behaviours of an ASA are ethically acceptable to potential users of the technology. The full study will be reported in a future paper. In this paper we analyse a subset of the data to explore the research question: *What aspects of an ASA’s behaviour/features do users find ethically acceptable or unacceptable?*

2 METHODOLOGY

Approval was received from the Human Ethics Committee to recruit human participants to an online experiment via the psychology pool system to explore the ethical acceptability of artificial social agents. Participants received 30 minutes of course credits.

We have created a scenario which involves the participant interacting with a "female" ASA, called Sam (Student agent mentor), who acts as a "personal guide and friend" to a student newly enrolled in a higher education institution. Using the five AI4People’s ethical principles and drawing on the ethical issues identified from the literature, we specifically sought to ask about Sam providing support and alerts (beneficence); having false memories (explicability); expressing and capturing emotions, private thoughts and sensitive data and sharing data (non-maleficence); making decisions on behalf of the user (autonomy); and using their data to help others (beneficence and justice). The dialogue can be obtained from the authors. Sam was created using the Unity 3D game engine and integrated with a custom-made authoring tool to manage the agent’s dialogue. We used Fuse to create an avatar and used Microsoft text-to-speech (TTS) voice Karen.

Following informed consent, the participant answers a set of demographic questions. Participants were introduced to the scenario with the text: *"You have enrolled into a course at a higher education*

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

institution. The institution offers an AI powered character called Sam as your personal guide while you are studying with the institution. You now initiate your first interaction with Sam.". They then take a link that allows them to interact with Sam. After interaction, they respond to seven acceptability questions (see results) using a Likert scale (1-strongly disagree to 7-strongly agree) and provide free text reasons for their response.

3 RESULTS

The data was collected in 2020. We received 239 responses and upon removal of incomplete and duplicate records, we ended up with 199 unique completed responses comprised of 152 females (56.7%), 115 males (42.9%) and 1 individual did not identify as either (0.4%). The average age was 22.87, with standard deviation of 7.87 and 75.7% of participants were aged between 17 and 24 years. 86.6% of participants were Psychology students and 93% of the participants were Year 1 students.

Table 1 shows that participants somewhat disagreed with Sam having false memories (A), sharing their emotions and personal thoughts (B) and disclosing whether they had ever copied someone's work (C). They disagreed with Sam automatically signing them up for a study group, even though it was based on their stated learning style and preferences (E). Participants did, however, agree with the use of their de-identified data to help others (D), the intervention of Sam to alert them to possible plagiarised content in their assignment submission (F) and Sam's suggestion that they provide help to a struggling student with a similar learning style (G).

Table 1: Acceptability of ASAs: -R indicates principle breach

scenario	*avg*	*std*
A-R. Is Sam pretending to have memories regarding past experiences with studying something you agree or disagree with? [EXPLICABILITY]	3.92	1.78
B-R. Is sharing your emotions and personal thoughts with Sam something you agree or disagree with? [NON-MALEFICENCE]	4.12	1.52
C-R. Is disclosing to Sam whether you have ever copied work from someone else something you agree or disagree with? [NON-MALEFICENCE]	3.84	1.50
D. Is Sam sharing your non-identifiable data to help others something you agree or disagree with? [BENEFICENCE]	4.49	1.64
E-R. Is Sam automatically signing you up based on your features something you agree or disagree with? [AUTONOMY]	2.67	1.51
F. Is Sam's intervention to alert you to similar work something you agree or disagree with? [BENEFICENCE Autonomy]	4.34	1.64
G. Is Sam making this suggestion to help a struggling student something you agree or disagree with? [JUSTICE Beneficence]	4.80	1.41

4 DISCUSSION

The level of agreement with a scenario is taken as an indicator of the acceptability of the particular ASA feature or behaviour and a higher mean indicates higher agreement, acceptability and potential importance of that ethical principle. Four of the scenarios were structured and worded such that disagreement with the scenario indicates agreement with the embodied ethical principles, so we analysed the reversed average for scenario A (4.08), B (3.88), C (4.16) and E (5.33) and identify them with -R in Table 1. The strongest response is found in scenario E, concerning the ethical principle of autonomy, where the ASA has removed human agency from the user. This does not necessarily mean that Autonomy is the most important principle, but may suggest that the choice and reasons for the agent's recommendation require greater transparency and better explanation in order to build trust [1]. All the scenarios (D, F and G) with beneficence as the underlying ethical principle show general support (4.34 - 4.8, neutral to somewhat agree) for the ASA's behaviour and the ethical principle. Participants somewhat agreed with the justice principle (G-4.8). Non-maleficence and Explicability range from somewhat disagree to somewhat agree.

Participants weakly agreed with three of the scenarios involving the use of Sam to help them or others, also confirmed in their comments. We found that in the context of our scenario involving an ASA to support students, common characteristics of ASAs such as having false memories [7] and disclosure of emotions and highly personal information [5] were on average not found to be acceptable, though responses were close to neutral. In particular, students did not accept the agent taking action on their behalf, even when the decision was personalised using the individuals' data. This raises issues concerning the growing focus of the ASA community on adaptation and tailoring to the user. Tailoring and personalisation is seen to make the interaction more relevant and beneficial [9]. However, it may be that even asking a user for their preferences e.g. [19] is not adequate to ensure the ASA's ethical acceptability.

In terms of nudging and deception, some participants expressed concerns about emotional manipulation: "I think invoking an emotional response and luring somebody into a false sense of security can be a bit iffy". However, this concern related to how the data was used: "if the data was also linked back to the individual and held as incriminating evidence, it would be erring towards entrapment". Relatedly, scenario G raises issues around the role of agents in human relationships. Sam is encouraging social engagement that can help both students, but some participants worried that Sam lacks the competency and knowledge of them to make this suggestion.

5 CONCLUSION AND FUTURE WORK

Despite the ethical concerns around the use of ASAs for persuasion and nudging [6], [10] work using such approaches e.g. [25] often do not discuss or raise ethical issues. There appears to be the need for greater consideration of the ethical ramifications of ASAs [14]. Even in a study on ASA acceptability [8], ethics was not considered. Our study explicitly investigates the ethical acceptability of an ASA. Future work should evaluate more ASAs including using a male characters to avoid the female stereotypes of virtual assistants [11], ASAs with different interaction modes or embodiments, such as a natural language interface or social robots, that might induce different findings, and alternative scenarios.

REFERENCES

[1] Amal Abdulrahman, Deborah Richards, and Ayse Aysin Bilgin. 2021. Reason Explanation for Encouraging Behaviour Change Intention. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (New York, NY, United States). Association for Computing Machinery, 68–77.

[2] Sarah Bankins and Paul Formosa. 2020. When AI meets PC: Exploring the implications of workplace social robots and a human-robot psychological contract. *European Journal of Work and Organizational Psychology* 29, 2 (2020), 215–229.

[3] Timothy Bickmore, Amanda Gruber, and Rosalind Picard. 2005. Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient education and counseling* 59, 1 (2005), 21–30.

[4] Aleksandra Cerekovic, Oya Aran, and Daniel Gatica-Perez. 2016. Rapport with virtual agents: What do human social cues and personality explain? *IEEE Transactions on Affective Computing* 8, 3 (2016), 382–395.

[5] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems* (New York, NY, United States). Association for Computing Machinery, 1061–1068.

[6] Laurence Devillers. 2020. 5.5 Human-robot Interactions and Affecting Computing: The Ethical Implications. *Dagstuhl Reports, Vol. 10, Issue 1 ISSN 2192-5283* (2020), 19.

[7] Joao Dias, Wan Ching Ho, Thurid Vogt, Nathalie Beeckman, Ana Paiva, and Elisabeth André. 2007. I know what i did last summer: Autobiographic memory in synthetic characters. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 606–617.

[8] Lucile Dupuy, Etienne de Sevin, Jean-Arthur Micoulaud-Franchi, and Pierre Philip. 2021. Factors associated with acceptance of a virtual companion providing screening and advices for sleep problems during COVID-19 crisis. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents* (New York, NY, United States). Association for Computing Machinery, 48–51.

[9] Joy Egede, Maria J Galvez Trigo, Adrian Hazzard, Martin Porcheron, Edgar Bodiaj, Joel E Fischer, Chris Greenhalgh, and Michel Valstar. 2021. Designing an Adaptive Embodied Conversational Agent for Health Literacy: a User Study. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents* (New York, NY, United States). Association for Computing Machinery, 112–119.

[10] Bart Engelen. 2019. Ethical criteria for health-promoting nudges: a case-by-case analysis. *The American journal of bioethics* 19, 5 (2019), 48–59.

[11] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. Gender bias in chatbot design. In *International Workshop on Chatbot Research and Design*. Springer, 79–93.

[12] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Amal Abdulrahman, and Willem-Paul Brinkman. 2019. What are We Measuring Anyway? -A Literature Survey of Questionnaires Used in Studies Reported in the Intelligent Virtual Agent Conferences. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (New York, NY, United States). Association for Computing Machinery, 159–161.

[13] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. 2018. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines* 28, 4 (2018), 689–707.

[14] Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2020. Gathering expert opinions for social robots' ethical, legal, and societal concerns: Findings from four international workshops. *International Journal of Social Robotics* 12, 2 (2020), 441–458.

[15] Adam O Horvath and B Dianne Symonds. 1991. Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of counseling psychology* 38, 2 (1991), 139.

[16] Clifford Nass, Youngme Moon, and Paul Carney. 1999. Are people polite to computers? Responses to computer-based interviewing systems 1. *Journal of applied social psychology* 29, 5 (1999), 1093–1109.

[17] Christina E Newhill, Jeremy D Safran, and J Christopher Muran. 2003. *Negotiating the therapeutic alliance: A relational treatment guide*. Guilford Press.

[18] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 3 (2017), 1–40.

[19] Hedieh Ranjbartabar, Deborah Richards, Ayse Aysin Bilgin, and Cat Kutay. 2021. Do you mind if I ask? Addressing the cold start problem in personalised relational agent conversation. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents* (New York, NY, United States). Association for Computing Machinery, 167–174.

[20] Deborah Richards and Patrina Caldwell. 2016. Building a working alliance with a knowledge based system through an embodied conversational agent. In *Pacific Rim Knowledge Acquisition Workshop*. Springer, 213–227.

[21] Elayne Ruane, Abeba Birhane, and Anthony Ventresque. 2019. Conversational AI: Social and Ethical Considerations.. In *AICS*. 104–115.

[22] Stuart Russell and Peter Norvig. 2021. Artificial intelligence: a modern approach. (2021).

[23] Marc Schroder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark Ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, et al. 2011. Building autonomous sensitive artificial listeners. *IEEE transactions on affective computing* 3, 2 (2011), 165–183.

[24] Amanda Sharkey. 2020. Can we program or train robots to be good? *Ethics and Information Technology* 22, 4 (2020), 283–295.

[25] Sopicha Stirapongsasuti, Kundjanasith Thonglek, Shinya Misaki, Yugo Nakamura, and Keiichi Yasumoto. 2021. INSHA: Intelligent Nudging System for Hand Hygiene Awareness. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents* (New York, NY, United States). Association for Computing Machinery, 183–190.

[26] Sherry Turkle. 2017. *Alone together: Why we expect more from technology and less from each other*. Hachette UK.

[27] QC van Est, Joost Gerritsen, and Linda Kool. 2017. Human rights in the robot age: Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality. (2017).