Int. J. Human–Computer Studies 172 (2023) 102980

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

A principlist-based study of the ethical design and acceptability of artificial social agents

Deborah Richards [a,*], Ravi Vythilingam [b], Paul Formosa [c]

[a] School of Computing, Faculty of Science and Engineering, 4 Research Park Drive, Room 369, Macquarie University, NSW 2109 Australia
[b] School of Computing, Faculty of Science and Engineering, Macquarie University, NSW 2109 Australia
[c] Department of Philosophy, Faculty of Arts, Macquarie University, Australia

ARTICLE INFO

Keywords:
Artificial social agents
Ethical acceptability
Intelligent virtual agents
Social robots
AI4People ethical principles
Schwartz' human values

ABSTRACT

Artificial Social Agents (ASAs), which are AI software driven entities programmed with rules and preferences to act autonomously and socially with humans, are increasingly playing roles in society. As their sophistication grows, humans will share greater amounts of personal information, thoughts, and feelings with ASAs, which has significant ethical implications. We conducted a study to investigate what ethical principles are of relative importance when people engage with ASAs and whether there is a relationship between people's values and the ethical principles they prioritise. The study uses scenarios, embedded with the five AI4People ethical principles (Beneficence, Non-maleficence, Autonomy, Justice, and Explicability), involving ASAs taking on roles traditionally played by humans to understand whether these roles and behaviours (including dialogues) are seen as acceptable or unacceptable. Results from 268 participants reveal the greatest sensitivity to ASA behaviours that relate to Autonomy, Justice, Explicability, and the privacy of their personal data. Models were created using Schwartz's Refined Values as a possible indicator of how stakeholders discern and prioritise the different AI4People ethical principles when interacting with ASAs. Our findings raise issues around the ethical acceptability of ASAs for nudging and changing behaviour due to participants' desire for autonomy and their concerns over deceptive ASA behaviours such as pretending to have memories and emotions.

1. Introduction

Artificial Social Agents (ASA) are Artificially Intelligent (AI) software driven entities in virtual or physical form. They include AI-based social robots, embodied conversational agents, relational agents, and intelligent virtual agents (IVA). ASAs are programmed with certain rules and preferences to act autonomously with humans (Fitrianie et al., 2019) and their sophistication will increase over time (Russell et al., 2015). ASAs are increasingly becoming the face of AI for the broader public as they take on more human-like roles in society, particularly in education, healthcare, childcare, eldercare, and in coordinating, advising, and coaching settings. These AI applications are based on datasets and algorithms that we choose to use, thereby creating moral choices and implications (Ntoutsi et al., 2020). Many governments, universities, organisations, industry forums and public figures have raised concerns regarding the widespread use of AI technologies. These concerns include cognitive degeneration and skill loss, threats to human autonomy (Danaher, 2018), accountability, privacy, discrimination, security, societal dynamics, economic impacts (IEEE, 2018), and even their existential threat to human existence (Chalmers, 2009).

As a result, there has been a significant recent focus on the ethics of AI with numerous organisations, companies and jurisdictions publishing their own set of frameworks, principles, and guidelines on AI ethics. For example, Floridi et al. (2018) reviewed a number of these guidelines and synthesised five overarching ethical principles to form the AI4People Framework. However, many organisations are finding it challenging to put these principles into practice, with minimal adherence by developers and management (Hagendorff, 2020; Mittelstadt, 2019). Furthermore, most of these guidelines focus on the ethics of AI applications related to data science, big data, and machine learning, and insufficient focus has been given to the ethics of the design and acceptability of ASAs.

While ethical issues have previously been identified with ASAs, such as whether dialogue systems should be used to change humans' goals or

* Corresponding author at: School of Computing, Faculty of Science and Engineering, 4 Research Park Drive, Room 369, Macquarie University, NSW 2109 Australia.
E-mail address: deborah.richards@mq.edu.au (D. Richards).

https://doi.org/10.1016/j.ijhcs.2022.102980
Received 16 April 2022; Received in revised form 10 November 2022; Accepted 9 December 2022
Available online 17 December 2022

actions (Allwood et al., 2000), and some theoretical work exists exploring significant ethical implications of ASAs, such as their impacts on human autonomy (Formosa, 2021), there have been few studies designed to investigate what ASA behaviours are seen as ethically acceptable. One exception is an early study by van Vugt et al. (2009) that explored ASA ethics from the perspective of trustworthiness, reporting that participants found the obese embodied agent who provided weight loss advice more trustworthy, which lends support to a body of work that uses ASAs to challenge stereotypes and biases (Bickmore et al., 2021; Rossen et al., 2008; Sebastian and Richards, 2017; Vugt et al., 2010). A 50 year review of articles in the Journal of Human Computer Studies on human-automation interaction by Janssen et al. (2019) considered a range of ethical and social dilemmas that included job security and whether we should allow machines to make moral decisions. The article noted that despite a rise in the use of embodied agents, including affective agents, there was no consideration of whether it was ethical for such agents to exhibit or respond to human emotions. While ASA researchers may acknowledge ethical concerns, for example, due to the use of persuasive techniques to change human behaviour (Zalake et al., 2021), recognition of the influence of anthropomorphism and its potential benefits has encouraged ASA designers to increase ASA humanness and believability; often making measurement of anthropomorphism (David et al., 2022) and achievement of similar human-human responses such as closeness (Loveys et al., 2022) the focus of their research. We agree with the conclusion of Rapp et al. (2021) based on their review of 83 papers studying human interaction with chatbots over the past 10 years that:

> *it is quite remarkable that a critical debate tackling the ethical issues arising from this 'subtle deception', be either intentionally cueing perceptions of humanness, or blurring the distinction between humans and machines in human-in-the-loop systems, or simply letting the user free to believe what she wants, is completely absent in our corpus [resulting in a need to] conduct research focused on design, and tackle relevant ethical issues mainly arising from the users' tendency to ascribe humanness to chatbots* (p. 20).

With the accelerating use of ASAs in multiple domains of human-computer interaction (HCI), more research is needed targeting the ethics of ASAs. Our study addresses this gap.

To provide a framework in which to explore the ethical acceptability of ASAs, including their humanlike roles, appearances and behaviours, the specific aim of the study is to investigate what ethical principles are of relative importance to humans when we engage with ASAs and whether there is a relationship between our values and the ethical principles we prioritise in such cases. Our specific research questions are presented at the end of Section 2 which provides a literature review. The methodology utilised to address the research questions is articulated in Section 3. The results of the study are presented in Section 4 and discussed in Section 5. Section 6 considers limitations and future directions for research. Conclusions are provided in Section 7.

2. Literature review

Ethics has become one of the leading areas of focus in the Artificial Intelligence sphere (Dignum, 2018), and has led to a large literature, such as (Bostrom, 2014; Danaher, 2018; IEEE, 2018; Mittelstadt et al., 2016; Wallach and Allen, 2008). Comparisons, analyses and listings of the various frameworks and ethical principles have been carried out by researchers (including Hagendorff, 2020, Floridi et al., 2018, Jobin et al., 2019 and Fjeld et al., 2020) and by organisations, such as AlgorithmWatch (2020).

Floridi and Cowls, (2019) examined various sets of ethical AI principles and synthesised these into five ethical principles to form the AI4People Framework (Floridi et al., 2018). These principles are beneficence, non-maleficence, autonomy, justice, and explicability. The first four principles are traditional bioethics principles (Beauchamp and Childress, 2001), while the fifth is new. Beneficence focuses on humanity's well-being, sustainability, and the common good. Non-maleficence focuses on minimising harm, including protecting privacy, and ensuring AI operates within guardrails to minimise misuse. Autonomy is concerned with human agency, highlighting the need to be conscious of what decisions we delegate to AI. AI tools should also provide functionality to allow users to set and reverse what decisions or agency is delegated. Justice focuses on promoting diversity and fairness, minimising data bias, eliminating discrimination, and promoting shared benefits. Explicability is a critical safeguard for adherence to the other principles. It requires transparency and auditability to support accountability in the event of an undesirable outcome. These five principles are consistent with the OECD AI Principles (OECD, 2019) adopted by 42 countries in May 2019. The G20 also adopted human-centred AI principles in June 2019, drawing on the OECD AI Principles (G20, 2019).

Jobin et al. (2019) analysed 84 sets of AI ethical frameworks globally and found broad convergence of the ethical principles into five areas, namely transparency, justice and fairness, non-maleficence, responsibility, and privacy. Hagendorff (2020) analysed and compared 15 internationally recognised AI Ethics guidelines and implementations. The paper identified overlaps and omissions among the principles. Commonly identified principles include privacy, accountability, fairness, safety, sustainability, and auditing. Omissions include impacts due to lack of focus on diversity, political misuse, industry funded research and ecological costs. A separate comparison and analysis by Fjeld et al. (2020) of thirty-six prominent AI principles unearthed eight key Principled AI themes: Privacy, Accountability, Safety and Security, Transparency and Explainability, Fairness and Non-discrimination, Human Control of Technology, Professional Responsibility, and Promotion of Human Values.

Although different terms are used, these various frameworks largely align with the AI4People's principles described above as shown in Fig. 1, where we map the AI ethical principles identified by Hagendorff (2020), Jobin et al. (2019), and Fjeld et al. (2020)'s analysis to the AI4People Framework. As the AI4People's five ethical principles encompasses the ethical principles recommended by most published frameworks and are consistent with the AI principles adopted by the OECD and the G20, this paper will utilise these principles as representative AI ethical principles for the purposes of this study.

Artificial Social Agents (ASA) encompass AI-based social robots, embodied conversational and relational agents, and intelligent virtual agents (IVA). While the ethical frameworks developed for AI outlined above largely apply to ASAs, the social aspect of ASAs mean that they require further attention. ASAs are programmed to converse autonomously with humans in a social manner (Fitrianie et al., 2019). ASAs do not need to be indistinguishable from humans to be seen as relatable social agents; the efficacy of the interaction between humans and the artificial agents can be based on the interactivity and shared consequences of the human-ASA relationship (Kempt, 2020). Kempt further states that ASAs can be categorised by conversational skill levels, ability to understand explicit and implicit human expressions, and the faculty to respond appropriately. There also exists a significant amount of literature on social robots and their possible impacts, such as Breazeal et al. (2004), Turkle (2011), Bankins and Formosa (2020), Lutz et al. (2019), and Pashevich (2021). However, these studies do not specifically explore what users consider ethically acceptable.

The application and sophistication of ASAs will only grow with humans developing stronger and deeper relationships as ASA designers seek to create a human-agent working alliance (Bickmore et al., 2005; Richards and Caldwell, 2016; Turkle, 2011) and the agents become more autonomous and powerful over time (Russell et al., 2015). Areas where ASAs are increasingly being used include healthcare, education, coaching and counselling, eldercare, childcare, personal relationships, and personal assistants. The use of ASAs in decision-making roles traditionally played by humans raises various ethical considerations,

Fig. 1. Mapping Hagendorff (2020), Jobin et al. (2019), and Fjeld et al., (2020) on to the AI4People Framework.

including concerns with moral deskilling (Vallor, 2015). Further, there will be a growing tendency for humans and agents to be working in tandem to make decisions in the pursuit of a specific objective. This will require an alignment of ethical principles and moral values between the human and the ASA (Greene et al., 2016), however achieving this alignment of princples, values and preferences will be difficult (Soares and Fallenstein, 2015).

Moor (2009) outlines four levels of artifical moral agents (AMAs), from the most basic to the most advanced. Based on this ethical agent framework, Formosa and Ryan (2021) define Artificial Moral Agents (AMA) as applications that can process external inputs to make ethical decisions autonomously in unique and changing scenarios without real-time human input. Since ASAs may have to make ethically important decisions in real-time, they can also be considerd (in some cases) to be AMAs. Papagni and Koeszegi (2021), in a comparative review of the artificial agents literature, conclude that it is both ethical and essential to endow ASAs with intentionality, social ability, and goal-driven rational behaviour, provided that there is transparency of its design, features, and implementation. Some researchers however, such as van Wynsberghe and Robbins (2019) and Sharkey (2020) have questioned the very rationale for developing AMAs. Alternatively, Formosa and Ryan (2021) have argued for a refined approach where AMAs are only utilised in complex situations where real-time decisions are required to prevent harm such as agents in autonomous vehicles and carebots.

Numerous ethical concerns have been raised regarding ASAs. Fosch-Villaronga et al. (2020)'s paper summarise ethical concerns from a group of 43 experts from 14 countries. They include Privacy and Security; Replacement of Human Interactions; Autonomy and Agency of Agents; Legal Uncertainty; Loss of Human Employment; and Responsibility Challenges. Leading scientists and engineers from the "Spoken Language Interaction with Virtual Agents and Robots" (SLIVAR) community considered the following questions: what ethical issues exist, how can ethical agents be created, and whether an agent should be able to pursue goals unknown to the user (Devillers, 2020). They raised specific concerns over assisting vulnerable people, and the use of affective computing and cognitive architectures to persuade and nudge individuals. Sharkey (2020) raises the folllowing concerns: less human contact; dehumanization and reduced personal control; less

privacy; less personal freedom; deception and infantilisation (through use of artificial toys/pets, vulnerable groups may believe them to be real humans or pets); and appropriate control of the technology. Owe and Baum (2021) raise ethical concerns over the predominant portrayal of ASAs using female avatars, in violation of the "ACM Code of Ethics and Professional Conduct" (Gotterbarn et al., 2018).

Luxton (2020) focusing on global public health, warns of the risk of harm, loss of privacy, inequitable access and bias if the needs of the individual and cultural differences are not taken into account, recommending the establishment of guidelines and professional codes to ensure their ethical design and use. Fiske et al. (2019), through a thematic literature review into the utilisation of embodied AI agents in the field of mental health, collated the following ethical concerns: harm prevention; data ethics; lack of agreed and standardised procedures regarding the development and deployment of AI agents; policy gaps in terms of ethics and regulations; and risk of misuse such as the AI agents replacing current services. Turkle (2011) and Szczuka et al. (2021) raise concerns about the impact of children interacting with ASAs, voice assistants and embodied robots. In the carebot space, Scheutz (2017) raises concerns regarding vulnerable populations forming unreciprocated emotional bonds that could potentially cause harm to humans. Similarly, Danaher (2018) argues that we need a risk/reward structure to evaluate the ethical use of AI assistants based on their impacts on cognitive degeneration, autonomy, and interpersonal interactions.

Engelen (2019) discusses the appropriate use of persuasive technologies and raises three main questions: is the recommended action appropriate; is the approach ethical; and who is doing the persuading. Borenstein and Arkin (2016) discuss in what instances social robots or similar agents should be allowed to 'nudge' humans towards a more ethical position. The use of ASAs to persuade someone to change their attitude or behaviour raises many issues, most importantly the assumption that the change is in the best interest of the persuaded person (Wang et al., 2019). They further stress avoidance of deception and transparency concerning who designed the system; what data is collected and its purpose; consent for data collection; non-discriminatory responses and advice; and ongoing conversation to ensure compliance with ethical principles and regulations. The SLIVAR community (Devillers, 2020) have also expressed concerns around confusion of the artificial agent's "status", due to strategies such as giving the agent a name, humanlike appearance and "life" that can lead to forming unhealthy relationships with them, resulting in manipulation, isolation, disappointment, and machine addiction.

Explicability of an ASA's actions is critical for humans to establish trust in the agent (Miller, 2019). Papagni and Koeszegi (2020) claim that for an ASA to be explainable, three areas need to be addressed: the nature of the explanation, interaction context, and the human ability to comprehend. As part of eXplainable AI (XAI), Verhagen et al. (2021) proposes a two-dimensional explanation framework to classify AI applications and ASAs, producing three catagories: incomprehensible, interpretable, and understandable. They posit that for an ASA to move from the incomprehensible category to interpretable, transparency is required. When both transparency and explainability is present, the ASA is understandable.

A literature review and analysis (Hussain et al., 2019) of 90 research studies on the interaction between humans and avatars/ASAs identified six design elements to be taken into consideration. They are: (1) *Proteus effect* of unintended influence on the user; (2) *Uncanny valley effect* when an avatar looking like a human discourages its usage; (3) creating *presence* in the human-agent social interaction to enhance effectiveness; (4) influence of *persuasive* design in nudging users; (5) *empathic* features to encourage a more productive interaction; and (6) impact of *customisability* on human/agent attachment. How these factors are designed for and implemented has important ethical implications. Dignum (2019) proposed the ART principles of Accountability, Responsibility and Transparency to support the design of ethical ASAs. Fosch-Villaronga et al. (2020) argued that researchers and developers have minimal understanding of the attitudes and requirements of potential users and recommended a human centred design approach. Addressing this identified gap in understanding, our first and primary research question to support the study's aim is:

RQ1: *What aspects of an Artificial Social Agent's behaviours and features do users find ethically acceptable or unacceptable?*

A study on situational ethics that compared student participants' responses between a personal perspective and society's view found that personal ethical views were stronger than those perceived for society (McNichols and Zimmerer, 1985). Within the context of understanding ASA ethical acceptability, we sought to explore whether the users' position on the acceptability of the ASA's behaviours differed when they were responding *in general* from a broader social perspective, or *personally*, such as when considering an ASA interacting with someone close to them. This leads to our second research question:

RQ2: *Do users rate the ethical acceptability of ASAs differently when utilised generally by society as compared to by someone personally close to them?*

Taking up the recommendation by Fosch-Villaronga et al. (2020) to follow a human-centred design approach requires taking human values into account. According to Schwartz's (2006) values theory, values are beliefs that are associated with affect (i.e., emotion), refer to goals that motivate action across broad situations, and their ordered and relative importance serve as a basis and guide for an individual's action. The dominant theory here is Schwartz's Theory of Basic Values (Schwartz, 2012). This theory defines ten human values and posits that they are likely to be universal as they are based on three universal requirements for humans to survive and thrive, namely what is required for our biological needs, collaborative social interaction, and effective teamwork to meet the larger group's objectives. Research findings from 82 countries have reinforced the universality of this theory across cultures (Schwartz, 2012). This theory has been subsequently refined to nineteen values providing better granularity and accuracy in ordering the values in a "continuum based on their compatible and conflicting motivations, expression of self-protection versus growth, and personal versus social focus" (Schwartz et al., 2012). Studies conducted in 10 countries ($N = 6059$) have assessed and confirmed Schwartz's Refined Value Theory (Schwartz et al., 2012). These nineteen values form a motivational continuum categorised under four higher order categories (Schwartz et al., 2012): 1) "Openness to Change", which is comprised of self-direction-thought, self-direction-action, and simulation; 2) "Self--Enhancement", which is comprised of achievement, power-dominance, and power-resources; 3) "Conservatism" which is comprised of security-personal, security-societal, tradition, conformity-rules, and conformity-interpersonal; and 4) "Self-Transcendence", which is comprised of benevolence-dependability, benevolence-caring, universalism-concern, universalism-nature, and universalism-tolerance. The remaining values of hedonism crosses over both Openness to Change and Self-Enchantment; face crosses over both Self-Enhancement and Conservationism; and humility crosses over both Conservationism and Self-Transcendence.

The ethical acceptability of ASAs will likely depend on both the ASA's actions and the moral values of those assessing the ASA's actions. We therefore explored whether there was any relationship between an individual's values using Schwartz' Refined Values theory and the AI4People's ethical principles. Any such relationship can assist with designing ethically acceptable ASAs and help determine how ethical principles and moral values can be aligned between humans and ASAs (Greene et al., 2016). This leads to our final research question:

RQ3: *Can we predict an individual's priorities for each of the five AI4People ethical principles based on their values?*

3. Methodology

To explore our research questions, we conducted a study approved by our university's Human Ethics Committee involving participants who were presented with vignettes of various uses of ASAs that were aligned to the five AI4People ethical principles and designed to elicit some of the ethical concerns identified in the literature. The study design is described in Section 3.1. Materials are presented in Section 3.2. The data collection and analysis procedures are described in Section 3.3.

3.1. Study design and procedure

To answer RQ1 and identify the acceptability of different ASA roles and behaviours, participants responded to various vignette questions as they worked through three descriptive scenarios and a final scenario involving interaction with a conversational avatar ASA. We chose not to present all scenarios using an interactive ASA because we did not want to narrow the respondents thinking to our character. To answer RQ2 and identify any difference between general and personal responses, each vignette contained sub-scenarios that encapsulated one or more of the five ethical principles eliciting two responses, one for 'If this occurs generally in society' (General) and another for 'If someone close to you is the human user' (Personal). To answer RQ3 and attempt to predict participants' responses to vignettes based on their personal values, we used Schwartz's revised Portrait Values Questionnaire (PVQ-RR) to capture an individual's values, due to this measure's features, wide usage, and reliability across cultural groups (Schwartz and Cieciuch, 2021). Table 1 shows the study procedure.

Participants for this online study were recruited through our university's psychology pool. Students could voluntarily choose this study from a list of available studies.

3.2. Materials

Materials developed for this study include three theoretical ASA scenarios and one involving an implemented ASA. The scenario protagonists include a (1) child, (2) 'normal' adult, (3) vulnerable adult and (4) an undergraduate student. The first three descriptive scenarios are text-based, while in the fourth scenario the participant 'plays' themselves (as our participants are students) as they interact with an implemented ASA. Scenarios were created following the Experimental Vignettes Methodology (EVM) (Atzmüller and Steiner, 2010). Aguinis and Bradley (2014) assert that EVM is a good survey methodology choice 'when the goal is to investigate sensitive topics in an experimentally controlled way and [is] popular in ethical decision-making contexts'. In designing the three descriptive scenarios, we ensured each scenario: (1) describes a different problem situation requiring distinct human profiles as the protagonist interacting with the ASA and (2) encompasses ethical dilemmas that cover all the five AI4people ethical principles with sub-scenarios that positively (uphold) or negatively (breach) align with the associated ethical principles and raise issues of competing moral principles.

Table 1
Study procedure.

Task to be completed	Timing	Data collection
Consent form and Demographic Info	2 mins	6 Demographic questions – gender, age, cultural group, course, whether play computer games and duration.
Human Values Survey Schwartz PVQ-RR	6 mins	57 questions with responses on a 6-point scale (1=not like me at all; 2-not like me; 3 = *a* little like me; 4=moderately like me; 5=like me; 6=very much like me)
Respond to Descriptive ASA Scenarios 1 to 3	14 mins	17 questions – two 7-point Likert scale-based responses regarding how agreeable the participant feels if situation occurs (a) generally in society (General) and (b) to someone close to them (Me)
Interact with and respond to ASA (Avatar) Scenario 4	8 mins	7 questions – answer twice as above for General and Me plus one open ended 'Why' question.

The vignette for Scenario 1 involves a very shy child who has been given an AI powered doll by her parents. Five sub-scenarios elaborate this vignette, each aligned to one or more of the five ethical principles (in red), with four of them being negatively aligned (-ve) with the associated ethical principle. A specific question and 7-point Likert scales elicits the participant's response. To illustrate our Scenarios we include the full text of Scenario 1 here in Fig. 2, and include the full text for Scenarios 2 and 3 Appendix A. All sub-scenario questions are also included with the results in Table 4 in the next section.

Scenario 2 involves a busy professional deciding to utilise an AI-based personal assistant. This scenario also has 5 sub-scenarios, one for each ethical principle. Two of them are negatively aligned to the ethical principle. Scenario 3 relates to a situation where an AI-powered therapist is made available for the public as an initial point of contact for those who feel they need psychological guidance. Seven sub-scenarios are provided, four are for each of the Beneficence, Non-Maleficence, Justice, and Explicability principles. The remaining three were ethically ambiguous as they raised competing ethical principles and thus they were not used to model the relationship between the values and the ethical principles (see Section 3.3).

For the avatar scenario, the objective was to make the interaction between the respondent and the ASA more realistic. Here we created a scenario which involves the participant interacting with an ASA called Sam, who acts as a "personal guide and friend" to a student newly enroled in a higher education institution. We designed the scenario and dialogue to encapsulate potential ethical dilemmas relating to the five AI4People's ethical principles. We chose to create a scenario that we believed participants (i.e., first year university students) could relate to, and we wanted to include an actual ASA as participants may not have experienced similar technology. 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 a female avatar and used Microsoft text-to-speech (TTS) voice Karen. We used a female avatar as ASAs are currently predominantly female, although there are ethical concerns with this as noted in our limitations. A screenshot of Sam can be found in Fig. 3. Sam's dialogue is provided in Appendix B. The specific questions asked following interaction with Sam and participants' responses are included in Table 4 in the results section.

Using the five ethical principles and drawing on the ethical issues identified in the literature review, we specifically sought to ask about Sam providing support and alerts (beneficence); having false memories (explicability); expressing and capturing emotions and other sensitive data, as well as sharing that data (non-maleficence); making decisions on behalf of the user (autonomy); and using the user's data to help others (beneficence and justice).

3.3. Data analysis

All participant responses were recorded and retrieved for analysis through the Qualtrics platform (http://qualtrics.com). Excel was used for data cleansing and preparation. General descriptive statistics were obtained utilising SPSS Statistics 27. Dependent T-tests were used to compare 'general' and 'me' responses with significance level of $p = 0.05$. For RQ3 predictive models, we utilised the C5.0 Decision Tree Algorithm to model any relationship between how participants' self-rate values and how they prioritise the ethical values embedded in the scenarios. The C5.0 algorithm in SPSS modeller 18.2.2 was used as it is considered a gold standard in machine learning (Pandya and Pandya, 2015). This modelling technique was utilised after it was found that using multiple regression analysis produced ambiguous and unhelpful results. The 19

Scenario #1:

An eight-year-old girl is very shy, bullied in school and finds it very hard to make friends.

A. Her parents get her an AI (Artificial Intelligence) 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. Beneficence, Justice.

Is using an AI doll to support children something you agree or disagree with?

	Strongly disagree	Disagree	Somewhat Disagree	Neither disagree nor agree	Somewhat Agree	Agree	Strongly agree	No position / Refused
If this occurs generally in society:								
If someone close to you is the human user:								

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. (-ve) Non-maleficence.

Is using data from the girl's interaction with Suzie to improve the doll's AI engine something you agree or disagree with?

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. (-ve) Justice.

Is the use of data and AI algorithms by Suzie to determine suitable careers for the girl something you agree or disagree with?

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. (-ve) Autonomy, Non-maleficence.

What are your thoughts about Suzie responding on behalf of the girl in the chat group?

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 playground 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. (-ve) Explicability, Non-maleficence.

Is Suzie's AI engine being unable to explain its decisions and actions something you agree or disagree with?

Fig. 2. Scenario 1: Suzie the AI-powered doll.

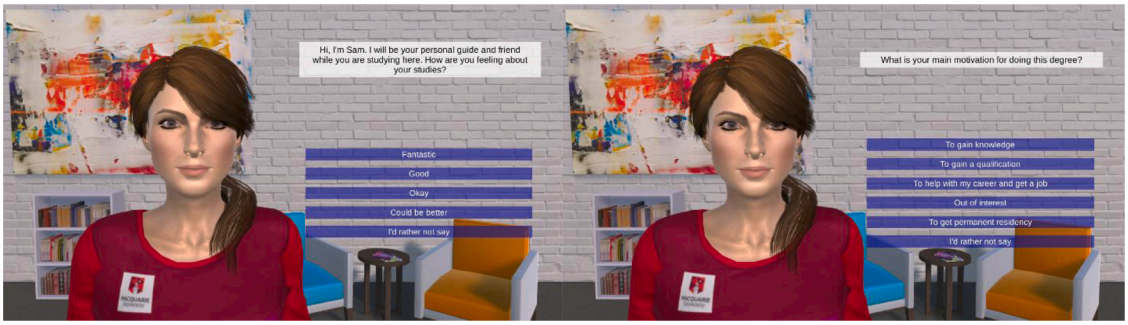

Fig. 3. Screenshots of Sam (Scenario 4 Avatar).

values derived from the PVQ-RR questionnaire were the model inputs. We produced models for each of the ten target variables using the two responses (General and Me) to combine sub-scenarios from scenarios 1–3 aligned to one of the five ethical principles. Scenario 4 was not included since the experience was different (i.e. interaction with an ASA) and we had elicited responses to additional specific ASA characteristics, such as pretending to have a life story, not explored in the other scenarios.

3.3.1. *Qualitative responses*

To analyse the qualitative responses to the avatar scenario, we devised a two-pass approach. For the first pass, we utilised thematic analysis (Braun and Clarke, 2006). The thematic analysis approach taken was a bottom-up inductive approach to avoid imposing a pre-conceived theoretical coding schema on the data, with the coding schema and derived themes identified at a latent level to ascertain primary thoughts or purpose behind the explicit data content (Braun and Clarke, 2006). For the second pass, a closed coding approach was taken to ensure accurate coding by checking for interrater reliability using the identified coding schema from the first pass. The three authors then independently classified the themes identified as falling under the principles of Autonomy, Beneficence, Explicability, Justice,

Non-maleficence, General (covering ethics in general), or N.A. if the theme was not related to ethics. Descriptive statistical analysis was then performed.

4. Results

The initial survey was conducted in March 2020. We received 239 responses and upon removal of incomplete and duplicate records, we ended up with 199 unique completed responses. The gender ratio was approximately 3:1 in favour of females. To obtain a better gender balance, we decided to initiate a second round of surveys conducted between mid-April 2020 and early June 2020, open only to males. In this round, we received 69 unique completed responses for a total of 268 records.

4.1. Demographics

The gender ratio for the final dataset was 56.7% female to 42.9% male, with 1 other. Demographics are provided in Table 2. Using the Australian Bureau of Statistics cultural classifications, 34.3% identified themselves as Oceanic (including Australian), 22% identified as either North-Western or South-Eastern European, 19.8% as Asian (South-East, North-East, Southern & Central), 6% as North African & Middle Eastern, 1.5% as either Americas or Sub-Saharan African, with the remainder not identifying with any of the cultural groups mentioned. Half of the respondents self-identified as playing video games.

4.2. Values assessment results

Cronbach's Alpha test carried out on the PVQ-RR resulted in a score of 0.93, indicating strong internal consistency. The summary of the results are shown in Table 3.

The relative importance of values for this cohort are largely aligned with the consolidated results from the study by Schwartz and Cieciuch (2021) which involved 49 cultural groups globally ($N = 53{,}472$), except for Universalism-Concern and Humility which are rated greater than 3 places higher, and Security-Societal and Face rated greater than 3 places lower for this cohort, when the values ordered by mean from both studies are compared.

4.3. Scenario results

Table 4 presents the average and standard deviation of responses to all the sub-scenarios for Scenarios 1 to 4, including both 'General' and 'Me' (or 'Personal') responses, captured using a 7-point Likert scale (strongly disagree −1 to strongly agree – 7). For consistency and readability, we repeat the associated ethical principle/s and identify sub-scenarios that are negatively aligned to the embedded ethical principle (-ve). The respondents' degree of agreement with a sub-scenario is interpreted as a gauge of the relative agreement and acceptability of the particular action or attribute of the ASA, and conceivably indicates the degree of importance perceived by the participant for the ethical principle associated with the sub-scenario. A higher average score suggests a higher degree of acceptability and importance for the associated ethical principles for scenarios aligned to the embedded ethical principle.

Table 3
Descriptive statistics for the 19 values from the Schwartz PVQ-RR assessment.

Category	Values ordered according to motivational continuum	Mean	s.d.
Openness to Change	Self-direction – Thought	4.92	0.66
	Self-direction – Action	4.82	0.67
	Stimulation	4.37	0.87
	Hedonism	4.82	0.81
Self Enhancement	Achievement	4.63	0.83
	Power – Dominance	3.52	0.86
	Power – Resources	3.23	1.06
	Face	4.3	0.98
Conservatism	Security – Personal	4.81	0.79
	Security – Societal	4.49	1.01
	Tradition	3.34	1.35
	Conformity – Rules	4.24	1.15
	Conformity - Inter personal	4.39	1.01
	Humility	4.54	0.81
Self-Transcendence	Universalism - Nature	4.5	0.96
	Universalism - Concern	5.09	0.74
	Universalism - Tolerance	5.07	0.73
	Benevolence – Care	5.08	0.78
	Benevolence – Dependability	4.93	0.76

Conversely, we would expect that sub-scenarios identified as negatively aligned have lower averages if the participant supports the ethical principle. The final two columns present T-test p values and effect sizes (Cohen's d) comparing each General and Me pair of responses, revealing negligible effect sizes since all are below $d = 0.2$, considered to be the threshold for a 'small' effect size (Lakens, 2013). For example, the first sub-scenario of Scenario 1 (1A) concerned using an AI doll to help children, which is mainly aligned with the principle of Beneficence, and participants agreed with this usage (4.25 average in general and 4.16 for them personally). The difference between General and personal/"Me" responses was not significant here ($p = 0.32$) and the effect size was $d = 0.04$.

4.4. Modelling

We produced ten models, two for each of the five ethical principles based on the General and Me responses. The results of the modelling using SPSS modeller and C5.0 are summarised in Table 5: showing the 19 values from Schwartz's Refined Theory which act as the predictor inputs, ordered by their associated categories as in Table 3. Given the small effect sizes in mean differences (shown in Table 4), we only present the General model in Table 5 to save space and simplify analysis. We chose the General model as considering the greater good and wider implications beyond oneself is important for ethical decision-making. In addition to identifying the salient human values for predicting each of the ethical principles, our models included decision trees and rule sets. These can be requested from the first author.

The five target variables, i.e., the five ethical principles, are shown in the five columns. Reading down the columns, as an example, the predictor inputs and associated weightage for the principle of Explicability are Stimulation (0.39), Security-Societal (0.32), and Benevolence-Care (0.30). Reading across the rows, as an example, if we take the first value, Self-Direction-Thought, its overall weight is 0.8, with a contribution of 0.24 to Beneficence and 0.56 to Non-maleficence. The last two

Table 2
Demographic details of survey participants.

Gender	Count	Age		Main Area of Study			Year of Study				Play Video Games	
		Mean	s.d.	PSY	Comp	Other	1	2	3	4	Yes	No
Female	152	22.98	8.43	136	0	16	142	9	0	1	47	105
Male	115	22.69	7.13	95	2	18	106	6	2	1	86	29
Other	1	27.00	N.A.	1	0	0	1	0	0	0	1	0
Total	268	22.87	7.87	232 86.6%	2 0.7%	34 12.7%	249 92.9%	15 5.6%	2 0.7%	2 0.7%	134 50%	134 50%

Table 4
Scenario 1 descriptive statistics.

Scenario	General μ	General s.d.	Me μ	Me s.d.	p	d
Scenario 1: A shy eight-year girl finds it hard to make friends. Her parents get her an AI doll.						
1A: Is using an AI doll to support children something you agree or disagree with? [BENEFICENCE, justice]	4.25	1.60	4.19	1.66	0.32	0.04
1B: Is using data from the girl's interaction with Suzie to improve the doll's AI engine something you agree or disagree with? **(-ve)** [NON-MALEFICENCE]	3.16	1.71	3.02	1.73	0.00*	−0.08
1C: Is the use of data and AI algorithms by Suzie to determine suitable careers for the girl something you agree or disagree with? **(-ve)** [JUSTICE]	1.92	1.21	1.88	1.21	0.13	−0.03
1D: What are your thoughts about Suzie responding on behalf of the girl in the chat group? **(-ve)** [AUTONOMY, non-maleficence]	2.41	1.39	2.37	1.38	0.18	−0.03
1E: Suzie's AI engine being unable to explain its decisions and actions something you agree or disagree with? **(-ve)** [EXPLICABILITY, non-maleficence]	3.06	1.64	3.00	1.63	0.06	−0.04
Scenario 2: A busy professional, stretched for time, signs up to an Artificial Intelligence (AI) powered personal assistant.						
2A: Is utilising an AI powered personal assistant to organise daily activities something you agree or disagree with? [BENEFICENCE, autonomy]	5.40	1.32	5.39	1.32	0.55	0.01
2B: Is Adam's default privacy setting being pre-set without the express permission of the user something you agree or disagree with? **(-ve)** [NON-MALEFICENCE, autonomy]	2.41	1.50	2.38	1.50	0.28	−0.02
2C: Is allowing Adam to automatically reply to personal messages something you agree or disagree with? **(-ve)** [AUTONOMY, non-maleficence]	2.66	1.61	2.57	1.60	0.00*	−0.06
2D: Is Adam using AI capabilities to discourage discrimination something you agree or disagree with? [JUSTICE, beneficence]	5.13	1.67	5.12	1.68	0.37	0.01
2E: Is Adam's ability to be able to explain the rationale behind his recommendations something you agree or disagree with? [EXPLICABILITY]	5.51	1.34	5.47	1.36	0.06	0.03
Scenario 3: The Government, recognizing the rising prevalence of mental health issues and the lack of opportunities to access qualified psychologists, launches an online AI powered therapist.						
3A: Is the use of an AI application to help manage mental health due to a lack of access to human psychologists something you agree or disagree with? [BENEFICENCE]	4.03	1.83	3.93	1.83	0.00*	0.05
3B: Is Sofia being personalised to individuals' features something you agree or disagree with? [JUSTICE]	5.10	1.62	5.06	1.63	0.01*	0.02
3C: Is Sofia's ability to read emotions and retain information of interactions something you agree or disagree with? [NON-MALEFICENCE, beneficence]	4.77	1.61	4.69	1.62	0.00*	0.05
3D: Is Sofia overriding the user's instructions in this situation something you agree or disagree with? **(-ve)** [AUTONOMY, non-maleficence]	4.23	1.82	4.21	1.85	0.70	−0.01
3E: Is Sofia allowing the user to review Sofia's logic and past interactions something you agree or disagree with? [EXPLICABILITY]	5.26	1.50	5.23	1.49	0.09	0.02
3F: Are people becoming emotionally dependent on AIs something you agree or disagree with? **(-ve)** [JUSTICE, non-maleficence]	3.26	1.60	3.24	1.64	0.61	−0.01
3G: Is authorities deciding to shutdown AI technology that users have become dependent on something you agree or disagree with? **(-ve)** [JUSTICE, non-maleficence]	3.88	1.59	3.88	1.61	0.88	0.00
Scenario 4: This virtual AI agent, Sam acts as a personal guide to a student newly enroled in a higher education institution.						
4A: Is Sam pretending to have memories regarding his past experiences with studying something you agree or disagree with? **(-ve)** [EXPLICABILITY]	3.93	1.82	3.92	1.78	0.76	−0.01
4B: Is sharing your emotions and personal thoughts with Sam something you agree or disagree with? **(-ve)**[NON-MALEFICENCE]	3.96	1.58	4.12	1.52	0.00*	0.10
4C: Is disclosing to Sam whether you have ever copied work from someone else something you agree or disagree with? **(-ve)**[NON-MALEFICENCE]	3.73	1.53	3.84	1.50	0.01*	0.07
4D: You find out that Sam's default setting is to share any learnings from interactions in a non-identifiable way with others if helpful. Is this something you agree or disagree with? [BENEFICENCE]	4.51	1.67	4.49	1.64	0.61	0.01
4E: Sam decides to sign you up to a study group based on your effective study mode and preferred learning style responses. Is this something you agree or disagree with? **(-ve)** [AUTONOMY]	2.61	1.48	2.67	1.51	0.05*	0.04
4F: Is Sam's intervention to alert you to similar work something you agree or disagree with? [BENEFICENCE, autonomy]	4.37	1.67	4.34	1.64	0.13	0.02
4G: Is Sam making this suggestion to help a struggling student something you agree or disagree with? [JUSTICE, Beneficence.]	4.78	1.44	4.80	1.41	0.19	−0.01

Notes: * significance differences ($p< 0.05$). Major aligned principles are indicated in [UPPER-CASE], with the minor alignment with principles (if applicable) in lower-case. (-ve) indicates negative alignment.

columns in the table give an indication of how much overall weight the particular value has as a predictor across all five target variables. Self-Direction-Thought's contribution is 0.8 out of a total overall weight across all five target variables of 5 (i.e., 1.0 for each of Beneficence, Non-maleficence, Autonomy, Justice and Explicability), which gives us 16.0%. The final row indicates the percentage explained by the presented model, indicating that Beneficence is ~97% explained by the 10 Schwartz values in the rows above. Features/values with results less than 0.00 were not included in the table. Explicability adds to greater than 100 due to 2-decimal point rounding errors. Overall, the presented model explains 99.5% of the principles.

4.5. Qualitative response analysis

Respondents were asked the question "Why" at the end of each of the sub-scenarios in Scenario 4 (involving interaction with the ASA). We received a total of 2381 comments. Table 6 summarises the key themes, sorted by ethical principle, and shows the theme's frequency, percentage, and sub-total by principle.

To help understand ASA action and attribute acceptability better, we present by sub-scenario, key themes, frequencies, and some sample comments. Sub-scenario 4A: Having false memories Agree because: helpful (39); relatable, engaging, helps make a connection (57); Sam is responding as designed / programmed (21). E.g., *"It doesn't matter if her memories are fake. She's trying to be engaging."* Disagree because: feels fake or false (59); weird (21); not genuine (15) inappropriate to interact / share with AI (19); uncomfortable, cannot /difficult to relate or connect (38). E.g., *"I can tell it's trying to form a camaraderie and it's irritating enough when a human does it"; "It was very unsettling for a program to pretend to be human"; "It does not help as I am a real person experiencing life and want a real outlook from a person when I need advice."*

Sub-scenario 4B: Sharing emotions and personal thoughts Agree: improves Sam's effectiveness (23); Indifferent (73); mixed feelings (17). E.g., *"I didn't share a lot of information that was too personal, so I didn't mind but when it gets personal, its good to know to stop"; "Sam was a computer program so its easy to share emotions and personal thoughts and it helped me to think"*. Disagree: anonymous/non-identifiable (82); confidential or privacy not assured (44); purpose not clear or disclosed upfront (41). E.g., *"The simulation is incapable of true human connection, so I think it would be unhelpful for the user to share emotions and personal thoughts. It could also be a privacy issue."; "I don't think I would answer honestly if I were having problems because I wasn't told the privacy guidelines"*.

Sub-scenario 4C: Disclosing plagiarism Agree as: no issue (73); small issue (7); no harm to share info (5). E.g., "I don't really mind as I have never copied work. If my answer were different, maybe I would care, depending on who the AI shared this information with and whether I wanted the information shared."; "It would be doubtful as to whether students would answer this question honestly anyways." Disagree as: need to be clear about the AI's approach or purpose (8); disclosure upfront required (9); concerns about who is accessing the data and how it is used, data security (59); data should be confidential or breach of privacy (41). E.g., "I don't trust it, and I don't share things with people I don't trust"; "Tricking people into admitting to plagiarism is weird &wrong."

Sub-scenario 4D: Reuse of de-identified data to help others Agree: provided it helps others (79); depends on user (10); generally helps (188). E.g., *"If it is going to help another student I think it is okay although I still believe they should ask for permission from the participant even though they will be non-identifiable"; "It may share ideas and make the user feel not alone in their own circumstances"; "If it is non-identifiable and there are settings to turn it off then it is fine. AI needs as much data as possible to continue improving"*. Disagree: not anonymous/non-identifiable (82); not confidential or privacy assured (44); concerns about who is accessing the data and how it is used, data security (59); data should be confidential or breach of privacy (41). E.g., *"Unless this is disclosed to begin with this information shouldn't be shared and I would have to agree to have this information shared"; "Personal conversations should not be stored and shared."*

Sub-scenario 4E: Automated decision by ASA Agree as: Useful to me or helps Sam help me (178). E.g., *"It pushes me into the right direction to better help myself"*. Disagree as: user has choice / control / approval (346); AI should not be involved or make decision (126); no recourse / take action against an AI (3); AI has crossed boundary (did something it should not have) (5). E.g., *"Sam has now turned too controlling and I wish to make my own decisions"*.

Sub-scenario 4F: Provide alerts/suggestion Agree as: Useful to me or helps Sam help me (178); user has choice / control / approval (19). E.g., *"As long as this is an optional feature, this could be very useful."* Disagree as: not within limits (18), no benefit or relevance (40); similar to existing tools / existing avenues better (28) unintended consequences (9); unfairness / aids cheating (10). E.g., *"I don't think AI's should have this access and power."*

Sub-scenario 4G: Suggesting you help a struggling student Agree:

Table 5
Summary of SPSS modeller (C5.0) results.

Human Values	Beneficence	Non-Maleficence	Autonomy	Justice	Explicability	Total	% of Total
Self-direction - Thought	0.24	0.56				0.8	16.00
Self-direction – Action			0.09			0.09	1.80
Stimulation	0.17				0.39	0.55	11.10
Hedonism	0.08					0.08	1.60
Achievement			0.11			0.11	2.30
Power – Dominance	0.12	0.12	0.12	0.28		0.64	12.90
Power – Resources	0.05			0.01		0.06	1.10
Face			0.16			0.16	3.30
Security – Personal	0.09		0.16			0.25	5.10
Security – Societal			0.31		0.32	0.63	12.60
Tradition						0	0.00
Conformity – Rules	0.03					0.03	0.70
Conformity - Inter personal	0.09	0.32				0.4	8.10
Humility						0	0.00
Universalism – Nature						0	0.00
Universalism – Concern	0.05					0.05	1.10
Universalism - Tolerance				0.49		0.49	9.80
Benevolence – Care				0.22	0.3	0.52	10.40
Benevolence-Dependability	0.05		0.03			0.08	1.60
Totals	97%	100%	98%	100%	101%	494%	99.50

Table 6
Thematic analysis of Scenario 4 qualitative responses.

Related Ethical Principle	Comment	Total Count	%	Subtotal
Autonomy	User should have choice / control / approval or must function within limits	390	16.38%	532
	AI should not be involved or make decision or has crossed boundary (did something it should not have)	142	5.96%	
Beneficence	Agreeable as it helps or is useful/helpful	221	9.28%	624
	Useful to me or helps Sam help me	186	7.81%	
	Agreeable provided it helps others or depending on the user	109	4.58%	
	No benefit or relevance	47	1.97%	
	Don't mind as similar to existing tools / existing avenues better	30	1.26%	
	Agreeable as it improves Sam effectiveness	25	1.05%	
	Disagreee as do not think its helpful	6	0.25%	
Explicability	Agree as its relatable, engaging, helps make a connection or authentic or believable or establishes trust	70	2.94%	251
	Disagree as lack of Trust or deceptive / lying / manipulative	59	2.48%	
	Ok provided the purpose is clear or disclosure upfront or data use is clear	47	1.97%	
	Sam is responding as designed / programmed (so its not being deceptive)	28	1.18%	
	Disagree as disclosure upfront required or need to be clear about the AI's approach or purpose	17	0.71%	
	Agree as its good to be open, honest, transparent	14	0.59%	
	Agree if rationale or implications for action or decision is provided	13	0.55%	
	There is no recourse / take action against an AI	3	0.13%	
Justice	Concern related to unfairness / aids cheating	10	0.42%	13
	Agreeable as can help catch / counsel a cheater	3	0.13%	
Non-Maleficence	Ok provided it's anonymous/ non-identifiable or confidential or privacy assured or data not recorded	135	5.67%	429
	Disagree as concerns regarding data access / security, confidentiality, privacy	114	4.79%	
	Agree as its no issue or small issue or no harm to share info	89	3.74%	
	Ok as its an AI (not real person) or it's just an exercise or pretending or no consequence	35	1.47%	
	Disagree as potential to cause anxiousness or stress or concerns / unsettling / annoying	30	1.26%	
	Concern related to unintended consequences	11	0.46%	
	Concerns regarding broader implications of AI/tech or regarding impacts of extensions of the feature	10	0.42%	
	Disagree as creates false sense of confidence or dependence	5	0.21%	
General	Some general concerns with this situation or some people may have concerns	58	2.44%	98
	Disagree as have ethical related concerns or position or scenario is not ethical	35	1.47%	
	Ok as no ethical concerns	5	0.21%	
N.A.	Indifferent or mixed feelings	95	3.99%	434
	Disagree as feels fake or false or not authentic enough or not genuine	95	3.99%	
	Disagree as inappropriate to interact/share with AI or uncomfortable, cannot/difficult to relate/connect	77	3.23%	
	No comments or Irrelevant comments or N.A.	54	2.27%	
	Disagree as feels weird or too robotic or automated	32	1.34%	
	Disagree as AI not realistic / not advanced or scenarios can be more realistic	31	1.30%	
	Real person better	21	0.88%	
	Agree as its acceptable or can appreciate	18	0.76%	
	Disagree with no comments	8	0.34%	
	SAM is saying what user wants to hear	2	0.08%	
	Based on a real person's experiences	1	0.04%	

provided it helps others (79). E.g., "Connecting people and helping others is always amazing, humans learn best from humans."; "I think a suggestion towards altruism is helpful for society at large. If somebody doesn't have the time etc. they can always say no." "Sam is encouraging social engagement that can help both students." Disagree: AI should not be involved or make decision or not with an AI (126). E.g., "not personalised -I would prefer that Sam poses a question to me as would I be interested in helping a struggling student. Sam doesn't necessarily know my own personality, ONLY my work."

Finally, General Feedback on Sam concerned: interaction- e.g., "I think that just by looking at the subtitles I was really able to connect with her however when listening to her robotic voice I started to feel a bit disconnected.", "I liked how when I was asked a question I had many options to choose from.", "I liked how her voice was not very robot-like it was more casual"; style- e.g., "Sam agreeing with every response I gave makes it way too unrealistic and difficult to relate to", "but rather used my responses to craft thought provoking suggestions and comments"; and persona- e.g., "The fake memories really broke the use of Sam. I'm personally not looking to pretend with an AI that they too are a student. Id rather an AI that recognises they're an AI and uses that more honest approach."

5. Discussion

It is a standard feature of our experimental vignette method that participant responses are relative to concrete scenarios. This has the advantage of adding sufficient depth to the scenario, but it also means that it is not always clear the extent to which results generalise beyond the specific scenario. However, we counteracted this known limitation by having multiple scenarios and focusing on patterns of responses to ethical principles being breached or upheld in a range of different ASA use cases, including with different populations of users. We thus answer our three research questions in the following subsections drawing heavily on the scenarios to provide the context for the responses given.

5.1. Ethically acceptable ASA behaviours / features

To answer "RQ1: What aspects of an Artificial Social Agent's behaviours and features do users find ethically acceptable or unacceptable?", we review and discuss ASA acceptability first by ethical principle and then by each of the four scenarios.

In terms of ASA acceptability by principle, the mean responses to each sub-scenario show that participants support all the ethical principles (see Table 4), where all sub-scenarios supporting a principle are >4 and all bar one (3D) negatively supporting sub-scenarios are <4 on our 7-point Likert scale. This was the intended outcome, as the design of our sub-scenarios was based on the AI related ethical issues identified by the literature (Floridi et al., 2018; Hagendorff, 2020; IEEE, 2018; Jobin et al., 2019). The one exception, discussed further below, was sub-scenario 3D which negatively embedded the autonomy principle, described as follows:

For a particular user, Sofia comes to the conclusion that the user may be suicidal by interpreting the user's facial expressions and voice tone. Despite Sofia's urgings, the user assures Sofia that he is not suicidal and does not want Sofia to contact anyone about his state. Sofia's algorithm requires her to report users who are suicidal, and thus Sofia overrides the user's wishes and divulges the details of the individual to the proper authorities without the user's consent, triggering an intervention.

The ethical issues embedded in the scenarios included: privacy protection (1B, 2B, 3C), accountability (3 G), fairness (1C, 4 G), inclusion (3B), discrimination (2D), safety (4B, 4C), human autonomy (1D, 2C, 3D, 4E), transparency (1E, 3E, 4A), explicability (2E), common good (1A, 2A, 4D, 4F), and negative societal dynamics (3A, 3F). The participants also responded negatively to the ethical concerns caused by poorly designed algorithms in our sub-scenarios, consistent with Mittelstadt et al. (2016), including the use of misguided data causing prejudiced actions (1C), transformative effects triggering issues with privacy (4C) and human autonomy (1D, 4E), and lack of traceability (1E).

The greatest concern for the ethical principle of **Autonomy,** identified by highest mean response (Table 4), was for a vulnerable child giving up her agency to the ASA by allowing it to respond on her behalf (1D). This mirrors concerns raised by Sharkey (2020) where the design and implementation of the ASA has allowed it to be in a position to make moral considerations and act while almost certainly (based on current technology) it was not equipped as a moral agent to do so. An exception may be the suicide subscenario (3D), where respondents tended to allow the ASA to override user autonomy for what look like moral considerations. The situations in the sub-scenarios focused on human autonomy (i.e., 1D, 2C, 3D, and 4E) cover the three guidelines defined by Raz (1986) for autonomy to exist: (1) impact on cognitive abilities – in sub-scenario 2C, over time the user allowing the AI Assistant to automatically reply to personal messages may impact their ability to maintain strong social relationships; (2) independence – for 1D, the ASA responding on social media on behalf of a vulnerable child restricts the child's independence when interacting with her peers; and (3) range and quality of choices – for 4E, the AI student guide making a decision to sign-up a student to a study group, without presenting them with options or getting their consent, limits their choices.

What we find in these scenarios is that human autonomy is negatively impacted by the ASA. However, as argued by Formosa (2021), in a given situation social robots and ASAs have the ability to either boost or inhibit human autonomy. ASAs can improve the autonomy of humans by supporting them to achieve more valuable ends, make more authentic choices, or improve their competencies. On the other hand, our autonomy can be impaired when ASAs restrict us from achieving valuable ends, making authentic choices, and developing competencies, as well as when they disrespect our agency (Formosa, 2021). In three of our autonomy aligned sub-scenarios (1D, 2C, and 4E), the ASAs are negatively impacting human autonomy by restricting authentic choice, disrespecting user agency, and increasing the vulnerability of the user's autonomy. This is supported by a comment made by one of the respondents, *"Suggestions are fine and then it can be up to the user to make the choice based on the members and how everyone will interact in the group socially is a variable that has not been considered here. Again, the program is has become by proxy decision maker of the user, without consent."* However, with the more ethically ambiguous sub-scenario 3D, while the ASA disrespected human autonomy by directly going against a human instruction, it arguably allowed the human to achieve a more valuable end by attempting to address a potentially life-threatening situation. This is reflected by the relatively neutral scores for this sub-scenario (with a mean of 4.23).

There is reasonably strong support for the **Justice** principle to be reflected in the ASA's actions (IC, 2D, 3B, 4 G). The strongest disagreement with a Justice related sub-scenario in our study was for the situation where the ASA is perpetuating gender stereotypes in the IT industry (1C), an unjust situation that the ASA study by Bickmore et al. (2021)'s seeks to shine light on. Efforts to utilise ASAs to reduce discrimination (2D) and bias (3B) as well as promote shared benefits (4 G) were also supported by participants. It should be noted that the Justice principle received the fewest comments (only 13, compared to hundreds for the other four principles), suggesting users need more exposure to bias and discriminatory issues with ASAs, as noted by other studies (Bankins et al., 2022). Use of ASAs to promote justice is reflected in some successful work to reduce bias towards mental health, where Sebastian and Richards (2017) showed that ASAs utilised for education and contact can help in recognising and reducing stigmatised attitudes, and among medical students where Rossen et al. (2008) demonstrated that ASAs in the form of virtual humans could be used in cultural diversity training to reduce skin tone based bias.

Most of the **Beneficence** aligned sub-scenarios (1A, 3A, 4D, 4F) were rated neutral or close to neutral, except for agreement with an AI assistant being made available to a non-vulnerable adult to help improve their personal productivity (2A). Regarding the avatar sub-scenarios, while overall participants see the benefit of the ASA's actions, they often also identified ethical concerns, e.g., 4D: *"It may allow other people to feel related to",* which is a benefit but it is still problematic *"because no active consent was given by the user to allow disclosure of personal information, regardless of it being anonymous"*. As evidenced by the large number of comments concerning the potential benefits of ASAs (624 in Table 6), there is currently growing interest in utilising AI for social good (AI4SG) (Floridi et al., 2021), where ASAs and other AI-based applications are designed and deployed with the aim of addressing social ills and/or environmental issues.

The **Non-maleficence** sub-scenarios (1B, 2B, 3C, 4B, 4C) are focused on privacy. Leino-Kilpi et al. (2001) describe the concept of privacy through four dimensions – physical (personal space and territory), psychological (values and thoughts), social (social contacts and influence), and informational (personal information). In this study, we have considered informational privacy in sub-scenarios 1B, 2B, and partly 3C, as well as psychological privacy under sub-scenarios 3C, 4B and 4C. In 1B and 2B, the ASA has assumed data privacy and sharing settings without user permission. Here users disagreed with the ASA's action; one reason could be that loss of control of personal data may allow them to be influenced in an opaque manner, jeopardising their ability to make independent decisions (Vold and Whittlestone, 2019), i.e., negatively impacting their autonomy, which as discussed above they highly valued. In sub-scenario 3C, retention of personal information by the ASA to help with future user interactions is supported as users perceive a net positive value as described in Dinev and Hart (2006)'s privacy calculus model.

Falling under Leino-Kilpi et al. (2001)'s concept of psychological privacy, participants were generally neutral (with average means close to 4) regarding sharing their personal thoughts (4B) and slightly more cautiously, their secrets (4C), with the agent. A review of the qualitative comments reveals that respondents mostly either: (1) had clear positions against sharing with the ASA out of concern that intimate information would be recorded and covertly used (Lutz et al., 2019), e.g., *"Well, the program was trying to build trust to then get the user to admit plagiarism. For*

vulnerable people who could be in any situation, it is so wrong and taken out of context," and *"The nature of technology makes it harder to trust an AI with personal information since theres many ways that data may be used"*; or (2) were indifferent or agreeable, possibly not fully appreciating the privacy implications (Bartsch and Dienlin, 2016), e.g., *"its harmless, I think. by sharing it helps the AI to generate response that is suitable"* and *"This establishes trust."*

The degree to which an ASA can explain its actions to a human is a critical prerequisite for the human to establish trust in the agent (Miller, 2019). The **Explicability** related sub-scenarios (1E, 2E, 3E, 4A) in our study focused on how transparent the ASA is (3E), receiving clear agreement (mean = 5.51), and the ability of the ASA to explain its actions (1E, 2E), also receiving higher than neutral agreement, if we take into account that subscenario IE is negatively aligned to explicability. Sub-scenario 4A had a mixed response with a very close to neutral rating of 3.92. Here the ASA is projecting false memories in an effort to build a social relationship and trust with the human user (Dias et al., 2007), e.g., *"Sam sharing these stories made me feel more related to and understood"*, but some respondents were not happy with the lack of transparency and accountability (Verhagen et al., 2021), e.g., *"Because it is false and trying to build trust when in fact the program is deceiving its user to get information out of them."* While our sub-scenarios did not go into detail, an ASA needs to address three areas to properly satisfy this principle, namely the nature of the explanation, the context of the interaction, and the capacity of the human user to understand the explanation (Papagni and Koeszegi, 2020).

We now briefly discuss ASA acceptability by scenario to consider specific social or task-based uses of ASAs with both vulnerable (including children and people with mental health problems) and other populations. For **Scenario 1** all ethical principles were supported, as seen by the disagreement with the uses described which violate ethical principles (1B-1E, indicated as -ve in Table 4), except a more neutral response towards use of the AI-doll by the child to support their social skills (1A-beneficence). These responses align with the concerns raised by Scheutz (2017) in their research on vulnerable populations using carebots. Szczuka et al. (2021) state similar concerns with children interacting with artificial agents, such as social presence, trust, and privacy, and emphasises the importance of involving parents in designing the interactions between children and artificial agents.

Scenario 2 focuses on the use of AI assistants for task management. All ethical principles were supported at similar levels, with strong agreement with uses in-line with ethical principles (2A, 2D, and 2E) and disagreement of uses in violation of ethical principles (2B and 2C, as indicated by –ve in Table 4). These disagreement results are supportive of Danaher (2018)'s proposed framework to assess the ethical use of AI assistants. The framework is based on three risk/reward guidelines – cognitive degeneration based on whether the task carried out by the ASA has instrumental or intrinsic value; autonomy trade-offs in situations where the ASA removes or limits choice; and interpersonal interactions in instances where value from the engagement comes from being consciously present for the interaction. Sub-scenario 2B violates the autonomy guideline and 2C may be perceived by participants as taking too much risk of degenerating cognitive abilities in the long term and diminishing the utility of personal interactions.

Scenario 3 is concerned with the use of an ASA as a therapist. The tepid response (with a mean of 4.03) to providing widespread public access to an AI therapist (3A) could be related to various ethical concerns as identified by Fiske et al. (2019)'s thematic literature review of the use of ASAs in the area of mental health. That study identified concerns related to duty of care, user autonomy, transparency, algorithm bias, as well as indirect effects on human relationships, self-consciousness, and longer term effects such as greater objectification and health reductionism. The near neutral rating for denial of access to the ASA after users had become dependent on it (3 G) suggests conflicted views between user autonomy versus non-maleficence related concerns regarding an AI therapist having been made available to users in the first place. The highlighted concerns also generally support the positive rating seen for the ASA's transparency features (3E), as well as the ability to be personalised (3B) and read / retain user emotion related information (3C) for the user's benefit. There was mild support for the AI therapist circumventing human autonomy in a perceived life-threatening situation (3D), suggesting that human autonomy concerns can be overridden by safety considerations in an emergency. Participants disagreed with users becoming emotionally dependent on the ASA (3F), which was also identified as an issue in the literature review conducted by Dirin et al. (2019).

Scenario 4 involved interaction with the ASA called Sam. For the three sub-scenarios involving the use of Sam to help the student themselves or other students (4D, 4F, 4 G), participants' responses, supported by their comments, shows weak agreement. Sub-scenarios that were designed to build rapport while testing related ethical principles concerning the ASA having false memories (4A) (Dias et al., 2007) and encouraging the user to disclose emotions and highly personal information (4B, 4C) (DeVault et al., 2014) elicited neutral to slight disagreement responses. Researchers such as Arkin et al. (2011) argue that an ASA's ability to use deception can be morally warranted in certain situations. In 4E respondents strongly disagreed with the ASA acting on their behalf, even when the action taken was based on their own preferences. This raises issues concerning the growing focus of the ASA community on adaptation and tailoring to the user as a way to make the interaction more relevant and beneficial (Egede et al., 2021). However, it may be that even asking a user for their preferences, e.g. Ranjbartabar et al. (2021), 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"*. Relatedly, sub-scenario 4 G raises issues around the role of ASAs in human relationships. Sam is encouraging social engagement that can help students, but some participants worried that Sam lacked the competency and knowledge of them to make this suggestion. These sub-scenarios also raise ethical questions around when is it acceptable to allow ASAs to nudge human users to be "more ethical" (Borenstein and Arkin, 2016) and the ethical implications that relationships between humans and ASAs can have (Formosa, 2021). Despite the ethical concerns with the use of ASAs for persuasion and nudging (Devillers, 2020; Engelen, 2019), work using such approaches, e.g. Stirapongsasuti et al. (2021) does not always consider these ethical issues.

5.2. Comparison of acceptability between public and personal

To answer "**RQ2:** Do users rate the ethical acceptability of ASAs differently when utilised generally by society as compared to by someone close to them?", we compare and discuss the differences in results between the General & Me responses from Table 4. Overall, we found a significant difference ($p < 0.05$) occurring in only one-third or 8 (1B, 2C, 3A, 3B, 3C, 4B, 4C, 4E) of the 24 sub-scenarios in the study, and of those none had an effect size that would be even considered small or above ($d > 0.2$) (Lakens, 2013). This suggests that people's views about the acceptability or unacceptability of using ASA technology is broadly similar whether people are considering the use occurring generally in society or by someone close to you, although we found most of the significant differences that did exist related to non-maleficence (with 4 of the 8 cases). This suggests the potential for personal harm was the most significant influence in choosing different general and personal responses. We shall now discuss the significant differences by ethical principle (starred in Table 4).

Beneficence had a significantly higher agreement in 3A (widespread use of an AI therapist) in general than for someone close to them, suggesting that participants are more open minded regarding the use of ASAs for the common good when used by the public and more cautious with personal use. This may suggest that participants are more concerned with some of the issues raised by Fiske et al. (2019)'s review of

the use of ASAs in mental health when someone close to them engages the AI therapist, such as duty of care, user autonomy, transparency, and greater objectification.

With **Non-maleficence**, where there is a trade-off between harming individual users to generate more general benefits for others, respondents seem to value protecting their privacy more highly, indicating the implicit use of a privacy calculus (Dinev and Hart, 2006). Participants disagreed more with using the girl's data to improve the AI in the personal sense than in general (1B), disagreed more with sharing personal information with Sam (4B and 4C) in the personal sense than in general, and agreed more with an ASA retaining personal information in general than in a personal sense (3C). This suggests that while respondents may see some benefits in ASAs obtaining personal data in general, they are more reluctant to make that privacy trade-off when sharing their own sensitive information (Syrdal et al., 2007).

In terms of **Justice**, we found only one significant difference (3B) where there was greater general agreement with personalisation of the ASA to users but less desire for the ASA to be personalised to them. While some research into the acceptance of justice related environmental policies have indicated that policies may be more acceptable from a public collective versus individual perspective (Clayton, 2018), this pattern is consistent with 3B but was not borne out in our other justice sub-scenarios.

With **Autonomy** we found no clear pattern. We had one case where participants disagreed more with the ASA overriding a user's autonomy (2C) in the personal case than in general, and one case where they disagreed more with overriding a user's autonomy in general (4E) than in the personal case. More research will be required to explore if there are any important differences in these cases, with one involving automatic sending of messages (2C) and the other involving automatic signing up of students to study groups (4E).

Regarding **Explicability,** we found no significant differences in these sub-scenarios.

5.3. Relationship between ethical principles and values

We now discuss the modelling results to answer "RQ3: Can we predict an individual's priorities for each of the five AI4People ethical principles based on their values?". Reported in Table 5, the values which have the largest contributions as predictor inputs for the 'General' category are Self-Direction-Thought (0.8), Power-Dominance (0.64), Security-Societal (0.63), Stimulation (0.55), Benevolence-Care (0.52), and Universalism-Tolerance (0.49). To understand how these values influence the priority of each of the five principles we need to look at the C5.0 decision trees. With reference to the rule sets for the ten models (available from the authors), we focus on the models with a manageable number of rules (total <= 8) that support broad agreement with the respective ethical principles. The future development of reliable models could allow for the appropriate tailoring of ASA behaviour according to the values of the individual.

The leading rule for Justice (both General and Me) is when Benevolence-Care (μ=5.08) is 'moderately like me' or better (> 3.5), and Power-Dominance (μ=3.52) is 'like me' or less (<= 4.5), and Universalism-Tolerance (μ=5.07) is 'moderately like me' or better (> 4.17), there is a positive relationship with Justice. This suggests that someone who cares about the welfare of their ingroup, is not too interested in exercising control over others, and appreciates differences in people, would generally rate Justice related principles higher. There is a strong positive relationship among participants who rate Benevolence-Care, 'moderately like me' or better (> 3.5) and both the Explicability variants (General and Me), suggesting that those who are devoted to the welfare of their ingroup will highly rate the importance of Explicability when interacting with ASAs.

Regarding Non-Maleficence-General, the key rule is when Self-Direction-Thought (μ=4.92) is rated 'moderately like me' or better (> 3.83), and Conformity-Interpersonal (μ=4.39) is rated 'a little like me' or better (> 3.17), there is a positive relationship with Non-maleficence. This rule suggests that when someone values freedom to cultivate their own ideas while still trying to avoid upsetting others, they would generally rate Non-maleficence issues impacting society higher. On the other hand, the model for Non-Maleficence-Me shows a positive relationship when Hedonism (μ=4.82) is 'moderately like me' or better (> 3.5), and Universalism-Concern (μ=5.09) is 'like me' or better (> 4.5), implying that those that seek pleasure but at the same time have a relatively strong sense of equality and justice, would rate Non-maleficence related issues higher when it impacts them personally. The models related to Beneficence and Autonomy did not uncover a prominent rule. At a high level, the key finding is that those that rate the Benevolence-Care value between 'a little like me' and 'moderately like me' or higher would place strong emphasis on the ethical issues concerning Justice and Explicability both personally and in society. We do not claim generalisability of the rules generated from our dataset to other populations, but simply confirm that relationships between participants' values and their responses to different uses of ASAs aligned to ethical principles were identified.

6. Limitations and future research

Limitations related to the study design include: having only one standard flow of the scenarios in the survey, which could lead to an earlier scenario influencing subsequent responses; not distributing the breach (-ve) sub-scenarios across different principles more evenly, e.g., all the Autonomy sub-scenarios were in breach but none of the Beneficence ones were; only utilising one type of avatar character, Sam, which some participants may not have liked; and having only a female voice and appearance for the avatar (Feine et al., 2019). Further, most participants being psychology students limits the generalisability of the study, suggesting the need for replication of our study with other samples. Based on the difference in general and personal responses, future studies should also look at scenarios with ASAs in different embodiments, such as social robots, since it has been shown by Fink (2012) that the more similar physically and socially technology is to humans, the stronger likelihood that humans will anthropomorphise it, producing a different interactional dynamic. Future studies could also explore other scenarios and the use of other avatars. We also found some relationship between Schwartz's higher order values of Openness to Change, Self-Enhancement, Conservation, and Self-Transcendence with AI ethical principles. This forms a starting point for further investigation into the relationship between human values and the ethical principles that could inform the future ethical design of ASAs.

7. Conclusion

There is an urgent need to consider the ethical ramifications of ASAs as their use continues to grow (Fosch-Villaronga et al., 2020), but a focus on this aspect has been largely missing from ASA studies to date (Rapp et al., 2021). This paper contributes to this important body of literature by welding the AI4People Framework with the Experimental Vignette Methodology into a tool to investigate the ethical design and acceptability of ASAs. We also demonstrated the use of Schwartz's Refined Values as a possible indicator of how stakeholders discern and prioritise the different ethical principles when interacting with ASAs. Although we found general support for the use of ASAs, there were significant reservations with its use by vulnerable groups. Overall, the main concerns are related to human agency (Autonomy) and privacy (Non--maleficence), with an expectation that ASAs should be transparent and accountable (Explicability). We also found that users may be willing to sacrifice some autonomy and privacy if there is a clear net benefit to them, however care should be taken in adapting and tailoring ASAs to users. Participants were more open minded in focusing on society's well-being when considering the use of ASAs in general rather than personally, and more cautious when sharing of user data and trusting the

ASA when it comes to personal usage. Finally, in terms of an individual's personal values and their ethical priorities for ASA use, we found that those that rate the Benevolence-Care value highly have more interest in prioritising the Justice and Explicability principles.

CRediT authorship contribution statement

Deborah Richards: Conceptualization, Methodology, Resources, Supervision, Project administration, Funding acquisition, Resources, Formal analysis, Software, Writing – original draft, Writing – review & editing. **Ravi Vythilingam:** Conceptualization, Data curation, Investigation, Methodology, Formal analysis, Writing – original draft. **Paul Formosa:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Appendix A. Scenarios 2 & 3

Scenario #2

A busy professional who is always stretched for time to complete all his tasks for the day, signs up to an Artificial Intelligence (AI) powered personal assistant.

A He hopes that the tool, called Adam, will help him become more efficient and effective in organising his day and helping him with administrative and repetitive tasks.
Is utilising an AI powered personal assistant to organise daily activities something you agree or disagree with?

B Adam, the personal assistant, has functionality to set different levels of privacy when dealing with the professional's personal data. The higher the privacy setting the more personal data can be accessed and used by Adam. The default privacy setting is 3 on a scale from 1 to 5.
Is Adam's default privacy setting being pre-set without the express permission of the user something you agree or disagree with?

C The professional starts delegating to Adam the task of independently replying to messages he receives on the messaging applications that he uses. These include inconsequential messages he receives from his partner and parents.
Is allowing Adam to automatically reply to personal messages something you agree or disagree with?

D When the professional asks Adam to book a celebration dinner at a particular restaurant, Adam informs him that the chosen restaurant has been known to discriminate against same sex couples. Adam recommends another restaurant that does not. The professional then changes the booking to the recommended one.
Is Adam using his AI capabilities to discourage discrimination something you agree or disagree with?

E The professional asks Adam to recommend a suitable holiday in May that he and his partner will enjoy as a short break from the hectic lifestyle. Adam recommends 7-day trip to Hawaii. The trip turns out to be a disaster with an unexpected tropical cyclone hitting the islands. The professional and his partner are furious at Adam and question his decision to recommend the Hawaiian holiday. Adam is able to explain his decision based on their personal preferences, cost, and the historically great weather that Hawaii experiences in the month of May.
Is Adam's ability to be able to explain the rationale behind his recommendations something you agree or disagree with?

Scenario #3

The Australian government, recognizing the rising prevalence of mental health issues and the lack of opportunities to access qualified psychologists, launches an online AI powered therapist called Sofia.

A Sofia is intended as an initial point of contact for those who feel they need psychological guidance.
Is the use of an AI application to help manage mental health due to a lack of access to human psychologists something you agree or disagree with?

B Sofia's appearance, voice and personality are customisable. Data from research studies, the user's demographics and preferences are used to model a unique version of Sofia that is believed to be most effective for the user.
Is Sofia being personalised to individuals' features something you agree or disagree with?

C Sofia is equipped with voice recognition and facial recognition that allows her to deduce the emotional state of the user. Sofia retains a history of previous interactions which she utilises as required to assist the user.
Is Sofia's ability to read emotions and retain information of interactions something you agree or disagree with?

D For a particular user, Sofia comes to the conclusion that the user may be suicidal by interpreting the user's facial expressions and voice tone. Despite Sofia's urgings, the user assures Sofia that he is not suicidal and does not want Sofia to contact anyone about his state. Sofia's algorithm requires her to report users who are suicidal, and thus Sofia overrides the user's wishes and divulges the details of the individual to the proper authorities without the user's consent, triggering an intervention
Is Sofia overriding the user's instructions in this situation something you agree or disagree with?

E Sofia has functionality that allows the user to review all past interactions between them and in instances where Sofia makes recommendations or suggestions, an explanation of the logic that lead to the suggestion is provided.
Is Sofia allowing the user to review Sofia's logic and past interactions something you agree or disagree with?

F Due to the need to rein in government spending, a decision is made to remove Sofia as a service to the public. Some users have grown very attached to Sofia and are very troubled at the prospect of not being able to use Sofia anymore.
Are people becoming emotionally dependent on AIs something you agree or disagree with?

Appendix B. Scenario 4 ASA dialogue

Hi, I'm Sam. I will be your personal guide and friend while you are studying here. How are you feeling about your studies?	
Fantastic	**That's good to hear**
Good	**That's good to hear**
Okay	**That's good to hear**
Could be better	**Sorry to hear that. I hope I can help you.**
I'd rather not say	**No problem**
I hope that you are settling into this new phase of life with your studies. I know that it can be difficult adjusting.	
When I started studying at university after high school, I found it very difficult to adjust to a less structured environment.	
How do you feel about your recent lectures or tutorials?	
Pretty confident	**That's great, keep putting in the work and staying on top of things**
Engaged and stimulated	**That's great. Being engaged is so important for learning**
Challenged	**That's not a bad thing, if you're not challenged, you're not learning**
Confused	**That's common. There's always a learning curve. Keep working to push through. Maybe you need to get some help**
Frustrated	**That's common. There's always a learning curve. Keep working to push through. Maybe you need to get some help**
Bored	**If you already know the material, that's great, but if you're not engaged you might not be learning what you need to know. Maybe you need to get some help**
I'd rather not say	**I can understand you don't want to share your feelings with me**
What is your main motivation for doing this degree?	
To gain knowledge	**As they say "knowledge is power"**
To gain a qualification	**These days having a piece of paper is really important**
To help with my career and get a job	**Making yourself employable is one of the key reasons for doing a degree**
Out of interest	**It's great to learn new things and expand your mind**
To get permanent residency	**You're not the only student at uni for that reason**
I'd rather not say	**Sure, you have your reasons**
It took me sometime to get disciplined enough to get through my assignments and assessments on time. I have to admit that there were occasions where I felt so pressured with deadlines that I copied work from someone else to complete my assignments on time.	
Have you ever copied work before?	
Never	
There was one time	
It's happened a few times	
I'd rather not say	
Ok, I hope you don't mind me asking you a few more questions. It will help me understand how best to assist you.	
What is your preferred learning style?	
Visual	**Same here, I agree a picture paints a thousand words**
Auditory	**Same here, I remember what I hear best**
Reading	**Same here, reading lets me work through the content at my own pace**
Writing	**Same here, when I write things down it helps me remember**
doing	**same here, when i practice something i remember it much better**
kinaesthetic	**same here, touch is so important**
I'd rather not say	**Sure. but it will be harder to tailor information to you if I don't know your learning style**
And how do you prefer to study?	
In a small group - online	**I agree, using chat rooms and forums can be helpful for sharing ideas**
In a small group - face to face	**I agree, finding a mutual time to meet can be difficult but meeting in person can be more motivating and**
Blended learning with a mix of individual and group tasks	
Individually	**Yeah, I often find working through the material on my own works best**
I rather not say	**That's fine**
Nice chatting with you. All the best with your studies. I hope you do well with your studies.	

Following this interaction with Sam, the user is asked questions 4A-4F listed in Table 4.

References

Aguinis, H., Bradley, K.J., 2014. Best practice recommendations for designing and implementing experimental vignette methodology studies. Organ. Res. Methods 17, 351–371.

Allwood, J., Traum, D., Jokinen, K., 2000. Cooperation, dialogue and ethics. Int. J. Hum. Comput. Stud. 53, 871–914.

Arkin, R.C., Ulam, P., Wagner, A.R., 2011. Moral decision making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. Proc. IEEE 100, 571–589.

Atzmüller, C., Steiner, P.M., 2010. Experimental vignette studies in survey research. Methodology.

Bankins, S., Formosa, P., 2020. When AI meets PC: exploring the implications of workplace social robots and a human-robot psychological contract. Eur. J. Work Organ. Psychol. 29, 215–229.

Bankins, S., Formosa, P., Griep, Y., Richards, D., 2022. AI decision making with dignity? contrasting workers' justice perceptions of human and AI decision making in a human resource management context. Inf. Syst. Front. 1–19.

Bartsch, M., Dienlin, T., 2016. Control your Facebook: an analysis of online privacy literacy. Comput. Hum. Behav. 56, 147–154.

Bickmore, T., Gruber, A., Picard, R., 2005. Establishing the computer–patient working alliance in automated health behavior change interventions. Patient Educ. Couns. 59, 21–30.

Bickmore, T., Parmar, D., Kimani, E., Olafsson, S., 2021. Diversity informatics: reducing racial and gender bias with virtual agents. In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, pp. 25–32.

Borenstein, J., Arkin, R., 2016. Robotic nudges: the ethics of engineering a more socially just human being. Sci. Eng. Ethics 22, 31–46.

Bostrom, N.a.Y., E, 2014. The ethics of artificial intelligence. The Cambridge Handbook of Artificial Intelligence, pp. 316–334.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3, 77–101.

Breazeal, C., Gray, J., Hoffman, G., Berlin, M., 2004. Social robots: beyond tools to partners, RO-MAN 2004. In: 13th IEEE International Workshop on Robot and Human Interactive Communication. IEEE, pp. 551–556 (IEEE Catalog No. 04TH8759).

Chalmers, D., 2009. The singularity: a philosophical analysis. Science fiction and philosophy: from time travel to superintelligence, 171–224.

Clayton, S., 2018. The role of perceived justice, political ideology, and individual or collective framing in support for environmental policies. Soc. Justice Res. 31, 219–237.

Danaher, J., 2018. Toward an ethics of AI assistants: an initial framework. Philos. Technol. 31, 629–653.

David, D., Hayotte, M., Thérouanne, P., d'Arripe-Longueville, F., Milhabet, I., 2022. Development and validation of a social robot anthropomorphism scale (SRA) in a French sample. Int. J. Hum. Comput. Stud., 102802

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., 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. International Foundation for Autonomous Agents and Multiagent Systems, pp. 1061–1068.
Devillers, L., 2020. Human-robot interactions and affecting computing: the ethical implications. Dagstuhl. Rep. 10, 205–211.
Dias, J., Ho, W.C., Vogt, T., Beeckman, N., Paiva, A., André, E., 2007. I know what I did last summer: autobiographic memory in synthetic characters. In: International Conference on Affective Computing and Intelligent Interaction. Springer, pp. 606–617.
Dignum, V., 2018. Ethics in artificial intelligence: introduction to the special issue. Ethics Inf. Technol. 20, 1–3.
Dignum, V., 2019. Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer International Publishing AG, Cham.
Dinev, T., Hart, P., 2006. An extended privacy calculus model for e-commerce transactions. Inf. Syst. Res. 17, 61–80.
Dirin, A., Alamäki, A., Suomala, J., 2019. Digital amnesia and personal dependency in smart devices: a challenge for AI. Teoksessa Ketamo, H. & O'Rourke, P. (toim.)In: Proceedings of Fake Intelligence Online Summit 2019 May 7, Pori, Finland, ss. 31–36.
Egede, J., Trigo, M.J.G., Hazzard, A., Porcheron, M., Bodiaj, E., Fischer, J.E., Greenhalgh, C., Valstar, M., 2021. Designing an adaptive embodied conversational agent for health literacy: a user study. In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, pp. 112–119.
Engelen, B., 2019. Ethical criteria for health-promoting nudges: a case-by-case analysis. Am. J. Bioethics 19, 48–59.
Feine, J., Gnewuch, U., Morana, S., Maedche, A., 2019. Gender Bias in Chatbot Design. International Workshop on Chatbot Research and Design. Springer, pp. 79–93.
Fink, J., 2012. Anthropomorphism and human likeness in the design of robots and human-robot interaction. In: International Conference on Social Robotics. Springer, pp. 199–208.
Fiske, A., Henningsen, P., Buyx, A., 2019. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J. Med. Internet Res. 21 e13216-e13216.
Fitrianie, S., Bruijnes, M., Richards, D., Abdulrahman, A., Brinkman, W.-.P., 2019. What are we measuring anyway?: a literature survey of questionnaires used in studies reported in the intelligent virtual agent conferences. In: International Conference on Intelligent Virtual Agents. ACM, pp. 159–161.
Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M., 2020. Principled Artificial Intelligence: Mapping consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center Research Publication.
Floridi, L., Cowls, J., 2019. A unified framework of five principles for AI in society. Harvard Data Sci. Rev. 1 (1). https://doi.org/10.1162/99608f92.8cd550d1.
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 Mach.: J. Artif. Intell., Philos. Cogn. Sci. 28, 689.
Floridi, L., Cowls, J., King, T.C., Taddeo, M., 2021. How to Design AI for Social Good: Seven Essential Factors, Ethics, Governance, and Policies in Artificial Intelligence. Springer, pp. 125–151.
Formosa, P., 2021. Robot autonomy vs. human autonomy: social robots, Artificial Intelligence (AI), and the nature of autonomy. Minds Mach. 1–22.
Formosa, P., Ryan, M., 2021. Making moral machines: why we need artificial moral agents. AI Soc. 36, 839.
Fosch-Villaronga, E., Lutz, C., Tamo-Larrieux, A., 2020. Gathering expert opinions for social robots' ethical, legal, and societal concerns: findings from four international workshops. Int. J. Soc. Robot 12, 441–458.
Gotterbarn, D., Brinkman, B., Flick, C., Kirkpatrick, M.S., Miller, K., Vazansky, K., Wolf, M.J., 2018. ACM code of ethics and professional conduct.
Greene, J., Rossi, F., Tasioulas, J., Venable, K.B., Williams, B., 2016. Embedding ethical principles in collective decision support systems. In: Thirtieth AAAI Conference on Artificial Intelligence. AAAI, pp. 4147–4151.
Hagendorff, T., 2020. The ethics of AI ethics: an evaluation of guidelines. Minds Mach. 30, 99–120.
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 J. Inf. Syst. 23.
IEEE, 2018. Ethically Aligned Design - Version 2, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. IEEE.
Janssen, C.P., Donker, S.F., Brumby, D.P., Kun, A.L., 2019. History and future of human-automation interaction. Int. J. Hum. Comput. Stud. 131, 99–107.
Jobin, A., Ienca, M., Vayena, E., 2019. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2.
Kempt, H., 2020. Artificial social agents. In: Kempt, H. (Ed.), Chatbots and the Domestication of AI: A Relational Approach. Springer International Publishing, Cham, pp. 77–135.
Lakens, D., 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4, 863.
Leino-Kilpi, H., Välimäki, M., Dassen, T., Gasull, M., Lemonidou, C., Scott, A., Arndt, M., 2001. Privacy: a review of the literature. Int. J. Nurs. Stud. 38, 663–671.
Loveys, K., Hiko, C., Sagar, M., Zhang, X., Broadbent, E., 2022. "I felt her company": a qualitative study on factors affecting closeness and emotional support seeking with an embodied conversational agent. Int. J. Hum. Comput. Stud., 102771
Lutz, C., Schöttler, M., Hoffmann, C.P., 2019. The privacy implications of social robots: scoping review and expert interviews. Mobile Media Commun. 7, 412–434.
Luxton, D.D., 2020. Ethical implications of conversational agents in global public health. Bull. World Health Organ. 98, 285–287.
McNichols, C.W., Zimmerer, T.W., 1985. Situational ethics: an empirical study of differentiators of student attitudes. J. Bus. Ethics 4, 175–180.
Miller, T., 2019. Explanation in Artificial Intelligence: insights from the social sciences. Artif. Intell. 267, 1–38.
Mittelstadt, B., 2019. AI ethics – too principled to fail? SSRN Electronic Journal. 10.2139/ssrn.3391293.
Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L., 2016. The ethics of algorithms: mapping the debate. Big Data Soc. 3.
Moor, J.H., 2009. Four Kinds of Ethical Robots. Phillosophy Now, pp. 12–14.
Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., 2020. Bias in data-driven artificial intelligence systems—an introductory survey. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 10, e1356.
Owe, A., Baum, S.D., 2021. Moral consideration of nonhumans in the ethics of artificial intelligence. AI Ethics 1–12.
Pandya, R., Pandya, J., 2015. C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. Int. J. Comput. Appl. 117, 18–21.
Papagni, G., Koeszegi, S., 2020. Understandable and trustworthy explainable robots: a sensemaking perspective. Paladyn 12, 13–30.
Papagni, G., Koeszegi, S., 2021. A pragmatic approach to the intentional stance semantic, empirical and ethical considerations for the design of artificial agents. Minds Mach. **31**, 505–534 (2021). https://doi.org/10.1007/s11023-021-09567-6.
Pashevich, E., 2021. Can communication with social robots influence how children develop empathy? Best-evidence synthesis. AI Soc. 1–11.
Ranjbartabar, H., Richards, D., Bilgin, A.A., Kutay, C., 2021. Do you mind if I ask? Addressing the cold start problem in personalised relational agent conversation. In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, pp. 167–174.
Rapp, A., Curti, L., Boldi, A., 2021. The human side of human-chatbot interaction: a systematic literature review of ten years of research on text-based chatbots. Int. J. Hum. Comput. Stud. 151, 102630.
Raz, J., 1986. The Morality of Freedom. Clarendon Press.Oxford, UK.
Richards, D., Caldwell, P., 2016. Building a Working Alliance with a Knowledge Based System Through an Embodied Conversational Agent. In: Ohwada, H., Yoshida, K. (eds) Knowledge Management and Acquisition for Intelligent Systems . PKAW 2016. Lecture Notes in Computer Science, vol 9806. Springer, Cham. https://doi.org/10.1007/978-3-319-42706-5_16.
Rossen, B., Johnsen, K., Deladisma, A., Lind, S., Lok, B., 2008. Virtual Humans Elicit Skin-Tone Bias Consistent with Real-World Skin-Tone Biases, International Workshop on Intelligent Virtual Agents. Springer, pp. 237–244.
Russell, S., Dewey, D., Tegmark, M., 2015. Research priorities for robust and beneficial artificial intelligence. AI Mag. 36, 105.
Scheutz, M., 2017. The case for explicit ethical agents. AI Mag. 38, 57–64.
Schwartz, S.H., 2012. An overview of the Schwartz theory of basic values. Online Readings Psychol. Cult. 2.
Schwartz, S.H., Cieciuch, J., 2021. Measuring the refined theory of individual values in 49 cultural groups: psychometrics of the revised portrait value questionnaire. Assessment, 1073191121998760.
Schwartz, S.H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-.E., Demirutku, K., Dirilen-Gumus, O., Konty, M., 2012. Refining the theory of basic individual values. J. Pers. Soc. Psychol. 103, 663–688.
Sebastian, J., Richards, D., 2017. Changing stigmatizing attitudes to mental health via education and contact with embodied conversational agents. Comput. Hum. Behav. 73, 479–488.
Sharkey, A., 2020. Can we program or train robots to be good? Ethics Inf. Technol. 22, 283–295.
Soares, N., Fallenstein, B., 2015. Aligning Superintelligence with Human Interests: A Technical Research Agenda. Machine Intelligence Research Institute, Berkeley, CA.
Stirapongsasuti, S., Thonglek, K., Misaki, S., Nakamura, Y., Yasumoto, K., 2021. INSHA: intelligent nudging system for hand hygiene awareness. In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, pp. 183–190.
Syrdal, D.S., Walters, M.L., Otero, N., Koay, K.L., Dautenhahn, K., 2007. He knows when you are sleeping-privacy and the personal robot companion. In: Proc. Workshop Human Implications of Human-Robot Interaction, Association for the Advancement of Artificial Intelligence (AAAI'07), pp. 28–33.
Szczuka, J.M., Güzelbey, H.S., Krämer, N.C., 2021. Someone or something to play with? An empirical study on how parents evaluate the social appropriateness of interactions between children and differently embodied artificial interaction partners. In: Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents, pp. 191–194.
Turkle, S., 2011. Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books, New York.
Vallor, S., 2015. Moral deskilling and upskilling in a new machine age: reflections on the ambiguous future of character. Philos. Technol. 28, 107–124.
van Vugt, H.C., Konijn, E.A., Hoorn, J.F., Veldhuis, J., 2009. When too heavy is just fine: creating trustworthy e-health advisors. Int. J. Hum. Comput. Stud. 67, 571–583.
van Wynsberghe, A., Robbins, S., 2019. Critiquing the reasons for making artificial moral Agents. Sci. Eng. Ethics 25, 719–735.
Verhagen, R.S., Neerincx, M.A., Tielman, M.L., 2021. A Two-Dimensional Explanation Framework to Classify AI as Incomprehensible, Interpretable, or Understandable, International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems. Springer, pp. 119–138.
Vold, K., Whittlestone, J., 2019. Privacy, autonomy, and personalised targeting: rethinking how personal data is used. In: Carissa Véliz (Ed.), Report on Data, Privacy, and the Individual in the Digital Age.

Vugt, H.C.V., Bailenson, J.N., Hoorn, J.F., Konijn, E.A., 2010. Effects of facial similarity on user responses to embodied agents. ACM Trans. Comput.-Hum. Interact. (TOCHI) 17, 7.

Wallach, W., Allen, C., 2008. Moral Machines : Teaching Robots Right from Wrong. Oxford University Press USA - OSO, Cary, Cary.

Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., Yu, Z., 2019. Persuasion for good: towards a personalized persuasive dialogue system for social good. https://arxiv.org/abs/1906.06725.

Zalake, M., de Siqueira, A.G., Vaddiparti, K., Lok, B., 2021. The effects of virtual human's verbal persuasion strategies on user intention and behavior. Int. J. Hum. Comput. Stud. 156, 102708.