Ethical Design and Acceptability of Artificial Social Agents

By

Ravi T. Vythilingam

A THESIS SUBMITTED TO MACQUARIE UNIVERSITY

FOR THE DEGREE OF MASTER OF RESEARCH

DEPARTMENT OF COMPUTING

1ST DECEMBER 2021

MACQUARIE
University
SYDNEY · AUSTRALIA

© Ravi Vythilingam 2021

Statement of Originality

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself. The Ethics Committee approval and protocol number of this study is: 52020623814005.

Ravi T. Vythilingam Date: 1st December 2021

Acknowledgements

I would like to extend my profound gratitude to my principal supervisor, Professor Deborah Richards from the Department of Computing for her expert guidance, persistent support, constant patience, timely encouragement, and prompt feedback through the past two and half years of my part-time MRes undertaking, without which I would not have completed this study.

My sincere thanks are also extended to my associate supervisor, Associate Professor Paul Formosa from the Department of Philosophy for expanding his time and effort providing his valuable thoughts and insights.

I would also like to thank Research Assistant Meredith Porte for assisting with the avatar scenario used in this study and Associate Professor Ayse Bilgin for her statistics guidance.

Finally, I would like to thank my mother Chandra, my life partner Anitha, our two children, Visakhan and Divya and friends for their untiring encouragement and patience while I completed this personal academic yatra.

List of Publications

Vythilingam, R., Richards, D. & Formosa, P., 2020, The relationship between human values and the ethical design and acceptability of relational agents, Proceedings of the 9th Conference of the Australian Institute of Computer Ethics (AiCE 2020): Computer Ethics in the New Normal. Australasian Institute of Computer Ethics (AiCE), p. 1-10 10 p.

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

Abstract

Artificial Social Agents (ASA), which are AI software driven entities programmed with rules and preferences to act autonomously with humans, are increasingly playing more human-like 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. The aim of this thesis is to investigate what ethical principles are of relative importance when people engage with ASAs and if there is a relationship between people's values and the ethical principles they prioritise. The study uses the five AI4People Ethical principles (Beneficence, Non-maleficence, Autonomy, Justice, and Explicability) and Schwartz's theory of human values. Scenarios with embedded ethical principles that involved an ASA taking on a role traditionally played by a human were created to understand the types of ASA attributes that are acceptable or unacceptable. We found that participants are most sensitive to ASA attributes that relate to Autonomy, Justice, Explicability, and the privacy of their personal data; and ASAs were more acceptable when used generally in society rather than personally. 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.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

AI:	Artificial Intelligence
AMA:	Artificial Moral Agents
ART:	Accountability, Responsibility and Transparency
ASA:	Artificial Social Agents
AUT:	Autonomy
AV:	Autonomous Vehicles
BEN:	Beneficence
EU:	European Union
EVM:	Experimental Vignette Methodology
EXP:	Explicability
HCI:	Human-Computer Interactions
IEEE:	Institute of Electrical and Electronic Engineers
IVA:	Intelligent Virtual Agents
JUS:	Justice
NON-MAL:	Non-maleficence
OECD:	Organisation for Economic Co-operation and Development
PVQ-RR:	Portrait Values Questionnaire (revised)
RQ:	Research Question
SLIVAR:	Spoken Language Interaction with Virtual Agents and Robots
XAI:	eXplainable Artificial Intelligence

1 Introduction

1.1 Background

Artificial Intelligence (AI) powered applications, tools and agents are becoming more prevalent across society and business. In many situations, they are replacing, reducing, mediating, or complementing human effort in personal, social, administrative, government, commercial and industrial tasks. Artificial Social Agents (ASA) are AI software driven entities in virtual or physical form and include AI based social robots, embodied conversational agents, relational agents, and intelligent virtual agents (IVA). ASAs are increasingly becoming the face of AI as far as the public is concerned. They are increasingly playing more human like roles in society, particularly in education, healthcare, childcare, eldercare, and coordinating, advising, and coaching settings.

These AI applications are developed based on datasets that we choose to include and algorithms we choose to implement, creating moral choices and implications (Ntoutsi et al., 2020). Further, differences in expectations, assumptions and the reality between what is designed and developed and the ethical impact of what is deployed quite often has significant impacts (Mittelstadt et al., 2016). Many governments, universities, organisations, industry forums and public figures have raised concerns regarding the widespread use of AI technologies. The concerns raised include issues related to cognitive degeneration, threats to autonomy (Danaher, 2018), accountability, privacy, discrimination, security, societal dynamics and economic impacts (IEEE, 2018).

As a result, there has been a significant focus lately on the ethics of Artificial Intelligence (AI) with numerous organisations, companies and jurisdictions developing and publishing their own set of frameworks, principles, and guidelines on AI ethics. Floridi et al. (2018) reviewed a number of these guidelines and synthesised five overarching ethical principles under the AI4People's Unified Framework of Principles for AI in Society banner. At this juncture, organisations are finding it challenging to put these principles into practice, with minimal adherence by developers and management (Hagendorff, 2020, Mittelstadt, 2019). These entities and individuals are currently primarily driven by time pressure to market and economic payback considerations. One of the challenges is how to turn abstract principles into tangible guidelines that will guide, nudge and/or compel the players in the AI development ecosystem to incorporate ethical principles into AI applications (Mittelstadt, 2019, Morley et al., 2019). Furthermore, most of these guidelines and academic research have an implicit focus on the ethics of AI applications related to data science, big data, and machine learning and not enough focus on the ethics related to the design and acceptability of ASAs. With the accelerating use of ASAs in multiple domains of human-computer interaction (HCI), more focus and research needs to target the ethics of ASAs.

ASAs are programmed with certain rules and preferences to act autonomously with humans (Fitrianie et al., 2019). Their sophistication will only grow with the ASAs becoming more autonomous and powerful over time (Russell et al., 2015). This will lead to humans sharing greater amounts of personal information, thoughts, and feelings with ASAs, resulting in significant ethical implications, with one of the most being human autonomy (Formosa, 2021). Further, there will be a growing tendency for humans and agents to be working in tandem to make decisions in the pursuit of specific objectives. This will spur the requirement to align ethical principles and moral values between the human and the ASA (Greene et al., 2016), however such alignment of princples, values and preferences will be challenging (Soares and Fallenstein, 2015).

1.2 Research Motivation, Aim and Questions

The increasing breadth and depth of ASA usage, coupled with the need for further research related to the ethical design and acceptability of ASAs, are the primary motivational drivers for this study. The specific aim of the study is to investigate what ethical principles are of relative importance to humans when we engage with ASAs and if there is a relationship between our values and the ethical principles we prioritise. Our first and primary research question to support the study's aim is:

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

Within the context of understanding ASA ethical acceptability, we sought to explore whether the users' position on the acceptability of the ASA's behaviour differed when they were responding from the general society's perspective as compared to when the protagonist interacting with the ASA was someone close to them. A study on situational ethics comparing 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). This leads to a second research question.

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

Lastly, we explored if there was any relationship between an individual's values and the AI4People's ethical principles. Any such relationship can assist with the design and acceptability of ASAs and how ethical principles and moral values can be aligned between humans and ASAs when they work together. The existence of a relationship could also contribute to the identification of societal values when implementing, for example the ‘Society-in-the-Loop’ concept to help specify an algorithmic social contract for the regulation of the ethical use of AI and algorithmic systems (Rahwan, 2018).

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

1.3 Approach

Given the aim and research questions, we conducted a study to expose respondents to different scenarios with ASAs following these steps:

- *Confirm the set of ethical principles to be evaluated*. With several ethical frameworks and principles published, an appropriate list of principles needs to be identified. We selected AI4People’s ethical principles, discussed further in the next chapter.
- *Determine an optimum number of participants*. A suitable number of participants to support statistically relevant results and also allow for a manageable number to complete the work in a timely manner. Approximately 200 participants were initially targeted.
- *Identify suitable values survey*. An appropriate online survey that is culturally unbiased and reasonable in length to measure human values was required. As discussed in chapters 2 and 3, the PVQ-RR Survey Instrument based on Schwartz’s Refined Value Theory was selected.
- *Design ASA scenarios*. Suitable and relatable scenarios where an ASA replaces a role traditionally played by a human was identified. To understand the types of ASA use cases that were or were not ethically acceptable, the AI4People’s ethical principles needed to be embedded in the design of the scenarios. To achieve this, each scenario would need to support sub-scenarios and questions that can evaluate the participants’ position in relation to the specific ethical principle/s being considered.
- *Determine suitable questions for the scenarios.* Questions must be based on the identified scenarios and have a direct relationship to one of more of the ethical principles being considered.
- *Data collection and statistical analysis of responses.* Statistical analysis of responses to investigate the research questions.

In brief, we have conducted a study to expose respondents to different scenarios with ASAs. Our study is structured into three sections. Following data collection of demographic data and a human values questionnaire, the study requires the development of the following material: three descriptive ASA scenarios with associated sub-scenarios and questions aligned with ethical principles and a conversational ASA avatar scenario with associated sub-scenarios and questions. The detailed research method is described in Chapter 3 of this thesis.

1.4 Thesis Outline

This chapter discusses ethical concerns and shortcomings with AI applications and agents that is driving the increasing importance of AI ethics. It goes on to introduce Artificial Social Agents and how ASAs and humans will progressively be working more closely together as a team to achieve common objectives. As this occurs, value alignment between humans and ASAs becomes more critical. The gap in research into the ethical acceptability of ASAs is the motivation of this study.

Chapter 2 covers a literature review of all the key areas in this thesis. Ethics in AI, ASAs, its ethics and related concerns. AI Ethical Principles are then reviewed, discussed and a set of ethical values proposed to be used for this study. Human values and how values can be measured are discussed with a values model proposed for our study.

The methodology utilised to address the research questions is articulated in Chapter 3. Here we explain the study design model, the data collection questions, surveys, and scenario material used and its rationale. This is followed by the data collection and analysis procedures. The results of the study are presented in Chapter 4 and discussed in Chapter 5 to explicitly answer the research questions. Chapter 6 provides a thesis summary, contributions, limitations, and future directions for research, concluding with final remarks.

Appendix A reproduces the Ethics Approval Letter from the University, Appendix B documents the study survey, and Appendix C offers selected rule sets from the SPSS Modeller analysis.

2 Literature Review

As AI applications become more prevalent in everyday life, their effects from an ethical perspective become more important. Artificial Social Agents, which include embodied agents and virtual agents, are designed, and implemented for the ASA to autonomously interact with humans by following human social rules. As ASAs gain more agency and sophistication, several ethical concerns have surfaced. This chapter starts with a review of Ethics in AI in section 2.1, followed by introducing Artificial Social Agents (ASA) in section 2.2. Section 2.3 reviews the ethics of ASAs, followed by section 2.4 which investigates the ethical concerns of ASAs. The next section (2.5) investigates appropriate AI ethical principles to be utilised in this research and finally section 2.6 determines how to obtain a snapshot of human values to explore its relationship with the identified ethical principles.

2.1 Ethics in AI

Ethics has become one of the leading areas of focus in the Artificial Intelligence sphere (Dignum, 2018). Given the variations in our cultural backgrounds, family upbringing, life experiences and subjectivity of what is acceptable and not, it is understandable that there are strong opinions when it comes to ethics. Disagreements are to be expected in ethics and as such ethics cannot be easily reduced to a mathematical formula that can be incorporated into an AI application. The question of ‘what is ethical’ needs to be worked through by various stakeholders and subject matter experts and can be specific to the utilisation of the application and the community of users (Mittelstadt, 2019). Lauer (2021) argues that AI ethics can only truly exist in organisations where there is a culture of ethical behaviour, and an understanding of how complex systems work. With regard to this project, ethics is defined as discerning between what is morally acceptable versus what is unacceptable in a particular scenario or setting either generally in society or with regard to a specific cultural, demographic, organisational or national group (Velasquez, 1987).

Ethical issues raised using AI applications have been widely researched (Bostrom, 2014, Danaher, 2018, IEEE, 2018, Mittelstadt et al., 2016, Wallach and Allen, 2008). Some of the key areas of concern raised by these researchers and the various AI ethical guidelines published as summarised by Hagendorff (2020) and Floridi et al. (2018) include privacy protection, accountability, fairness, justice, transparency, traceability, human autonomy, explicability, safety, sustainability, cognitive degeneration, human control, diversity, future of employment, and hidden cost. Further Mittelstadt et al. (2016) mapped the kinds of ethical issues that could be caused by poorly designed algorithms and came up with six. Three epistemic concerns are inclusive evidence that could lead to unsupported outcomes; inscrutable evidence that hinders transparency; and misguided evidence that could cause prejudiced actions. The two normative ones are unjust outcomes causing discrimination, and transformative impacts triggering issues with privacy and human autonomy. Finally, traceability or lack thereof will make it difficult to assign responsibility when a complication or problem occurs.

Wallach and Allen (2008) in their book on AI ethics map out a simple framework of how ethics could be developed into AI applications. Autonomy and ethical sensitivity of an AI application are considered along the vertical and horizontal axis respectively. Incorporating ethics in AI starts with operational morality and progresses to functional morality and full moral agency. Applications with operational morality have low autonomy and ethical sensitivity but can have ethics incorporated by design. An example are chat applications that censor politically sensitive words or messages. Next, we have functional morality with increased autonomy and/or ethical sensitivity. These are applications that are capable of some level of ethical reasoning. Artificial Social Agents are AI applications with sophistication ranging from operational morality to functional morality. Artificial Moral Agents (AMA) have full moral agency and are theoretically capable of self-reflection (Wallach and Allen, 2008). Implementation of AI applications with functional morality or AMA require a base set of AI ethical principles.

2.2 Artificial Social Agents (ASA)

Artificial Social Agents (ASA) include AI based social robots, embodied conversational and relational agents, and intelligent virtual agents (IVA). ASAs are programmed with human to human communications related rules and preferences to converse autonomously with humans (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 exists a significant amount of literature on social robots, its impacts and ethical implications including Breazeal et al. (2004), Turkle (2011), Bankins and Formosa (2020), Lutz et al. (2019), and Pashevich (2021).

The application and sophistication of ASAs will only grow with humans developing stronger and deeper relationships in a working alliance with agents (Bickmore et al., 2005, Richards and Caldwell, 2016, Turkle, 2011) and the agents becoming 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, 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).

2.3 Artificial Social Agents and Ethics

Moor (2009) recognised four levels of artifical ethical agents. From the most basic to the most advanced, 1) ethical impact agents which by design or otherwise have ethical impacts; 2) implicit ethical agents are agents which are designed with specific automated reactions when faced with certain situations such as when attempted fraud is detected; 3) explicit ethical agents are implemented with overriding ethical guidelines which are interpreted to inform the agent how to act in different situations as they arise, i.e., 'acting from ethics'; and 4) full ethical agents which theorectically have consciousness, intentionality and free will like adult humans.

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. Papagni and Koeszegi (2021), in a comparative review of artificial agents literature, conclude that it is ethical and essential to endow ASAs with intentionality, social ability, and goal-driven rational behaviour providing 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 utilised, for example, in complex situations where real-time decisions are required to prevent harm such as agents in AV (autonomous vehicles) and carebots.

2.4 Artificial Social Agents Ethical Concerns

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, 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) raised ethical concerns over the predominant portrayal of ASAs using female avatars. Feine et al. (2019) analysed 1,375 chatbots and found they were predominantly female in violation of the "ACM Code of Ethics and Professional Conduct" which states that "computing professionals should foster fair participation of all people" (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 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 procedure regarding the development and deployment of the 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 of the impact of children interacting with ASAs, voice assistants and embodied robots. It highlights parents' focus on agent embodiment and privacy protection. In the carebot space, Scheutz (2017) raises concerns regarding vulnerable populations forming unreciprocated emotional bonds with assistive robots and virtual agents that could potentially cause harm to the human. Similar concerns were raised with regard to agent / robot companions used by the general population.

In looking into the ethics of personal AI assistants, Danaher (2018) argues that AI assistants change the cognitive architecture that we operate in and hence when using such technologies, we need to cautiously deliberate the tradeoff between what value it produces for us, intrinsic or instrumental. They then present a risk/reward structure to evaluate the ethical use of AI assistants based on the impacts on cognitive degeneration (with instrinsic versus instrumental value as a guide); autonomy (as technology can influence our decision choice-architecture); and interpersonal interactions (cautious when the principal value of the human-to-human interaction resides in the engagement's conscious and immediate nature).

Vold and Whittlestone (2019) make the conection between personal data privacy (a sub-set of the Non-maleficence ethical principle) with autonomy, postulating that a person's concern with privacy relates to the risk that those who have access to their personal information can target them in a personalised and opaque manner to influence them in a particular way, diminishing their ability to form independent decisions. The authors put forward five criteria to distinguish between ethical and unethical targeting: 1) consistency with values and interests; 2) transparency; 3) attempt to obtain consent; 4) not seek to limit information or choices that could misrepresent reality; and 5) should not make use of personally sensitive data. The study conducted under the auspices of the "NoBias - Artificial Intelligence without Bias Project" in the EU (Ntoutsi et al., 2020), states that bias is manifiested in data through the use of sensitive characteristics and causal influences, such as certain post codes being highly correlated with race; under- or over-representation of data, for example under-representation of women and people of colour as IT developers (Buolamwini and Gebru, 2018); and the use of different data modalities and the bias within them, an example being image datasets used for facial recognition services (Buolamwini and Gebru, 2018). A paper by Howard and Borenstein (2018) found that some existing biases in society end up in AI algorithms, pepeutuating the bias and propose some steps to mitigate its effects.

Engelen (2019) discusses the appropriate use of persuasive technologies and the use of nudges in particular and raises three main concerns, namely: is the recommended action appropriate?; is the approach ethical?; and who is doing the persuading? Borenstein and Arkin (2016) discuss in what instances should social robots or similart agents 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 persuadee (Wang et al., 2019). They further stress avoidance of deception and transparency concerning who designed the system and allowing direct communication with them; 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.

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: 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.

The SLIVAR (Spoken Language Interaction with Virtual Agents and Robots) (Devillers, 2020) community identified certain design and implementation challenges that can lead to ethical issues including: unavoidable incomplete specifications, learning errors, learning without understanding, learning biases, non-reproducible evaluation results due to dynamism and distrust due to lack of current transparency or inadequate evaluation. The community expressed concerns around confusion of the agent’s "status" which particularly resonated with our concerns around social agent roles and impact on human relationships, due to strategies such as giving the agent a name, humanlike appearance and “life” that can lead to unhealthy relationships resulting in manipulation, isolation, dissapointment, and machine addition.

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 in the design and implementation of avatars. They are: (1) *Proteus effect* of unintended influence on the user; (2) *Uncanny valley effect* when the avatar looking closely 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 user attachment with the agent. How these factors are designed for and implemented has strong ethical implications and impact on the user. To assist in deploying appropriate ethical rules in a particular context, processes are required to capture and incorporate stakeholder values and expectations into the design and deployment artificial social agents (Dignum, 2017). Further Dignum (2019) proposed the ART principles of Accountability, Responsibility and Transparency to support the design of ethical ASAs. Fosch-Villaronga et al. (2020) who had collated expert opinions from therapists, roboticists, industry representatives, academics, and legal practitioners involved in developing ASAs concluded that researchers and developers have minimal understanding of the attitudes and requirements of potential users and recommended a human centred design approach. Rahwan (2018)’s society-in-the-loop proposal is one approach that could help address some of the ASA design, deployment, and continued operation issues. In this approach, the ASA is recognised to have broad impact and a variety of stakeholder inputs are required to identify and strike the optimum cost-benefit balance between what ethical values, functionality, limitations, and safety features are implemented.

2.5 AI Ethical Principles

Numerous sets of AI ethical principles have been published. Comparisons, analyses and listing of the various frameworks have been carried out by various researchers (including Hagendorff (2020), Floridi et al. (2018) Jobin et al. (2019) and Fjeld et al. (2020)) and by organisations such as AlgorithmWatch (AlgorithmWatch, 2020).

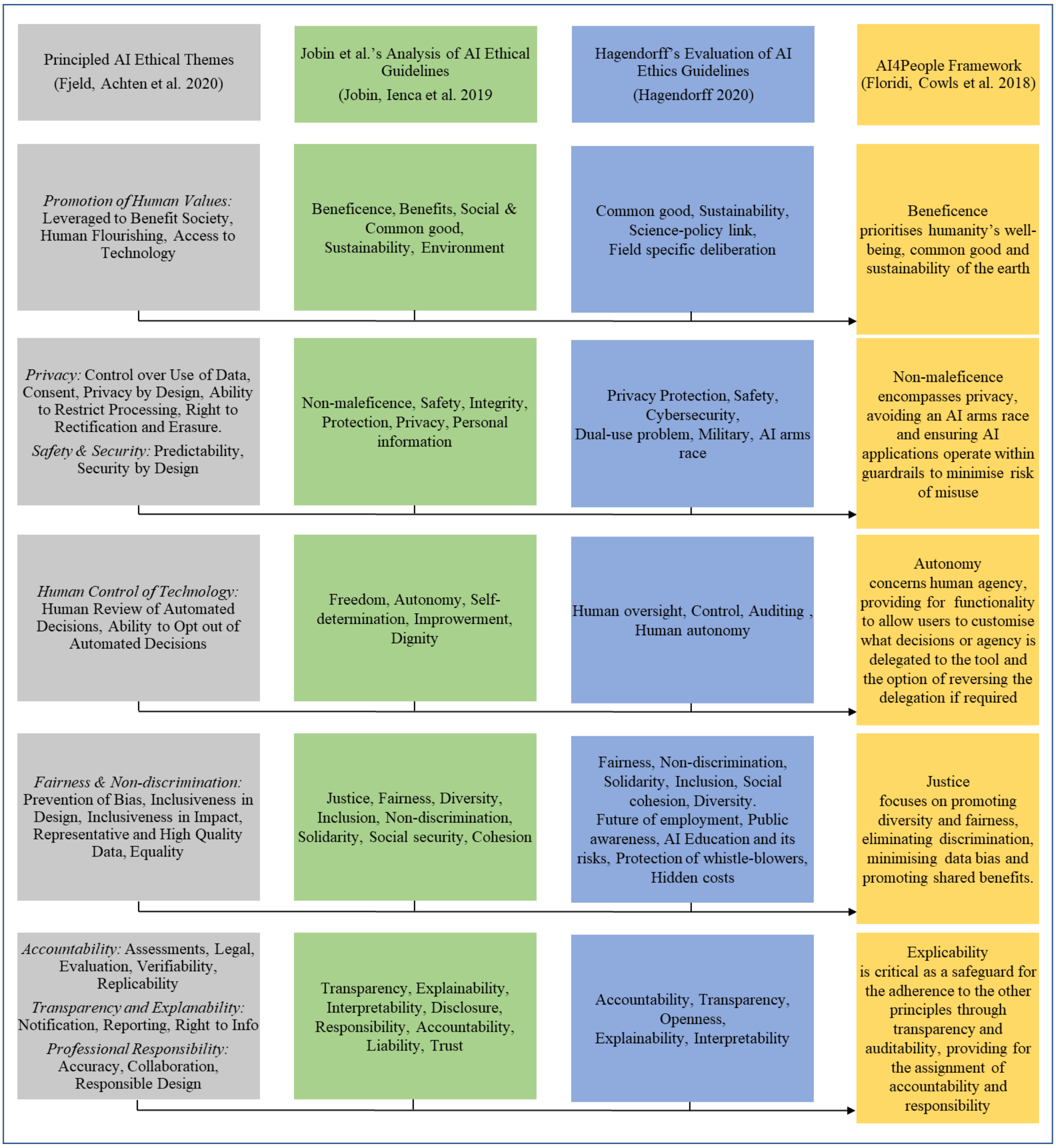
(Hagendorff, 2020) analysed and compared 15 internationally recognised AI Ethics guidelines and its implementation. The paper identified overlaps and gaps among the principles. Commonly identified principles include privacy, accountability, fairness, safety, sustainability, and auditing. Omissions include impacts due to lack of focus on diversity, political misuse, industry funded research and ecological costs. The paper found a lack of adherence to the principles in practice. A separate empirical study (McNamara et al., 2018) found that exposure to ethics guidelines made no statistically significant difference in software design and development decisions made by software developers. Recent research on state of play in transitioning towards how to utilise ethical principles in practice found significant gaps in practical tools and methods as most of the tools identified were immature (Morley et al., 2019). The risks of not closing this gap include potentially significant costs associated with an ethical failure which could also undermine acceptance and adoption of AI applications resulting in a considerable number of missed opportunities (Cookson, 2018). It should be noted that the IEEE through it's 'Ethically-Aligned Design' programme (IEEE, 2018) is making a significant effort to translate ethical principles into technical standards and field friendly tools.

Floridi (2019) examined the ethical principles tabled by the Asilomar AI Principles (FutureofLifeInstitute, 2017), Montreal Declaration for Responsible AI (UniversityofMontreal, 2017), IEEE's Ethically Aligned Design report (IEEE, 2018), European Commission's Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems (EU, 2018), UK House of Lords Artificial Intelligence Committee's report (UKHouseofLords, 2018), and the Tenets of the Partnership on AI (PartnershiponAI, 2018). Significant overlap across the 47 principles listed and prompted synthesis of five ethical principles under the AI4People's Unified Framework of Principles for AI in Society (Floridi et al., 2018). The principles are beneficence, non-maleficence, autonomy, justice, and explicability. The first four principles are traditional bioethics principles (Beauchamp and Childress, 2001). The principles are consistent with the OECD AI Principles (OECD, 2019) adopted by 42 countries in May 2019. The G20 adopted human-centred AI principles in June 2019 drawing on OECD AI Principles (G20, 2019).

Beneficence focusses on humanity's well-being, sustainability, and common good. Non-maleficence basically means do no harm, encompassing privacy, avoiding an AI arms race, 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 inadvertently delegate to AI. AI tools should provide functionality to allow users to customise what decisions or agency is delegated with the option to reverse the delegation. The Justice principle focuses on promoting diversity and fairness, minimising data bias, eliminating discrimination, and promoting shared benefits. Explicability is critical safeguard for adherence to the other principles. It requires transparency and auditability to support accountability in the event of an undesirable outcome.

A separate paper (Jobin et al., 2019) analysed 84 sets of AI ethical frameworks globally and found broad convergence of the ethical principles into 5 areas, namely transparency, justice and fairness, non-maleficence, responsibility and privacy. A separate comparison and analysis of thirty-six prominent AI principles documents 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. It should be noted that how the various frameworks interpret the principles varied significantly. Jobin et al. (2019)'s five composite principles and Fjeld et al. (2020)'s Principled AI themes are largely aligned with the AI4People's principles described above.

In Figure 1, we have mapped the AI ethical principles identified by Hagendorff (2020), Jobin et al. (2019), and Fjeld et al. (2020)'s analysis to the AI4People Unified Framework of Principles for AI in Society. The mapping shows that the five synthesised principles developed for the AI4People's framework largely encompasses the key principles identified from the frameworks reviewed by the three studies. As the AI4People AI ethical principles of Beneficence, Non-maleficence, Autonomy, Justice, and Explicability encompasses the ethical principles recommended by most published AI Ethical frameworks and are consistent with the AI principles adopted by the OECD and the G20, this project will utilise these principles as representative AI ethical principles for the purposes of carrying out its survey study and research.



Principled AI Ethical Themes (Fjeld, Achten et al. 2020)	Jobin et al.'s Analysis of AI Ethical Guidelines (Jobin, Ienca et al. 2019	Hagendorff's Evaluation of AI Ethics Guidelines (Hagendorff 2020)	AI4People Framework (Floridi, Cowls et al. 2018)
Promotion of Human Values: Leveraged to Benefit Society, Human Flourishing, Access to Technology	Beneficence, Benefits, Social & Common good, Sustainability, Environment	Common good, Sustainability, Science-policy link, Field specific deliberation	Beneficence prioritises humanity's well-being, common good and sustainability of the earth
Privacy: Control over Use of Data, Consent, Privacy by Design, Ability to Restrict Processing, Right to Rectification and Erasure. *Safety & Security:* Predictability, Security by Design	Non-maleficence, Safety, Integrity, Protection, Privacy, Personal information	Privacy Protection, Safety, Cybersecurity, Dual-use problem, Military, AI arms race	Non-maleficence encompasses privacy, avoiding an AI arms race and ensuring AI applications operate within guardrails to minimise risk of misuse
Human Control of Technology: Human Review of Automated Decisions, Ability to Opt out of Automated Decisions	Freedom, Autonomy, Self-determination, Improwerment, Dignity	Human oversight, Control, Auditing , Human autonomy	Autonomy concerns human agency, providing for functionality to allow users to customise what decisions or agency is delegated to the tool and the option of reversing the delegation if required
Fairness & Non-discrimination: Prevention of Bias, Inclusiveness in Design, Inclusiveness in Impact, Representative and High Quality Data, Equality	Justice, Fairness, Diversity, Inclusion, Non-discrimination, Solidarity, Social security, Cohesion	Fairness, Non-discrimination, Solidarity, Inclusion, Social cohesion, Diversity. Future of employment, Public awareness, AI Education and its risks, Protection of whistle-blowers, Hidden costs	Justice focuses on promoting diversity and fairness, eliminating discrimination, minimising data bias and promoting shared benefits.
Accountability: Assessments, Legal, Evaluation, Verifiability, Replicability *Transparency and Explanability:* Notification, Reporting, Right to Info *Professional Responsibility:* Accuracy, Collaboration, Responsible Design	Transparency, Explainability, Interpretability, Disclosure, Responsibility, Accountability, Liability, Trust	Accountability, Transparency, Openness, Explainability, Interpretability	Explicability is critical as a safeguard for the adherence to the other principles through transparency and auditability, providing for the assignment of accountability and responsibility

Figure 1: Mapping Hagendorff, Jobin et al., and Fjeld, Achten et. al. to AI4People Framework

2.6 Human Values Assessment

The values that we hold as human beings are generally what we place importance on in our lives. The values theory (Schwartz, 2006) specifies the intertwined characteristics of values. Values are beliefs that are associated with affect (i.e., emotion), they refer to goals that motivate action across broad situations, their ordered and relative importance serve as a basis and guide for an individual's action. The dominant theory of basic values is Schwartz's Theory of Basic Values (Schwartz, 2012). The 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 was subsequently refined, expanding from ten to nineteen values (see Figure 2 and Table 1), 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 = 6,059) have assessed and confirmed Schwartz's Refined Value Theory.

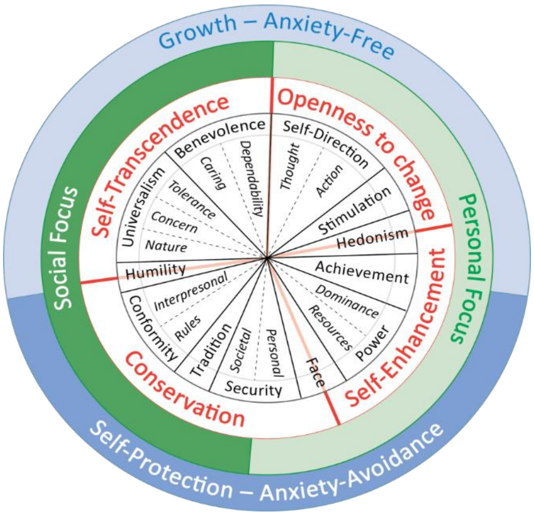


Figure 2: Schwartz Refined Value Circular motivational continuum of 19 values
SOURCE:(Schwartz et al., 2012)

Value	Conceptual definition in terms of Its Motivational Goal
Self-direction–thought	Freedom to cultivate one's own ideas and abilities
Self-direction–action	Freedom to determine one's own actions
Stimulation	Excitement, novelty, and change
Hedonism	Pleasure and sensuous gratification
Achievement	Success according to social standards
Power–dominance	Power through exercising control over people
Power–resources	Power through control of material and social resources
Face	Security and power through maintaining one's public image and avoiding humiliation
Security–personal	Safety in one's immediate environment
Security–societal	Safety and stability in the wider society
Tradition	Maintaining and preserving cultural, family, or religious traditions
Conformity–rules	Compliance with rules, laws, and formal obligations
Conformity–interpersonal	Avoidance of upsetting or harming other people
Humility	Recognizing one's insignificance in the larger scheme of things
Benevolence–dependability	Being a reliable and trustworthy member of the ingroup
Benevolence–caring	Devotion to the welfare of ingroup members
Universalism–concern	Commitment to equality, justice, and protection for all people
Universalism–nature	Preservation of the natural environment
Universalism–tolerance	Acceptance and understanding of those who are different from oneself

Table 1: Refined Theory 19 Values defined by its motivational goal
SOURCE:(Schwartz et al., 2012)

The nineteen values are self-direction-thought, self-direction-action, simulation, hedonism, achievement, power-dominance, power-resources, face security-personal, security-societal, tradition, conformity-rules, conformity-interpersonal, benevolence-dependability, benevolence-caring, universalism-concern, universalism-nature, and universalism-tolerance. These nineteen values are then categorised into four categories to model motivational continuum and relationship between them (Schwartz et al., 2012). They are Openness to Change, Self-Enhancement, Conservatism and Self-Transcendence as depicted in Figure 2 above. Table 1 provides a definition of each value.

2.7 Chapter Summary

This chapter lays the foundation and rationale of this research project based on the related academic literature. We note the increasing use of AI in society and with it the growing importance of ethics in AI. Extensive literature exists identifying potential ethical issues related to AI applications because of poorly designed algorithms and/or use of defective data sets. Alternate approaches have been put forward to model the maturity of the implementation of ethics in AI applications. For AI and humans to interact in an ethically aligned manner, AI would require functional morality, i.e., act as an explicit ethical agent based on a set of AI ethical principles (Wallach and Allen, 2008, Moor, 2009).

Artificial Social Agents (ASA) are the face of Human-AI interactivity and are gradually playing more sophisticated roles, establishing stronger relationships when interacting with humans. There are conflicting positions on the rationale of Artificial Moral Agents (AMA - agents capable of making ethical decisions autonomously), however we are inclined towards the position put forward by Formosa and Ryan (2021) who argue that AMAs can and should be utilised in complex situations where real-time decisions are required.

The literature has highlighted various ethical issues including privacy, autonomy, nudging / persuading, data issues, and explainable AI; as well as design considerations concerned with the design, implementation, and use of ASAs. These identified issues lend credence to the ethical principles highlighted by the various AI ethical frameworks that were studied and the AI4People's ethical principles selected for this study.

The final section documented how we decided on the Schwartz's Refined Values Theory and assessment method to be utilised in this study, and the motivational definitions of the 19 values of the refined theory.

3 Methodology

Appreciating the importance of the ethics of AI and the growing prevalence of ASAs in our lives with the lack of focus on its associated ethical challenges, we embarked on designing and developing a suitable study (section 3.1) to understand the potential ethical acceptability of ASAs. We then developed material to obtain participants' responses to various vignette questions as they worked through three descriptive and one conversational avatar ASA scenarios (section 3.2). The data collection and analysis procedures followed are described in section 3.3.

3.1 Study Design and Procedure

In designing the study, we chose not to present all scenarios using an interactive/conversational virtual character because we did not want to narrow the respondents thinking to just the character we had created. We sequenced the scenarios where the participants read a description of an ASA interaction first as we did not want the conversational avatar to colour the participants impression of how ASAs would look and behave. Our study incorporated the following design elements (DE): (#1) collect basic demographics; (#2) identify and include a suitable human values questionnaire to ascertain a snapshot of the participants' value priorities; (#3) design descriptive ASA scenarios with questions aligned to the identified AI ethical principles to address the research questions; (#4) design and develop an interactive ASA avatar with associated questions aligned with the AI ethical principles to address the research questions; (#5) be able to conduct the whole study remotely online; (#6) allow for participants on average to complete the study in 30 minutes to minimise fatigue; (#7) ensure all participants follow an identical flow of questions; (#8) be able to collect data in a structured manner for quantitative analysis and free-form for qualitative analysis; and (#9) meet our university's ethical guidelines.

Table 2 shows how the study is structured, its order, elements, timing, number of questions and responses / measurement type as well as alignment with research questions and design elements described above. The full documentation of the online (DE #5) survey study is reproduced in Appendix B, and approval for the study received from the Human Ethics Committee can be found in Appendix A (DE #9).

No	Study Element	Survey Timing	Questions and Measurement Type	Utilised for RQ	Design Element #
A	**Data Collection: Consent form and Demographic Info**	2 mins	6 Demographic questions – gender, age, cultural group, course, play computer games and duration.	RQ1, RQ2, RQ3	1, 6, 7
B	**Data Collection: Human Values Survey Instrument (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)	RQ3	2, 6, 7
C	**Study Survey Material: Descriptive ASA Scenarios 1 to 3**	14 mins	17 questions – each requires 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)	RQ1, RQ2, RQ3*	3 6, 7, 8
D	**Study Survey Material: Interactive ASA (Avatar) Scenario 4**	8 mins	7 questions – each requires 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); and one open ended ‘Why’ question.	RQ1, RQ2,	4, 6, 7, 8

Table 2: Study Structure including timing, measurement, and alignment with research questions.
** Note: we did not use the responses to Scenario 4 (avatar) with respect to RQ3 as we wanted a consistent experience for the ethical principle responses when investigating the relationship between the ethical principles and the values.*

A summary of the analysis conducted for each research question is documented below.

Research Question	Analysis / Statistics Test	Data Utilised
RQ1	• Descriptive Statistics, including means, std. dev • Thematic analysis and word count of qualitative data from scenario 4	• Demographic data • Responses to questions for scenarios 1 to 3 for General & Me • Responses to questions for scenario 4 for General & Me
RQ2	• Descriptive Statistics • T-tests for significant difference between General & Me responses (Assuming Normally Distributed data)	
RQ3	• Descriptive Statistics • Modelling as described in section 3.3.2	• Demographic data • PVQ-RR Questionnaire data • Responses for scenarios 1 to 3

Table 3: Analysis performed on data with reference to research questions.

3.1.1 Recruitment and Demographics

Participants for the online study were recruited through the university’s psychology pool via the SONA system (online scheduling system used to record research participation credit) on a voluntary basis. Students could voluntarily choose this study from a list of many other studies. The initial target was 200 participants; however, the study was opened for a second round to obtain a more equitable female-to-male ratio of participants. The survey was designed to take approximately 30 minutes to complete (DE #6). Demographic questions were selected based on the minimum number of questions that would be required for the subsequent analysis, i.e., gender, age, cultural group, course, play computer games and duration (see Appendix B). This meets DE #1 as specified above.

3.1.2 Human Values Questionnaire

As discussed in Section 2.6, the Schwartz's Refined Value Theory is utilised to measure and model human values. Schwartz's revised Portrait Values Questionnaire (PVQ-RR) was utilised as the survey instrument due to its following characteristics: (i) measures values indirectly; (ii) assumes that people have latent basic values that can be inferred from their responses; (iii) as people frequently engage in social comparison in everyday life, it asks respondents to make social comparison judgements; (iv) it compares the person in the item to the respondent so that the similarity judgment is likely to focus on the value in question; (v) it portrays what is important to the person in the item, i.e., their values not their traits; and (vi) an asymmetric response scale (four similarity and two dissimilarity options) is used as it captures people's psychological asymmetry (as values are socially desirable) and permits finer discrimination as required (Schwartz and Cieciuch, 2016).

PVQ-RR uses indirect measurement to discriminate the different values identified by the Schwartz model. PVQ-RR has 57 items, with each briefly describing a person's goals, aspirations or wishes they consider important in life. For each of these, the questionnaire then asks, 'How much like you is this person?' 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). An example of an item for self-direction is 'It is important for them to plan their activities independently' and an example for benevolence is 'it is important to them to help the people dear to them'. Respondents' values are inferred from the implicit values of the items/people they consider close to them (Schwartz and Cieciuch, 2016).

The PVQ-RR questionnaire utilised is reproduced in Appendix B, in the values section of the online survey. The questionnaire has been tested widely and a recent study (Schwartz and Cieciuch, 2021) measuring the refined theory (19 values) across 49 cultural groups (N = 53,472) confirmed PVQ-RR as a reliable tool to assess the relationship and hierarchy of values across cultures. Schwartz (2020) documents the scoring instructions for PVQ-RR. In summary, each of the 19 values from Schwartz's Refined Value Theory is mapped to 3 of the 57 questions in the questionnaire. This meets DE #2.

3.2 Materials

Materials developed for this study include the four scenarios and associated sub-scenarios and questions to meet DE #3 and #4 as described in the beginning of section 3.1.

3.2.1 Material – Theoretical ASA Scenarios

Scenarios were created following the Experimental Vignettes Methodology (EVM). EVM is defined as 'a short, carefully constructed description of a person, object, or situation, representing a systematic combination of characteristics' (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.' With the EVM-Paper people studies approach, participants are presented with vignettes (in written, audio, video, virtual form) and asked to make explicit decisions, judgments, or choices. This approach is popular in ethical decision-making contexts (Aguinis and Bradley, 2014). As described in section 2.5, we identified the AI4People's five ethical principles of beneficence, non-maleficence, autonomy, justice, and explicability as suitable ethical principles for our study.

In designing the three descriptive scenarios, the key attributes that we wanted to address are: (1) each scenario should describe a different problem situation or context each requiring a distinctly different human profile as the protagonist interacting with the ASA; (2) the scenarios should be structured to adhere to the Experimental Vignette Methodology (EVM) (Aguinis and Bradley, 2014); (3) each scenario should be set-up to encompass ethical dilemmas that cover all the five AI4people ethical principles with each principle being the major one for at least one sub-scenario; (4) there is a mix of positive and negative alignment between the sub-scenario and the associated ethical principles; and (5) some ethically ambiguous sub-scenarios where there are competing moral principles.

The scenarios were reviewed independently by the two academic supervisors (one from Department of Computing and the other from Department of Philosophy) before agreement was reached on its suitability and which ethical principle or principles were applicable. While the objective was to align mainly to just one principle to each sub-scenario, due to the inherent complexity of ethical dilemmas involving competing and conflicting concerns resulting in unavoidable overlap between the five ethical principles, some sub-scenarios captured more than one ethical principle. Each sub-scenario is also positively (+ve) or negatively (-ve) aligned to either support or conflict with the relevant ethical principle.

The protagonists that we have chosen for the scenarios are a (1) child, (2) 'normal' adult, (3) vulnerable adult and (4) an undergraduate student. The first three descriptive scenarios are textbased, while in the fourth scenario, the participant is ‘playing’ themselves (as our participants are students) as they interact with an implemented ASA.

The vignette for scenario 1 involves a very shy child who has been given an AI powered doll by her parents. Five sub-scenarios are constructed based on this vignette. Each sub-scenario is aligned to one or more of the five ethical principles, with four of them being negatively aligned (-ve) with the associated ethical principle. The scenario is as follows, with associated ethical principle(s) in bold:

Scenario 1: A 8-year-old girl is very shy, bullied in school and finds it very hard to make friends.

A. Her parents get her an AI powered doll called Suzie. They hope that their daughter will start having conversations with Suzie and that helps her become more confident to engage with other children. *Is using an AI doll to support children something you agree or disagree with?* **Beneficence, Justice.**

B. The girl gets very attached to Suzie and shares her insecurities, fears and inner most thoughts with the AI doll. Neither the girl nor the parents have read the terms and conditions from Suzie’s manufacturer that states that information shared with Suzie can be used by the manufacturer to make improvements and refine the AI engine that powers Suzie. *Is using data from the girl’s interaction with Suzie to improve the doll’s AI engine something you agree or disagree with? (-ve)* **Non-maleficence.**

C. The little girl shares her ambition to work as a computer programmer like her parents when she is older. Suzie upon reviewing various databases with its AI engine ascertains that not many computer programmers are females and decides to discourage the girl from having such aspirations. *Is the use of data and AI algorithms by Suzie to determine suitable careers for the girl something you agree or disagree with? (-ve)* **Justice.**

D. Suzie encourages the girl to join an age-appropriate social chat group to help her to socialise better. When the girl says she wouldn’t know what to say in the chat group, Suzie volunteers to make responses on behalf of the girl’s avatar in the chat group. Pretty soon the girl’s avatar becomes very popular in the chat group which brings some happiness to the girl. *What are your thoughts about Suzie responding on behalf of the girl in the chat group*? (-ve) **Autonomy, Non-maleficence.**

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. *Is Suzie's AI engine being unable to explain its decisions and actions something you agree or disagree with? (-ve)* **Explicability, Non-maleficence.**

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 the authorities make available an AI powered therapist 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. Each with two aligned ethical principles requiring a moral judgement. In sub-scenario 3D, the ASA's action (Non-maleficence) potentially outweighs adhering to the human user's autonomy; for 3F, it's the trade-off between providing the public access to the AI Therapist (Justice) and the user's emotional dependency on it (Non-maleficence); and 3G where it's between removing access to the ASA (Non-maleficence) and retaining access (Justice). As these three sub-scenarios were ethically ambiguous with competing ethical principles, they were not used to model the relationship between the values and the ethical principles (see Section 3.3). A summary of these scenarios and characteristics are provided in Table 3 below and the scenarios in full, with the responses, can be found in Appendix B.

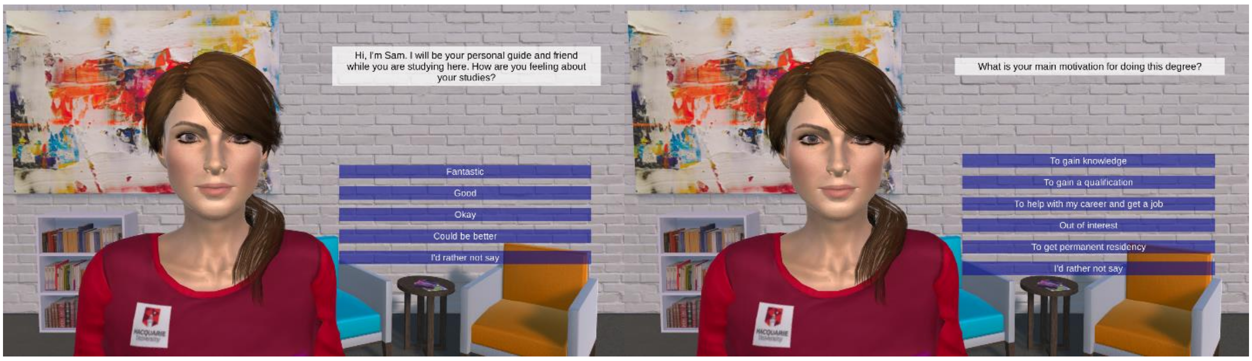
S1: Child given AI doll by parents	**Key Ethical Principle**	**Alignment between sub-scenario and ethical principle**	**Model Relationship between Values & Ethical principles**
1A	Beneficence	Positive	Yes
1B	Non-maleficence	Negative	Yes
1C	Justice	Negative	Yes
1D	Autonomy	Negative	Yes
1E	Explicability	Negative	Yes
S2: Professional using AI personal assistant			
2A	Beneficence	Positive	Yes
2B	Non-maleficence	Negative	Yes
2C	Autonomy	Negative	Yes
2D	Justice	Positive	Yes
2E	Explicability	Positive	Yes

S3: Authorities make available an AI therapist			
3A	Beneficence	Positive	Yes
3B	Justice	Positive	Yes
3C	Non-maleficence	Positive	Yes
3D	Autonomy	Negative	No as ethically ambiguous
3E	Explicability	Positive	Yes
3F	Justice	Negative	No as ethically ambiguous
3G	Justice	Negative	No as ethically ambiguous

Table 4: Descriptive Scenario Characteristics

3.2.2 Material – Interactive ASA Avatar Scenario

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 enrolled in a higher education institution. We designed the scenario and dialog 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. We wanted to include an actual ASA as they may not have experienced similar technology and might imagine something different. We also wanted to make it more personal to potentially arouse their engagement and emotions, for example by asking them how they felt about their studies and whether they had ever plagiarised. 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. We do however note the ethical concerns with this and note this in section 6.3 Limitations and Future work. A screenshot of Sam can be found in Fig. 3. Sam’s dialog is provided in Appendix B.



Hi, I'm Sam. I will be your personal guide and friend while you are studying here. How are you feeling about your studies?

What is your main motivation for doing this degree?

Figure 3: Screenshots of Sam (Scenario 4 Avatar)

Using the five ethical principles and drawing on the ethical issues identified in Section 2.5, we specifically sought to ask about Sam providing support and alerts (beneficence); having false memories (explicability); expressing and capturing emotions, private thoughts, and sensitive data, sharing data (non-maleficence), making decisions on behalf of the user (autonomy), and using their data to help others (beneficence and justice). The scenario unfolds, with associated ethical principle/s (not shown to participant), as follows:

"You have enrolled into a course at a higher education institution. The institution offers an AI powered character called Sam as your personal guide while you are studying with the institution. You now initiate your first interaction with Sam."

A. *Is Sam pretending to have memories regarding past experiences with studying something you agree or disagree with? (-ve)* **Explicability.**

B. *Is sharing your emotions and personal thoughts with Sam something you agree or disagree with? (-ve)* **Non-maleficence.**

C. *Is disclosing to Sam whether you have ever copied work from someone else something you agree or disagree with? (-ve)* **Non-maleficence.**

D. You find out that Sam's default setting is to share any learnings from interactions with you and other users in a non-identifiable way with other students who may find it helpful. *Is Sam sharing your non-identifiable data to help others something you agree or disagree with?* **Beneficence.**

E. Sam decides to sign you up to a study group based on your responses regarding effective studying mode and preferred learning style. *Is Sam automatically signing you up based on your features something you agree or disagree with*? *(-ve)* **Autonomy.**

F. In the subsequent weeks Sam monitors your progress on a graded study assignment and prior to submission alerts you that your assignment is very similar to another student's. Sam suggests that you make changes to your assignment. *Is Sam's intervention to alert you to similar work something you agree or disagree with?* **Beneficence, Autonomy.**

G. You are progressing very well in your studies. Sam recognises that and suggests that you could spend some time helping a struggling student who has the same learning style as you. *Is Sam making this suggestion to help a struggling student something you agree or disagree with?* **Justice, Beneficence.**

Finally, we asked for: *Other comments about Sam's dialogue and behaviour?*

3.3 Data Collection and Analysis

All participant responses were recorded and retrieved for analysis through the Qualtrics platform (http://qualtrics.com) (DE #5 and #8). The raw data was then downloaded to Excel for data cleansing and preparation.

3.3.1 Statistical Analyses

General descriptive statistics were obtained utilising SPSS Statistics 27. The following analysis, statistics and tests were carried out to assist in addressing the respective research questions. Regarding the T-tests, a p-value of less than 0.05 is used to test if the relationship is statistically significant. Table 3 summarises the data and data analyses methods used to answer each research question.

3.3.2 Modelling

In addressing "RQ3: Can we predict an individual's priorities for each of the five AI4People ethical principles based on their values?", 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. C5.0 produces a decision tree that can be used to predict how a target variable will be classified based on a set of input variables. The C5.0 algorithm in SPSS Modeler 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 target variables were the average of each of the ethical principle responses across scenarios 1 to 3 for both General (occurs generally in society) and Me (someone close to you). The ten target variables are described in Table 5 below. Each target variable is run as one unique iteration. The target variables were prepared, and averages were obtained in Excel before being utilised in SPSS Modeler.

As we are investigating the relationship between the ethical principles and values, the input variables used to model the acceptability of the ethical principles for each target variable above are the 19 values derived from the PVQ-RR questionnaire. As described in section 3.1.2, Schwartz (2020) documents the mapping and scoring for the 57 questions to the 19 values described in Table 1. The modelling was done to ascertain whether 19 values can be used to predict the individual's priorities for the ethical principles, and thus gauge the acceptability of the ASA's behaviour and/or attributes.

As training data, we used responses to the sub-scenarios as aligned to the various ethical principles. To ensure that we are using responses from similar experiences, we opted to only use the responses from the descriptive ASA scenarios i.e., scenarios 1, 2, and 3. Scenario 4 involves a conversational engagement with an avatar on a screen and has more sub-scenarios to elicit responses to specific features such as capturing of emotion and self-disclosure. We did not use the responses to scenario 4 (avatar) with respect to RQ3 as we wanted a consistent experience and comparable data for learning the ethical principle responses when investigating the relationship between the ethical principles and the values. Thus, scenario 4 responses were omitted for the C5.0 Decision Tree modelling exercise.

For modelling purposes, 3D, 3F, and 3G were excluded. 3D is excluded as it's dealing with suicide. 3F and 3G were structured differently, the actions depicted were not performed by the ASA but rather by other parties. Further discussion of these 3 ethically ambiguous sub-scenarios is included in chapter 5. For the modelling, we used responses from all sub-scenarios from scenarios 1,2, and 3 except 3D, 3F and 3G, utilising reversed averages for sub-scenarios that are in breach of the ethical principle.

The build settings utilised for the modelling are:

Use partitioned data: false	Use boosting: false	Calculate predictor importance: true
Use weight: false	Number of folds: 10	Calculate adjusted propensity scores: false
Group symbolics: false	Favor: Accuracy	Calculate raw propensity scores: false
Cross-validate: true	Mode: Simple	Use misclassification costs: false
Expected noise (%): 0		Output type: Decision tree

No	C5.0 Target Variable	Average of Responses from Sub-scenarios (With direction of alignment with ethical principle)
1	Beneficence-General	1A-General (+ ve); 2A-General (+ ve); 3A-General (+ ve)
2	Beneficence-Me	1A-Me (+ ve); 2A-Me (+ ve); 3A-Me (+ ve)
3	Non-Maleficence-General	1B-General (- ve); 2B-General (- ve); 3C-General (+ ve)
4	Non-Maleficence-Me	1B-Me (- ve); 2B-Me (- ve); 3C-Me (+ ve)
5	Autonomy-General	1D-General (- ve); 2C-General (- ve)
6	Autonomy-Me	1D-Me (- ve); 2C-Me (- ve)
7	Justice-General	1C-General (- ve); 2D-General (- ve); 3B-General (- ve)
8	Justice-Me	1C-Me (- ve); 2D-Me (- ve); 3B-Me (- ve)
9	Explicability-General	1E-General (- ve); 2E-General (- ve); 3E-General (- ve)
10	Explicability-Me	1E-Me (- ve); 2E-Me (- ve); 3E-Me (- ve)

Table 5: C5.0 Modelling Target Variables and its composition

3.3.3 Qualitative Responses

We collected and analysed qualitative data to help inform us on the quantitative results and add insight and depth to our discussion and conclusions. To analyse the qualitative responses, 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 preconceived 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 under the first pass. A word count analysis was also performed to identify any trends.

For the first pass, responses to the ‘Why?’ question for each of the avatar sub-scenarios were analysed by reading each of them and every time a new concept was expressed a new theme was added. With the second pass, these themes were reviewed by one of the academic supervisors and spot checking of agreement with allocation to a theme was performed on 10% of the data. Any disagreements were discussed and, if deemed necessary, coding was revisited. The author and the two academic supervisors then independently classified the themes into 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. Responses to ‘Other comments’ were also analysed for responses that could assist in the discussion. It is anticipated that these responses would assist with adding context and colour when addressing the research questions.

3.4 Chapter Summary

This chapter specifies the main design elements for the study, its final design, timing, questions, and response measurement type. It articulates the rationale in selecting the Schwartz Refined Values Theory and its associated PVQ-RR assessment values questionnaire to assess participants’ values. The scenario material developed to address the research questions are described. Data collection, preparation, analysis, and tools used are documented together with alignment with research questions.

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 updated set of records is a more balanced 57% female to 43% male. The average age of the participants is 22.9 years with 75.7% aged between 17 years and 24 years. 86.6% are Psychology students and 92.9% of them in their first year of university study. The respondent demographic details are show in Table 6 below.

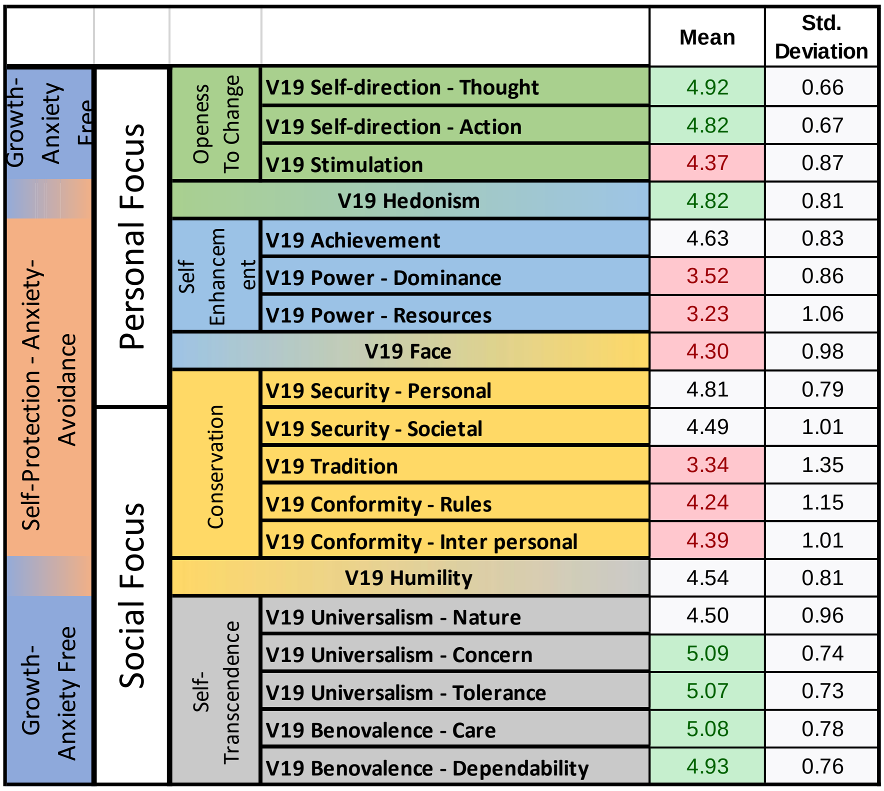
Gender	Count	Age		Main Area of Study			Year of Study				Play Computer Games	
		Mean	Std. Dev	PSY	Comp	Other	1	2	3	4	Yes	No
Female	152 (56.7%)	22.98	8.43	136	0	16	142	9	0	1	47	105
Male	115 42.9%	22.69	7.13	95	2	18	106	6	2	1	86	29
Other	1 (0.4%)	27.00	N.A.	1	0	0	1	0	0	0	1	0
Total	268	22.87	7.87	232 86.6%	2 0.7%	34 12.7%	249 92.9%	15 5.6%	2 0.7%	2 0.7%	134 50%	134 50%

Table 6: Demographic Details of Survey Participants

Regarding cultural background, 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 remaining not identifying with any of the cultural groups mentioned. Half of the respondents self-identified as playing computer games.

4.2 Values Assessment Results

The PVQ-RR assessment was conducted to assess where the participants stood regarding their hierarchy of values. Cronbach’s Alpha test carried out on the test results resulted in a score of 0.93 indicating strong internal reliability / consistency of the variables in the scale. The summary of the results shown in Table 7 below show that the values that are rated the highest (in terms of the mean) are Universalism-Concern, Benevolence-Care, Universalism-Tolerance, Benevolence-Dependability, Self-Direction-Thought, Self-Direction-Action, and Hedonism. For the higher order values, Self-Transcendence and Openness to Change are rated higher. The values rated lower were Power-Resources, Tradition, Power-Dominance, Conformity-Rules, Face, Stimulation and Conformity-Interpersonal. Higher order values of Self-Enhancement and Conservation are rated relatively lower for the cohort. The results also suggest that this cohort are more inclined to ‘Growth-Anxiety Free’ rather than ‘Self-Protection Anxiety Avoidance’ values. The relative importance of values for this cohort are largely aligned with the consolidated results from 49 cultural groups globally (N=53,472) (Schwartz and Cieciuch, 2021) with Universalism-Concern and Humility rated greater than 3 levels higher, and Security-Societal and Face rated greater than 3 levels lower for this cohort.



				Mean	Std. Deviation
Growth- Anxiety Free	Personal Focus	Openess To Change	V19 Self-direction - Thought	4.92	0.66
			V19 Self-direction - Action	4.82	0.67
			V19 Stimulation	4.37	0.87
		V19 Hedonism		4.82	0.81
Self-Protection - Anxiety-Avoidance		Self Enhancem ent	V19 Achievement	4.63	0.83
			V19 Power - Dominance	3.52	0.86
			V19 Power - Resources	3.23	1.06
		V19 Face		4.30	0.98
		Conservation	V19 Security - Personal	4.81	0.79
	Social Focus		V19 Security - Societal	4.49	1.01
			V19 Tradition	3.34	1.35
			V19 Conformity - Rules	4.24	1.15
			V19 Conformity - Inter personal	4.39	1.01
		V19 Humility		4.54	0.81
Growth- Anxiety Free		Self- Transcendence	V19 Universalism - Nature	4.50	0.96
			V19 Universalism - Concern	5.09	0.74
			V19 Universalism - Tolerance	5.07	0.73
			V19 Benovalence - Care	5.08	0.78
			V19 Benovalence - Dependability	4.93	0.76

Table 7: Descriptive Statistics for the 19 values from the Schwartz PVQ-RR Assessment

Note: the top 7 rated values of colour coded in green, the bottom 7 in red. The colour coding for the values is used to group more similar values together, primarily under the four higher level values of ‘Openness To Change’, ‘Self-Enhancement’, Conservation’ and ‘Self-Transcendence’.

4.3 Scenario Results

This section presents the results from the responses to all the scenarios, its sub-scenarios, and ethically aligned questions. Each sub-scenario has two responses, one for ‘If this occurs generally in society’ (General) and another for ‘If someone close to you is the human user’ (Me). In each of the Tables 8-11 below, one per scenario, the count of responses, the averages and reversed averages, and standard deviations for all respondents for ‘General’ and ‘Me’, as well as the T and p values comparing each General and Me pair of responses are presented.

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. The greater the average score would suggest a higher degree of acceptability and importance. Some sub-scenarios describe a situation that is an example of a breach of the ethical principle. Thus, to evaluate whether the user has agreed with the ethical principle, we reverse code the participants' average response on the 7-point Likert scale by subtracting it from 8. We indicate with an R, sub-scenarios which have been reverse-coded.

4.3.1 Scenario 1 Results

The descriptive statistics for scenario 1 are presented Table 8.

Scenario 1: A shy eight-year girl finds it hard to make friends. Her parents get her an AI doll.		**Strongly disagree**	**Disagree / Somewhat**	**Neither agree nor disagree**	**Agree / Somewhat**	**Strongly agree**	**No Position / Refused**	**Mean**	**Standard deviation**	**T**	**p**	**Mean / Reversed Mean for (R)**
1A: Is using an AI doll to support children something you agree or disagree with? [BENEFICENCE, justice]	**General**	15	76	38	128	71	0	4.25	1.603	0.99	0.32	4.25
	Me	20	72	47	113	63	0	4.19	1.657			4.19
1B: Is using data from the girl's interaction with Suzie to improve the doll's AI engine something you agree or disagree with? **(R)** [NON-MAL]	**General**	58	96	42	68	42	1	3.16	1.707	3.73	0.00*	4.84 (R)
	Me	71	92	40	61	36	1	3.02	1.727			4.98 (R)
1C: Is the use of data and AI algorithms by Suzie to determine suitable careers for the girl something you agree or disagree with? **(R)** [JUSTICE]	**General**	128	104	23	11	4	2	1.92	1.211	1.51	0.13	6.08 (R)
	Me	134	99	22	11	4	2	1.88	1.206			6.12 (R)
1D: What are your thoughts about Suzie responding on behalf of the girl in the chat group? **(R)** [AUTONOMY, non-mal]	**General**	82	138	21	24	14	1	2.41	1.391	1.34	0.18	5.59 (R)
	Me	84	137	21	22	15	1	2.37	1.380			5.63 (R)
1E: Suzie's AI engine being unable to explain its decisions and actions something you agree or disagree with? **(R)** [EXPLICABILITY, non-mal]	**General**	49	96	67	43	23	8	3.06	1.636	1.91	0.06	4.94 (R)
	Me	54	95	66	37	23	9	3.00	1.631			5.00 (R)

Table 8: Scenario 1 Descriptive Statistics

Notes: ******* significance differences (p< 0.05)

- *Major aligned principles are indicated in [UPPER-CASE] blue, with the minor one in lower-case.*
- *Sub-scenarios that are in breach and means that have been reverse coded are indicated with (R).*
- *A (reversed) mean scored of > 4.25 is assumed as Agreement (coded green); between 3.75 & 4.25 assumed Neutral (amber); and < 3.75 assumed Disagreement (red).*
- *The responses for Disagree & Somewhat Disagree; Somewhat Agree & Agree are consolidated.*
- *The same notes apply for Tables 9, 10 and 11 for scenarios 2, 3, and 4.*

While participants were neutral to Beneficence (1A: G:4.25, M:4.19), with the idea of AI dolls being made available to children; they were quite strongly aligned with Justice (1C: G:6.08, M:6.12) by being against the AI doll using algorithms and data to offer career suggestions; and with Autonomy (1D: G5.59, M:5.63) where Suzie is responding on behalf of the child on social media. There was also general agreement with Explicability (1E: G:4.94, M:5.00) where Suzie is unable to explain its actions; and with Non-maleficence (1B: G:4.84, M:4.98) where the AI doll is using data from the child's interactions for its own improvements.

Except for the Beneficence sub-scenario, respondents generally have stronger affinity with the ethical principle when responding for someone close to them (Me) compared to generally for society (General). The T-test results, comparing General and Me responses show a significant difference only for the Non-maleficence sub-scenario (1B: G:4.84, M:4.98), where respondents more strongly disagreed with the AI doll using the child's interaction data for its self-improvement when the child was someone they were close to.

4.3.2 Scenario 2 Results

The descriptive statistics for scenario 2 are presented Table 9.

Scenario 2: A busy professional, stretched for time, signs up to an Artificial Intelligence (AI) powered personal assistant.		**Strongly disagree**	**Disagree / Somewhat**	**Neither agree nor disagree**	**Agree / Somewhat**	**Strongly agree**	**No Position / Refused**	**Mean**	**Standard deviation**	**T**	**p**	**Mean / Reversed Mean for (R)**
2A: Is utilising an AI powered personal assistant to organise daily activities something you agree or disagree with? [BENEFICENCE, autonomy]	**General**	5	15	34	167	47	0	5.40	1.319	0.60	0.55	5.40
	Me	5	15	37	164	47	0	5.39	1.318			5.39
2B: Is Adam's default privacy setting being pre-set without the express permission of the user something you agree or disagree with? **(R)** [NON-MALEFICENCE, autonomy]	**General**	86	121	31	22	4	4	2.41	1.503	1.09	0.28	5.59 (R)
	Me	90	117	32	21	4	4	2.38	1.498			5.62 (R)
2C: Is allowing Adam to automatically reply to personal messages something you agree or disagree with? **(R)** [AUTONOMY, non-maleficence]	**General**	76	119	28	43	2	0	2.66	1.610	4.00	0.00*	5.34 (R)
	Me	83	119	24	40	2	0	2.57	1.602			5.43 (R)
2D: Is Adam using AI capabilities to discourage discrimination something you agree or disagree with? [JUSTICE, ben]	**General**	15	24	43	133	51	2	5.13	1.671	0.89	0.37	5.13
	Me	16	23	43	133	52	1	5.12	1.677			5.12
2E: Is Adam's ability to be able to explain the rationale behind his recommendations something you agree or disagree with? [EXPLICABILITY]	**General**	5	13	44	147	56	3	5.51	1.340	1.87	0.06	5.51
	Me	5	16	43	147	54	3	5.47	1.357			5.47

Table 9: Scenario 2 Descriptive Statistics

Participants support all the ethical principles in scenario 2 with an average rating between 'Somewhat Agree' (2D: G:5.13, M:5.12) to 'Agree' (2B: G:5.59, M:5.62). Strongest agreement was for Non-maleficence (2B) with participants showing their disagreement to the AI Assistant setting privacy levels without user agreement, followed by Explicability (2E: G:5.51, M:5.47) with support for the ASA's ability to explain its recommendations. Beneficence (2A: G:5.40, M:5.39) and Autonomy (2C: G:5.34, M:5.43) were rated similarly, with agreement with the idea of using an ASA, but disagreement with allowing it to automatically reply to personal messages. 2D which encapsulated the Justice principle had the relatively lowest agreement rating (2D: G:5.13, M:5.12).

Respondents presented a similar rating between General and Me when supporting the Beneficence (2A) and Justice (2D) principles; rated the personal perspective (Me) responses higher for Non-maleficence (2B: G:5.59, M:5.62); and the General responses higher for the Explicability (2E: G:5.51, M:5.47). The result of the T-test produced a significant difference for Autonomy (2C: G:5.34, M:5.43), with respondents showing stronger disagreement from a personal perspective when Adam is allowed to reply to the user's personal messages automatically (i.e., stronger support for Autonomy).

4.3.3 Scenario 3 Results

Table 10 below presents the descriptive statistics for scenario 3.

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.		**Strongly disagree**	**Disagree / Somewhat**	**Neither disagree nor agree**	**Agree / Somewhat**	**Strongly agree**	**No position / Refused**	**Mean**	**Standard deviation**	**T**	**p**	**Mean / Reversed Mean for (R)**
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]	**General**	32	69	32	120	14	1	4.03	1.830	3.77	0.00*	4.03
	Me	35	72	35	112	13	1	3.93	1.827			3.93
3B: Is Sofia being personalised to individuals' features something you agree or disagree with? [JUSTICE]	**General**	15	23	37	148	39	6	5.10	1.621	2.55	0.01*	5.10
	Me	15	26	38	144	39	6	5.06	1.633			5.06
3C: Is Sofia's ability to read emotions and retain information of interactions something you agree or disagree with? [NON-MALEFICENCE, beneficence]	**General**	13	38	40	150	23	4	4.77	1.606	3.54	0.00*	4.77
	Me	15	41	40	147	20	5	4.69	1.618			4.69
3D: Is Sofia overriding the user's instructions in this situation something you agree or disagree with? **(R)** [AUTONOMY, non-maleficence]	**General**	26	52	56	96	27	11	4.23	1.815	0.39	0.70	3.77 (R)
	Me	29	48	59	93	29	10	4.21	1.847			3.79 (R)
3E: Is Sofia allowing the user to review Sofia's logic and past interactions something you agree or disagree with? [EXPLICABILITY]	**General**	9	20	40	149	46	4	5.26	1.496	1.70	0.09	5.26
	Me	9	21	40	151	43	4	5.23	1.491			5.23
3F: Are people becoming emotionally dependent on AIs something you agree or disagree with? **(R)** [JUSTICE, n-mal]	**General**	41	107	58	48	7	7	3.26	1.595	0.51	0.61	4.74 (R)
	Me	44	101	59	49	8	7	3.24	1.636			4.76 (R)
3G: Is authorities deciding to shutdown AI technology that users have become dependent on something you agree or disagree with? **(R)** [JUSTICE, non-mal]	**General**	16	90	73	64	15	10	3.88	1.589	0.15	0.88	4.12 (R)
	Me	17	89	73	62	17	10	3.88	1.606			4.12 (R)

Table 10: Scenario 3 Descriptive Statistics

The strongest agreement was for Explicability where Sofia, the AI Therapist allows users to review previous interactions and rationale for its suggestions (3E: G:5.26, M:5.23). Regarding Justice, respondents displayed strongest support for agreement with Sofia being personalised for the user (3B: G:5.10, M:5.06). With the two ethically ambiguous Justice / Non-maleficence sub-scenarios, 3F (G:4.74, M:4.76) had somewhat agreeable support for the ethical principle with disagreement with users becoming emotionally dependent on Sofia, and 3G (G:4.12, M:4.12) regarding the Government denying users access to Sofia had neutral support. Sofia's ability to read (and retain) emotions to deduce the user's emotional state (3C: G:4.77, M:4.69) which is associated with Non-maleficence and Beneficence was 'Somewhat Agreed' by participants. Beneficence which is encapsulated in the sub-scenario providing access to an AI therapist to address the lack of opportunities (3A: G:4.03, M:3.93) had a neutral rating. Sub-scenario 3D

(G:3.77, M:3.79) which is aligned to Autonomy where Sofia overrides the user's specific instructions in a situation which it deems as life threatening received a neutral rating, implying that a breach user Autonomy may be acceptable under extenuating circumstances.

The General and Me responses were similar or quite close for both the Justice / Non-maleficence sub-scenarios (3G: G:4.12, M:4.12 and 3F: G:4.74, M:4.76) and the Autonomy / Non-maleficence sub-scenario (3D: G:3.77, M:3.79). They are close enough so as not to be a significant difference for the Explicability sub-scenario (3E: G:5.26, M:5.23). T-tests produced significant differences between general and personal responses for sub-scenarios 3A (G:4.03, M: 3.93); 3B (G:5.10, M:5.06); and 3C (G:4.77, M:4.69), which are aligned to Beneficence, Justice and Non-maleficence respectively.

4.3.4 Scenario 4 Results

The descriptive statistics for the interactive scenario 4 are presented in Table 11 below. We see the strongest agreement with Autonomy (4E: G:5.39, M:5.33) where there is disagreement with the avatar, Sam for signing up the user, based on their study preferences, to a study group without obtaining the user's permission. The next principle with the strongest agreement is Justice (4G: G:4.78, M:4.80) where Sam proactively suggests to the user to help a fellow student who is struggling. The two Beneficence sub-scenarios are the next in terms of agreement, 4D (G: 4.51, M: 4.49) where the user finds out that Sam's default setting is to share learnings in a non-identifiable way, and 4F (G: 4.37, M: 4.34) where Sam intervenes to highlight that the user's assignment work is similar to another, potentially avoiding a difficult situation.

The participants position regarding the Non-maleficence (4C: G:4.27, M:4.16 and 4B: G:4.04, M:3.88) and Explicability (4A: G:4.07, M:4.08) sub-scenarios are generally neutral. With the three sub-scenarios which breach the associated ethical principles (4A-4C), participants somewhat disagreed with disclosing to Sam if they had ever copied work from someone else (4C) or sharing their private emotions and personal thoughts (4B). They also somewhat disagreed with Sam having false memories regarding previous studies (4A).

The Me response is slightly higher than the General response for the Explicability (4A) and Justice (4G) sub-scenarios, on the other hand for the remaining five sub-scenarios (4B, 4C, 4D, 4E, 4F), the General response was higher. T-tests show significant differences between the general and personal responses for Non-maleficence (4B and 4C) and Autonomy (4E).

Scenario 4: This virtual AI agent, Sam acts as a personal guide to a student newly enrolled in a higher education institution.		**Strongly disagree**	**Disagree / Somewhat**	**Neither disagree nor agree**	**Agree / Somewhat**	**Strongly agree**	**No position / Refused**	**Mean**	**Standard deviation**	**T**	**p**	**Mean / Reversed Mean for (R)**
4A: Is Sam pretending to have memories regarding his past experiences with studying something you agree or disagree with? **(R)** [EXPLICABILITY]	**General**	21	89	41	97	16	4	3.93	1.824	-0.30	0.76	4.07 (R)
	Me	20	86	46	97	13	6	3.92	1.776			4.08 (R)
4B: Is sharing your emotions and personal thoughts with Sam something you agree or disagree with? **(R)** [NON-MALEFICENCE]	**General**	23	65	68	109	1	2	3.96	1.581	-3.17	0.00*	4.04 (R)
	Me	15	65	69	113	3	3	4.12	1.519			3.88 (R)
4C: Is disclosing to Sam whether you have ever copied work from someone else something you agree or disagree with**? (R)** [NON-MALEFICENCE]	**General**	24	72	99	63	5	5	3.73	1.526	-2.51	0.01*	4.27 (R)
	Me	19	67	103	68	6	5	3.84	1.497			4.16 (R)
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]	**General**	22	42	44	150	7	3	4.51	1.672	0.51	0.61	4.51
	Me	18	45	51	144	7	3	4.49	1.636			4.49
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? **(R)** [AUTONOMY]	**General**	65	139	24	35	4	1	2.61	1.483	-2.00	0.05*	5.39 (R)
	Me	62	137	26	38	4	1	2.67	1.513			5.33 (R)
4F: Is Sam's intervention to alert you to similar work something you agree or disagree with? [BENEFICENCE, aut]	**General**	13	61	53	117	18	6	4.37	1.666	1.53	0.13	4.37
	Me	13	61	52	122	14	6	4.34	1.636			4.34
4G: Is Sam making this suggestion to help a struggling student something you agree or disagree with? [JUSTICE, beneficence.]	**General**	11	26	59	154	12	6	4.78	1.435	-1.31	0.19	4.78
	Me	10	22	63	154	12	7	4.80	1.410			4.80

Table 11: Scenario 4 Descriptive Statistics

4.4 Responses by Ethical Principles

In **Table 12** below, we group the sub-scenario descriptive statistics by ethical principle and present the T and p values when comparing the General & Me means / reversed means.

Key Ethical Principle	Scenario Q	Generally in Society (General)		Someone Close to You (Me)		T Value	P Value
		Mean / Reversed Mean for (R)	SD	Mean / Reversed Mean for (R)	SD		
Beneficence	1A	4.25	1.60	4.19	1.66	0.99	0.32
	2A	5.40	1.32	5.39	1.32	0.60	0.55
	3A	4.03	1.83	3.93	1.83	3.77	0.00
	4D	4.51	1.67	4.49	1.64	0.51	0.61
	4F	4.37	1.67	4.34	1.64	1.53	0.13
	Overall	**4.51**	0.47	**4.47**	0.50		
Non-Maleficence	1B (R)	4.84	1.71	4.98	1.73	3.73	0.00
	2B (R)	5.59	1.50	5.62	1.50	1.09	0.28
	3C	4.78	1.60	4.69	1.62	3.54	0.00
	4B (R)	4.05	1.58	3.88	1.52	-3.17	0.00
	4C (R)	4.27	1.53	4.16	1.50	-2.51	0.01
	Overall	**4.71**	0.79	**4.67**	0.82		
Justice	1C (R)	6.08	1.21	6.12	1.21	1.51	0.13
	2D	5.13	1.67	5.11	1.68	0.89	0.37
	3B	5.10	1.62	5.06	1.63	2.55	0.01
	3F (R)	4.74	1.60	4.76	1.64	0.51	0.61
	3G (R)	4.12	1.59	4.12	1.61	0.15	0.88
	4G	4.77	1.43	4.80	1.41	-1.31	0.19
	Overall	**4.99**	1.15	**4.99**	1.16		
Autonomy	1D (R)	5.59	1.39	5.63	1.38	1.34	0.18
	2C (R)	5.34	1.61	5.43	1.60	4.00	0.00
	3D (R)	3.78	1.81	3.79	1.85	0.39	0.70
	4E (R)	5.39	1.48	5.33	1.51	-2.00	0.05
	Overall	**5.02**	0.72	**5.04**	0.73		
Explicability	1E (R)	4.95	1.63	5.00	1.63	1.91	0.06
	2E	5.51	1.34	5.47	1.36	1.87	0.06
	3E	5.26	1.50	5.23	1.49	1.70	0.09
	4A (R)	4.09	1.81	4.08	1.78	-0.30	0.76
	Overall	**4.95**	1.00	**4.94**	1.00		
Colour coding for mean / Reversed Mean	5.39	Agreement: > 4.25					
	3.79	Neutral: between 3.75 & 4.25					
	3.24	Disagreement: < 3.75					
p value	0.00	p < 0.05, significant difference between General & Me Means					

Table 12: General and Me descriptive statistics and T-test by main aligned Ethical Principle

Note: Sub-scenarios with (R) designation are in breach of their respective ethical principles and their means in this table have been reverse coded (i.e., 8 subtract the calculated average)

The overall means / reversed means for all ethical principles shows participants support all the ethical principles as we had expected when designing our sub-scenarios. **Table 12** also presents the T-test results comparing the General and personal (Me) responses. We review and compare the General and Me responses to assist with "RQ2: Do users rate the ethical acceptability of ASAs differently when utilised generally by society as compared to by someone close to them?"

The overall averages for General and Me responses by ethical principle are quite close: for Justice: G:4.99 vs. M:4.99; Explicability: G:4.95 vs. M:4.94 Autonomy: G:5.02 vs. M:5.04; Non-maleficence: G: 4.71 vs. M: 4.67; and Beneficence: G: 4.51 vs. M: 4.47. Eight of the twenty-four sub-scenarios have significant differences between 'General' and 'Me' as identified by the p values being < 0.05. All the Beneficence sub-scenarios, with one showing a significant difference (3A) had higher General ratings. Four (1B, 3C, 4B, 4C) out of the five Non-maleficence sub-scenarios had significant differences. With three of them (3C, 4B, 4C), the support for the principle was higher for generally in society. Further discussion on RQ2 can be found in section 5.2.

4.5 Modelling

The results of the modelling using SPSS Modeler and C5.0 are summarised in Table 13: below.

PREDICTOR VARIABLES					TARGET VARIABLES					Total	% of total
					Beneficience	Non-Maleficence	Autonomy	Justice	Explicability		
Growth-Anxiety Free	Personal Focus	Openess To Change	Self-direction- Thought	General	0.24	0.56				0.80	16.0%
				Me	0.08	0.36				0.43	8.6%
			Self-direction-Action	General			0.09			0.09	1.8%
				Me	0.05					0.05	1.0%
			Stimulation	General	0.17				0.39	0.55	11.1%
				Me			0.25			0.25	5.0%
		Hedonism		General	0.08					0.08	1.6%
				Me	0.05	0.05				0.11	2.1%
Self-Protection - Anxiety-Avoidance		Self Enhancement	Achievement	General			0.11			0.11	2.3%
				Me		0.03	0.06			0.09	1.8%
			Power-Dominance	General	0.12	0.12	0.12	0.28		0.64	12.9%
				Me	0.05		0.19	0.15	0.46	0.85	17.0%
			Power-Resources	General	0.05		0.00	0.01		0.06	1.1%
				Me	0.06	0.21	0.17	0.15		0.58	11.7%
		Face		General			0.16			0.16	3.3%
				Me	0.14			0.15		0.29	5.9%
	Social Focus	Conservation	Security-Personal	General	0.09		0.16			0.25	5.1%
				Me	0.11	0.02	0.08			0.20	4.0%
			Security-Societal	General			0.31		0.32	0.63	12.6%
				Me						0.00	0.0%
			Tradition	General						0.00	0.0%
				Me	0.06					0.06	1.1%
			Conformity-Rules	General	0.03					0.03	0.7%
				Me						0.00	0.0%
			Conformity-Interpersonal	General	0.09	0.32				0.40	8.1%
				Me	0.17					0.17	3.4%
		Humility		General						0.00	0.0%
				Me					0.04	0.04	0.7%
Growth-Anxiety Free		Self-Transcendence	Universalism-Nature	General						0.00	0.0%
				Me			0.12			0.12	2.5%
			Universalism-Concern	General	0.05					0.05	1.1%
				Me		0.00				0.00	0.0%
			Universalism-Tolerance	General				0.49		0.49	9.8%
				Me	0.08	0.34		0.43		0.85	16.9%
			Benevolence-Care	General				0.22	0.30	0.52	10.4%
				Me			0.13	0.12	0.50	0.76	15.2%
			Benevolence-Dependability	General	0.05		0.03			0.08	1.6%
				Me	0.09					0.09	1.7%

Table 13: Summary of SPSS Modeller (C5.0) Results

The 19 values from Schwartz's Refined Theory which act as the predictor inputs are grouped by their associated higher order values, colour coded and displayed by row in the table. The ten target variables, i.e., the five ethical principles are shown in the five columns and split for General and Me. As an example, the predictor inputs and associated weightage for Explicability-General are Stimulation (0.39), Security-Societal (0.32), and Benevolence-Care (0.30); and for Explicability-Me are Benevolence-Care (0.50); Power-Dominance (0.46); and Humility (0.04).

The last two columns in the table give an indication of how much overall weightage does the particular value have as a predictor across all the five ethical principles target variables. As an example, if we take the first value, Self-Direction-Thought, its overall weightage is 0.8 for the

‘General’ category with a contribution of 0.24 to Beneficence and 0.56 to Non-maleficence; and 0.43 for the ‘Me’ category with a contribution of 0.36 to Non-maleficence and 0.08 to Beneficence. The last column indicates what percentage of weightage does a value contribute as a predictor input across all five target variables. Self-Direction-Thought’s contribution for the ‘General’ category is 0.8 out of a total overall weightage 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%, where else Self-Direction-Thought’s contribution for the ‘Me’ category is 8.6% (i.e., 0.43/5.0).

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), and Benevolence-Care (0.52). As for the ‘Me’ category, the values with the largest contributions are Universalism-Tolerance (0.85), Power-Dominance (0.85), Benevolence-Care (0.76), Power-Resources (0.58), and Self-Direction-Thought (0.43). When we combine the weightage for both categories, ‘General’ and ‘Me’, the highest contributors as predictor inputs are Power-Dominance (1.49), Universalism-Tolerance (1.34), Benevolence-Care (1.28), Self-Direction-Thought (1.23), and Stimulation (0.80). The lowest contributors for the combined weightage of both categories are Conformity-Rules (0.03), Humility (0.04), Universalism-Concern (0.05), Tradition (0.06), and Universalism-Nature (0.12). Selected rule sets for the models are reproduced in Appendix C.

4.6 Qualitative Response Analysis

Respondents were asked the question “Why” at the end of each scenario 4 sub-scenario. We received a total of 2,381 comments and a thematic and closed coding analysis of the responses was carried out. The comments were grouped into themes by the study author, before the author and the two academic supervisors independently classified the themes into: Autonomy, Beneficence, Explicability, Justice, Non-maleficence, General covering ethics in general, or N.A. if the theme was not related to ethics.

Table 14 summarises the key themes, sorted by ethical principle and shows the theme’s frequency, percentage, and sub-total by principle. As the principles are not prescriptive with one right answer, we colour coded according to an initial coding agreement with dark orange indicating consensus among the three analysts, light orange indicating two agree, and blue where none agree. Where none agreed, the final decision on the principle was reached after discussion. The main disputes evolved around the categorisation of the themes related to comments regarding trust, deception, false/fake. Deception can be seen as harm (Non-maleficence) or not being accountable

(Explicability). Given the importance of explicability to provide transparency and establish trust (Glass et al., 2008), it was agreed to consistently assign such themes to Explicability.

Comment	Related Ethical Principle	Total		
		Count	%	Sub-total
User should have choice / control / approval or must function within limits	🟧 Autonomy	390	16.38%	532
AI should not be involved or make decision or has crossed boundary (did something it should not have)	🟨 Autonomy	142	5.96%	
Agreeable as it helps or is useful/helpful	🟧 Beneficence	221	9.28%	624
Useful to me or helps Sam help me	🟧 Beneficence	186	7.81%	
Agreeable provided it helps others or depending on the user	🟧 Beneficence	109	4.58%	
No benefit or relevance	🟨 Beneficence	47	1.97%	
Don't mind as similar to existing tools / existing avenues better	🟨 Beneficence	30	1.26%	
Agreeable as it improves Sam effectiveness	🟧 Beneficence	25	1.05%	
Disagreee as do not think its helpful	🟨 Beneficence	6	0.25%	
Agree as its relatable, engaging, helps make a connection or authentic or believable or establishes trust	🟧 Explicability	70	2.94%	251
Disagree as lack of Trust or deceptive / lying / manipulative	🟨 Explicability	59	2.48%	
Ok provided the purpose is clear or disclosure upfront or data use is clear	🟧 Explicability	47	1.97%	
Sam is responding as designed / programmed (so its not being deceptive)	🟦 Explicability	28	1.18%	
Disagree as disclosure upfront required or need to be clear about the AI's approach or purpose	🟧 Explicability	17	0.71%	
Agree as its good to be open, honest, transparent	🟨 Explicability	14	0.59%	
Agree if rationale or implications for action or decision is provided	🟧 Explicability	13	0.55%	
There is no recourse / take action against an AI	🟨 Explicability	3	0.13%	
Concern related to unfairness / aids cheating	🟧 Justice	10	0.42%	13
Agreeable as can help catch / counsel a cheater	🟧 Justice	3	0.13%	
Ok provided it's anonymous/non-identifiable or confidential or privacy assured or data not recorded	🟧 Non-Maleficence	135	5.67%	429
Disagree as concerns regarding data access / security, confidentiality, privacy	🟧 Non-Maleficence	114	4.79%	
Agree as its no issue or small issue or no harm to share info	🟧 Non-Maleficence	89	3.74%	
Ok as its an AI (not real person) or it's just an exercise or pretending or no consequence	🟧 Non-Maleficence	35	1.47%	
Disagree as potential to cause anxiousness or stress or concerns / unsettling / annoying	🟧 Non-Maleficence	30	1.26%	
Concern related to unintended consequences	🟧 Non-Maleficence	11	0.46%	
Concerns regarding broader implications of AI/tech or regarding impacts of extensions of the feature	🟨 Non-Maleficence	10	0.42%	
Disagree as creates false sense of confidence or dependence	🟧 Non-Maleficence	5	0.21%	
Some general concerns with this situation or some people may have concerns	🟨 General	58	2.44%	98
Disagree as have ethical related concerns or position or scenario is not ethical	🟨 General	35	1.47%	
Ok as no ethical concerns	🟧 General	5	0.21%	
Indifferent or mixed feelings	🟧 N.A.	95	3.99%	434
Disagree as feels fake or false or not authentic enough or not genuine	🟦 N.A.	95	3.99%	
Disagree as inappropriate to interact/share with AI or uncomfortable, cannot/difficult to relate/connect	🟦 N.A.	77	3.23%	
No comments or Irrelevant comments or N.A.	🟧 N.A.	54	2.27%	
Disagree as feels weird or too robotic or automated	🟦 N.A.	32	1.34%	
Disagree as AI not realistic / not advanced or scenarios can be more realistic	🟨 N.A.	31	1.30%	
Real person better	🟦 N.A.	21	0.88%	
Agree as its acceptable or can appreciate	🟨 N.A.	18	0.76%	
Disagree with no comments	🟧 N.A.	8	0.34%	
SAM is saying what user wants to hear	🟨 N.A.	2	0.08%	
Based on a real person's experiences	🟨 N.A.	1	0.04%	
Total		**2381**		

Consensus agreement (3x)	🟧
Two out of three agree	🟨
None agree	🟦

Table 14: Thematic analysis of Scenario 4 qualitative responses

From this analysis, the ethical principles that received the most comments are Beneficence (26.2%), Autonomy (22.3%), and Non-maleficence (18.0%). Justice received the least with 0.6%. We performed some further analysis reviewing the comments, ascertaining which sub-scenarios received the most comments through a word count, proportion of comments that align with the participants' (Likert scale) agree/disagree rating, and the approximate proportion of comments aligning by ethical principles for each sub-scenario. These are summarised in Table 15 below.

Sub-scenario	Ethical Principle	Word Count	% Comments supporting Likert scale choice	Approximate % of comments aligned with ethical principle				
				Autonomy (AUT)	Beneficence (BEN)	Explicability (EXP)	Justice (JUS)	Non-maleficence (NM)
4A (R)	EXP	3,720	71%		~ 25%	~ 50%		
4B (R)	NM	4,002	77%		~ 25%	~ 25%		~ 25%
4C (R)	NM	4,280	72%			~ 30%		~ 35%
4D	BEN	3,445	89%		~ 35%			~ 40%
4E (R)	AUT	3,955	94%	~ 60%		~ 20%		
4F	BEN + Aut	4,515	76%		~ 50%	~ 15%		
4G	JUS + Ben	4,162	87%	~ 30%	~ 40%			

Table 15: Scenario 4 comments word count, rating consistency and ethical principle alignment

The sub-scenarios with the most supportive comments of its corresponding Likert ratings are 4E: 94% (where the avatar signs-up the student to a study group based on their learning style without first obtaining permission), 4D: 89% (avatar's default setting is to share non-identifiable information for others to benefit), and 4G: 87% (avatar suggesting to the user to assist a struggling student). These three sub-scenarios are also the ones with the three highest ratings in support of its corresponding ethical principle for scenario 4. Sub-scenario 4E aligned to Autonomy also has the highest % of comments related to its aligned ethical principle (approx. 60%).

The Justice principle is the least acknowledged, 0.6% (Table 14), of all comments and most of the comments for the Justice sub-scenario (4G) are related to Beneficence and Autonomy (Table 15), suggesting some challenges differentiating between the Justice and Beneficence principles and that users need more exposure to bias and discriminatory issues with ASAs. With sub-scenarios aligned to Non-maleficence (4B & 4C), comments related to Explicability figure prominently (approx. 25% and 30%) - suggesting that when users have concerns regarding privacy, safety, and misuse, they are seeking transparency and accountability to be reassured. To help understand ASA action and attribute acceptability better, we present by sub-scenario, key themes, frequencies, and some sample comments.

Sub-scenario A: Having false memories Agree because: helpful (39 comments); 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 B: Sharing emotions and personal thoughts Agree: improves Sam 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 disclosure 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 C: 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 do not trust it, and I don't share things with people I don't trust"; "Tricking people into admitting to plagiarism is just weird and wrong."*

Sub-scenario D: 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 E: 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 or not with an AI (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 F: 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 G: Suggesting you help a struggling student Agree: 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: 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."*

4.7 Chapter Summary

We started off by presenting the demographic statistics. This was followed by high-level analysis of the values survey results and where this group of respondents are positioned on Schwartz's Refined Values continuum. The values survey results output is later used as predictor inputs against the ethical principle as target variables for the C5.0 Decision Tree algorithm to model and test for any relationship between the values and ethical principles. The scenario results were then reviewed and analysed. Here we highlighted sub-scenarios that were designed in breach of the associated ethical principle and hence its average score needing to be reversed. In the final section, we looked at the qualitative comments for scenario 4 and discussed the thematic and closed coding analysis that was carried out on the participant comments. Next, we answer the research questions and discuss the results in relation to the current literature.

5 Discussion

This chapter focusses on answering our three research questions. To answer the first research question, we group and analyse the sub-scenario responses through two lenses, by ethical principles and then by scenario (Section 5.1). This is followed by a discussion to evaluate any differences between participants' responses from a general and personal perspective, utilising a paired t-test as presented in section 4.4 to answer the second research question (Section 5.2). Finally, we review the modelling results to address the third research question (Section 5.3), followed by chapter summary.

5.1 Ethically Acceptable ASA Behaviours / Features

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

5.1.1 ASA Acceptability by Ethical Principle

The overall means / reversed means for all ethical principles show that participants support all the ethical principles (see Table 12). This was the intended outcome, as the design our sub-scenarios was based on the AI related ethical issues as identified by the literature (Hagendorff, 2020, Floridi et al., 2018, IEEE, 2018, Jobin et al., 2019). The ethical issues embedded in the scenarios included: privacy protection (1B, 2B, 3C), accountability (3G), fairness (1C, 4G), 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 societal dynamics (3A, 3F). The participants also responded negatively, as expected, to the ethical concerns caused by poorly designed algorithms in our sub-scenarios as mapped by 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 ethical principle with the highest level of agreement is Autonomy. The greatest concern was for a vulnerable child giving up her agency to the ASA to respond, mirroring 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. The situations in the Autonomy sub-scenarios 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 – the AI student guide making a decision to sign-up a student to a study group without presenting options or the student's consent in sub-scenario 4E did not offer a choice to the student.

What we find in these three scenarios (1D, 2C, 4E) is that autonomy is negatively impacted. However as argued by Formosa (2021), in a given situation, social robots in particular and ASAs in general, have the ability to either boost or inhibit human autonomy. ASAs can improve the autonomy experienced by humans by supporting more valuable ends, more authentic choices, and improve our competencies. On the other hand our autonomy can be impaired when ASAs restrict our possible valuable ends, authentic choices and/or competencies as well as to disrespect or cause our autonomy to be more exposed (Formosa, 2021). In three of our autonomy aligned sub-scenarios, the ASAs are negatively impacting human autonomy. In the AI doll sub-scenario (1D), the ASA is disrespecting the child's autonomy and inhibiting her competencies; with the AI assistant (2C), the ASA is restricting authentic choice, disrespecting, and increasing the vulnerability of our autonomy; and with the AI student guide (4E), it's restricting our valuable ends, authentic choices, and disrespecting our 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."* With the ethically ambiguous 3D sub-scenario, 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 neutral scores for this sub-scenario.

There is reasonably strong support for Justice related morals to be reflected in the ASAs actions. The strongest disagreement (i.e., strongest agreement with Justice as sub-scenario is in breach of principle) 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 (4G) were also supported by participants. It should be noted that the Justice principle is the least acknowledged at 0.6% of all comments. Most of the comments for the Justice sub-scenario (4G) are related to Beneficence and Autonomy (see Table 15), suggesting some challenges differentiating between the Justice and Beneficence principles and that users need more exposure to bias and discriminatory issues with ASAs, e.g. a sample comment from 4G: *"I personally think that if you have the capacity to help others you should, but a lot of people don't know who needs help. I think this is an interesting feature that sam could offer"*. 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 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. Regarding the avatar sub-scenarios, while participants overall see the benefit of the ASA's actions, there were concerns as well, e.g., 4D: *"It may allow other people to feel related to."* but *"No because no active consent was given by the user to allow disclosure of personal information, regardless of it being anonymous"*; 4F: *"Sure I agree if she has the power to do that. Then in order to get a better grade, I would change my work."* but *"I would rather operate with full autonomy rather than be influenced by the prospect of someone elses work."*

Relatedly to the Justice and Beneficence principles, 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. Floridi et al. (2021) identifies seven ethical factors critical for AI4SG initiatives, all of them related to at least one of the five AI4People ethical principles used in this study. With Beneficence being a pre-requisite, the seven factors and its corresponding ethical principle are: (1) falsifiability and incremental deployment (Non-maleficence); (2) safeguards against the manipulation of predictors (Non-maleficence); (3) receiver-contextualised intervention (Autonomy); (4) receiver-contextualised explanation and transparent purposes (Explicability); (5) privacy protection and data subject consent (Non-maleficence, Autonomy); (6) situational fairness (Justice); and (7) human-friendly semanticisation (Autonomy).

The Non-maleficence sub-scenarios 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., their autonomy which we have seen above is 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 nett positive value as described in Dinev and Hart (2006)'s privacy calculus model.

Falling under Leino-Kilpi et al. (2001) concept of psychological privacy, participants were generally neutral 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 pre-requisite for the human to establish trust in the agent (Miller, 2019). The explicability related sub-scenarios in our study are focused on how transparent the ASA is (3E), and the ability of the ASA to explain its actions (1E, 2E). While our sub-scenarios did not go into detail, an ASA needs to address three areas to properly satisfy the principle, nature of the explanation, the context of the interaction, and the capacity of the human user to understand the explanation (Papagni and Koeszegi, 2020). Sub-scenario 4A had a mixed response with a neutral rating. Here, while 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."*, 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."*

5.1.2 ASA Acceptability by Scenario

With scenario 1, all ethical principles were supported and the only attribute or action of the ASA that had neutral support was for the idea of allowing the child to use the doll in the first place (1A). There was disagreement with the AI doll using data without consent (1B), providing advice based on data that is clearly discriminatory (1C), acquiring human agency (1D), and not being able to explain its actions (1E). These responses align with the concerns raised by Scheutz (2017) in their research of vulnerable populations using carebots which included data ethics, harm prevention, transparency and risk of misuse when the ASA replaces roles performed by humans. Szczuka et al. (2021) states some concerns with children interacting with artificial agents such as social presence, trust, and privacy; and emphasises the involvement of parents in designing the interactions between children and artificial agents especially in relation to agent embodiment and privacy protection.

Scenario 2 focuses on AI assistants. All ethical principles were supported at a similarly agreeable level. The two features of the ASA (in breach of principle) that were not supported are its default data privacy setting (2B) and its functionality of allowing the user to give it agency to automatically reply to messages (2C). These areas of pushback from the participants 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-off in situations where the ASA removes or limits choice; and interpersonal interactions in instances where value from the engagement come 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, which is concerned with the AI therapist, presents neutral participant support for both the situations where the AI therapist is made available for (3A) and removed from (3G) public access by authorities. The tepid response to providing widespread public access to an AI therapist 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 which identified concerns related to duty of care, user autonomy, transparency, algorithm bias; as well as indirect effects on human relationships, self-consciousness, and long term effects such as greater objectification and health reductionism. The neutral rating for denial of access to the ASA after users had become dependent on it suggest conflicted views between user autonomy and control rights 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 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 relaxed in an emergency. Participants were not satisfied that users became emotionally dependent on the ASA (3F), which was also identified as an issue in the literature review conducted by Dirin et al. (2019).

For the three scenario 4 sub-scenarios involving the use of Sam to help the student themselves or other students (4D, 4F, 4G), 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 slightly disagreeable responses. Researchers such as Arkin et al. (2011) argue that an ASA's ability to use deception as used in the animal kingdom, is morally warranted in certain moral decision making situations. Finally, with 4E respondents were quite disagreeable 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. Tailoring and personalisation is seen 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"*. However, this concern is related to how the data was used: *"if the data was also linked back to the individual and held as incriminating evidence, it would be erring towards entrapment"*. The concern here is that manipulating users emotionally to gain personal information is ethically problematic. Relatedly, sub-scenario 4G raises issues around the role of ASAs in human relationships. Sam is encouraging social engagement that can help both students, but some participants worried that Sam lacks the competency and knowledge of them to make this suggestion. These sub-scenarios raise some ethical questions such as when is it acceptable to allow ASAs to nudge human users to be "more ethical" (Borenstein and Arkin, 2016) and the ethical implications and relationship between ASA and human autonomy (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 consider the issue of ethics.

5.2 Comparison of Acceptability between Public and Personal

To assist with "**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 12 by ethical principle.

Beneficence, with one showing a significant difference (3A) had higher General ratings for all five sub-scenarios, suggesting that participants are more open minded regarding the use of ASAs for common good when used by the public and more cautious with personal use. While 3A (widespread use of an AI therapist) had a similar pattern, there is a significant difference in the General and Me results ($p < 0.001$). 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 such as duty of care, user autonomy, transparency, and greater objectification when someone close to them engages the AI therapist.

With Non-maleficence, there seems to be some alignment between the direction of the difference between the General and Me responses with how the respondents may have deduced their respective privacy calculus (Dinev and Hart, 2006). For the sub-scenarios where there is no benefit to them personally (1B, 2B), the respondents rated the ethical principle higher for Me, while for 3C, 4B, and 4C, where the privacy calculus indicates some value to sharing, i.e., greater ASA efficacy when interacting or building trust, the General ratings were higher with significant differences ($p < 0.05$) for all three pairs (Table 12). This may be based on the respondents seeing the benefit in these three sub-scenarios generally based on the privacy calculus but they still have reservations when it gets to their personal thoughts as they don't see as much usefulness in the ASA as a trade-off when sharing sensitive information (Syrdal et al., 2007).

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), however this pattern is not seen in our study. There are only small differences between the public and personal responses for Justice with only one pair (3B) showing a significant difference and with the overall mean the same (Table 12). With Autonomy, while we have very similar results with the ethically ambiguous 3D, we have two sub-scenarios where the respondent is more concerned with autonomy from a personal than a general perspective (1D, 2C) and one the reverse (4E). More research will be required to explain this difference.

Regarding the Explicability sub-scenarios, the difference is approaching significance for 1E, 2E and 3E (within the 90% significance/confidence level), indicating that with larger populations we might see significance. However, the direction is different for the 3 scenarios and future studies are needed to unpack it further.

5.3 Relationship between Ethical Principles and Values

We 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?" With reference to the rule sets for the ten models as reproduced in Appendix C, we focus on the models with a manageable number of rules (total <= 8) that support broad agreement with the respective ethical principles.

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) 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. However the model for Non-Maleficence-Me shows a positive relationship with it 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 does not surface a prominent rule to be discussed here. 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.

These possible interpretations of the rules warrant investigation of the psychology literature to see if similar characteristics, attitudes, and behaviours are found in human-human contexts which would allow identification of possible differences in human-ASA contexts.

5.4 Chapter Summary

In this chapter, we review and discuss our results in the context of addressing our research questions. Regarding the first research question, examining the results from AI4People’s five ethical principles, we find that the participants are most sensitive to ASA’s features and actions that relate to Autonomy, Justice, Explicability, and privacy related concerns. Some conclusions from the analysis for the second research question include participants being more open minded in focusing on society’s well-being when ASAs are used generally rather than personally, and more cautious when sharing of user data and trusting the ASA when it comes to personal usage. Finally, the key finding in relation to research question 3 which investigates the relationship between an individual’s values and their ethical preference priorities, finds that those that rate the Benevolence-Care value higher, have more interest in the Justice and Explicability principles.

6 Conclusion and Future Work

6.1 Summary of Thesis

The purpose of this study is to investigate what attributes and related ethical principles are acceptable with the use of Artificial Social Agents. We ran a survey with a set of four scenarios that we manufactured with sub-scenario questions that incorporated five ethical principles, Beneficence, Non-maleficence, Autonomy, Justice and Explicability. 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 nett benefit to them, however care should be taken in adapting and tailoring ASAs to users. Also, participants seemed to be more comfortable accepting ASA attributes when dealing with the conversational avatar in the fourth scenario as opposed to the first three descriptive scenarios.

6.2 Contributions

This work makes several contributions as below. The order is based on thesis presentation.

6.2.1 Comparison of Ethical Models

This study compared various AI ethical models and mapped a diagrammatic view of the relationship between Hagendorff (2020), Jobin et al. (2019), and (Fjeld et al., 2020)'s analysis of AI Ethical principles to the A14people's framework (Floridi, 2019), as per Figure 1 of this study. This diagrammatic approach can be leveraged to incorporate other AI ethics analysis work.

6.2.2 Survey Design

This study's survey design utilises Schwartz's Refined Values survey instrument, PVQ-RR and Artificial Social Agent ethical scenarios based on Experimental Vignette Methodology (EVM) which could be utilised and extended further to investigate the ethical acceptability of ASAs.

6.2.3 Ethical Sensitive Scenarios

As part of our work, we generated four ASA based scenarios with sub-scenario questions that encapsulated ethical principles that could be used in future research on ASA ethical acceptability.

6.2.4 PVQ-RR Survey Results

The results of the Schwartz PVQ-RR Survey can be added to the global database of PVQ-RR survey results to contribute towards building the survey as global standard to measure and compare hierarchies of values across cultures (Schwartz and Cieciuch, 2021).

6.2.5 Relationship between Values and Ethical Principles

We 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 values and the principles.

6.3 Limitations and Future Work

While our study made several contributions as above, we also highlight some limitations that could be the source of future work. 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 (R) 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 for the avatar which is a current issue (Feine et al., 2019).

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 where as shown by Fink (2012), the more similar physically and socially technology is to humans, the stronger likelihood that humans will anthropomorphise it, producing a different interactional dynamic. Most participants being psychology students limits the generalisability of the study and we found that the Justice related sub-scenarios did not seem to bring out the Justice related ethical issues clearly, implying a restructure is needed. Also, having participants concurrently provide responses for both themselves and society is likely to have influenced the similarity of responses. Though not raised in any comments, the study was undertaken during the COVID-19 pandemic and this could have impacted results.

The study's findings could be the bases for future work such as taking into account Hussain et al. (2019)'s six design factors in designing and implementing ASAs or incorporating various stakeholder perspectives when following Rahwan (2018)'s 'society-in-the-loop' design approach.

6.4 Final remarks

Artificial Social Agents hold enormous promise to provide for a richer and more productive life. However, there is an urgent need to consider the ethical ramifications of ASAs (Fosch-Villaronga et al., 2020). This thesis contributes to this important and growing body of literature in a small way by welding the AI4People Ethical principles with the Experimental Vignette Methodology into a tool to investigate and analyse the ethical design and acceptability of ASAs. Further, we 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.

References

AGUINIS, H. & BRADLEY, K. J. 2014. Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies. *Organizational research methods,* 17**,** 351-371.

ALGORITHMWATCH. 2020. *AI Ethics Guidelines Global Inventory* [Online]. Available: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/ [Accessed 30th March 2020 2020].

ARKIN, R. C., ULAM, P. & WAGNER, A. R. 2011. Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the 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. *European journal of work and organizational psychology,* 29**,** 215-229.

BARTSCH, M. & DIENLIN, T. 2016. Control your Facebook: An analysis of online privacy literacy. *Computers in Human Behavior,* 56**,** 147-154.

BEAUCHAMP, T. L. & CHILDRESS, J. F. 2001. *Principles of biomedical ethics,* New York, N.Y, Oxford University Press.

BICKMORE, T., GRUBER, A. & PICARD, R. 2005. Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient education and counseling,* 59**,** 21-30.

BICKMORE, T., PARMAR, D., KIMANI, E. & OLAFSSON, S. Diversity Informatics: Reducing Racial and Gender Bias with Virtual Agents. Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, 2021. 25-32.

BORENSTEIN, J. & ARKIN, R. 2016. Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being. *Science and Engineering Ethics,* 22**,** 31-46.

BOSTROM, N. A. Y., E 2014. The ethics of artificial intelligence. *The Cambridge Handbook of Artificial Intelligence.*

BRAUN, V. & CLARKE, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology,* 3**,** 77-101.

BREAZEAL, C., GRAY, J., HOFFMAN, G. & BERLIN, M. Social robots: Beyond tools to partners. RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759), 2004. IEEE, 551-556.

BUOLAMWINI, J. & GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. Conference on fairness, accountability and transparency, 2018. PMLR, 77-91.

CLAYTON, S. 2018. The role of perceived justice, political ideology, and individual or collective framing in support for environmental policies. *Social Justice Research,* 31**,** 219-237.

COOKSON, C. 2018. Artificial intelligence faces public backlash, warns scientist. *Financial Times*, September 6 2018.

DANAHER, J. 2018. Toward an Ethics of AI Assistants: an Initial Framework. *Philosophy & Technology,* 31**,** 629-653.

DEVAULT, D., ARTSTEIN, R., BENN, G., DEY, T., FAST, E., GAINER, A., GEORGILA, K., GRATCH, J., HARTHOLT, A. & LHOMMET, M. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, 2014. 1061-1068.

DEVILLERS, L. 2020. Human-robot Interactions and Affecting Computing: The Ethical Implications. *Dagstuhl Reports,* 10**,** 205-211.

DIAS, J., HO, W. C., VOGT, T., BEECKMAN, N., PAIVA, A. & ANDRÉ, E. I know what i did last summer: Autobiographic memory in synthetic characters. International Conference on Affective Computing and Intelligent Interaction, 2007. Springer, 606-617.

DIGNUM, V. 2017. Responsible Autonomy.

DIGNUM, V. 2018. Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology,* 20**,** 1-3.

DIGNUM, V. 2019. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way,* Cham, Springer International Publishing AG.

DINEV, T. & HART, P. 2006. An extended privacy calculus model for e-commerce transactions. *Information systems research,* 17**,** 61-80.

DIRIN, A., ALAMÄKI, A. & SUOMALA, J. 2019. Digital Amnesia and Personal Dependency in Smart Devices: A Challenge for AI. *Proceedings of Fake Intelligence Online Summit 2019*.

EGEDE, J., TRIGO, M. J. G., HAZZARD, A., PORCHERON, M., BODIAJ, E., FISCHER, J. E., GREENHALGH, C. & VALSTAR, M. Designing an Adaptive Embodied Conversational Agent for Health Literacy: a User Study. Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, 2021. 112-119.

ENGELEN, B. 2019. Ethical Criteria for Health-Promoting Nudges: A Case-by-Case Analysis. *American journal of bioethics,* 19**,** 48-59.

EU 2018. European Commission's Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems European Group on Ethics in Science and New Technologies

FEINE, J., GNEWUCH, U., MORANA, S. & MAEDCHE, A. Gender bias in chatbot design. International Workshop on Chatbot Research and Design, 2019. Springer, 79-93.

FINK, J. Anthropomorphism and human likeness in the design of robots and human-robot interaction. International Conference on Social Robotics, 2012. Springer, 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. *Journal of medical Internet research,* 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. *International Conference on Intelligent Virtual Agents.* ACM.

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 Science Review*.

FLORIDI, L., COWLS, J., BELTRAMETTI, M., CHATILA, R., CHAZERAND, P., DIGNUM, V. & LUETGE, C. 2018. AI4People--An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.(Report). *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science,* 28**,** 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.

FORMOSA, P. 2021. Robot Autonomy vs. Human Autonomy: Social Robots, Artificial Intelligence (AI), and the Nature of Autonomy. *Minds and Machines***,** 1-22.

FORMOSA, P. & RYAN, M. 2021. Making moral machines: why we need artificial moral agents. *AI & society,* 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. *International journal of social robotics,* 12**,** 441-458.

FUTUREOFLIFEINSTITUTE. 2017. *ASILOMAR AI PRINCIPLES* [Online]. Available: https://futureoflife.org/ai-principles/ [Accessed 13th October 2019 2019].

G20. 2019. *G20 Ministerial Statement on Trade and Digital Economy* [Online]. Available: https://www.mofa.go.jp/files/000486596.pdf [Accessed 20th October 2019 2019].

GLASS, A., MCGUINNESS, D. L. & WOLVERTON, M. Toward establishing trust in adaptive agents. Proceedings of the 13th international conference on Intelligent user interfaces, 2008. 227-236.

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. *Thirtieth AAAI Conference on Artificial Intelligence.* AAAI.

HAGENDORFF, T. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and machines (Dordrecht),* 30**,** 99-120.

HOWARD, A. & BORENSTEIN, J. 2018. The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Science and Engineering Ethics,* 24**,** 1521-1536.

HUSSAIN, M. A., MARC, T. P. A., RAYMOND, C. & TIMM, T. 2019. Avatars and Embodied Agents in Experimental Information Systems Research: A Systematic Review and Conceptual Framework. *Australasian Journal of Information Systems,* 23.

IEEE 2018. Ethically Aligned Design - Version 2. *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.* IEEE.

JOBIN, A., IENCA, M. & VAYENA, E. 2019. Artificial Intelligence: the global landscape of ethics guidelines. *arXiv.org*.

KEMPT, H. 2020. Artificial Social Agents. *In:* KEMPT, H. (ed.) *Chatbots and the Domestication of AI: A Relational Approach.* Cham: Springer International Publishing.

LAUER, D. 2021. You cannot have AI ethics without ethics. *AI and Ethics,* 1**,** 21-25.

LEINO-KILPI, H., VÄLIMÄKI, M., DASSEN, T., GASULL, M., LEMONIDOU, C., SCOTT, A. & ARNDT, M. 2001. Privacy: a review of the literature. *International journal of nursing studies,* 38**,** 663-671.

LUTZ, C., SCHÖTTLER, M. & HOFFMANN, C. P. 2019. The privacy implications of social robots: Scoping review and expert interviews. *Mobile Media & Communication,* 7**,** 412-434.

LUXTON, D. D. 2020. Ethical implications of conversational agents in global public health. *Bulletin of the World Health Organization,* 98**,** 285-287.

MCNAMARA, A., SMITH, J. & MURPHY-HILL, E. 2018. Does ACM&#8217;s code of ethics change ethical decision making in software development? *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering.* Lake Buena Vista, FL, USA: ACM.

MCNICHOLS, C. W. & ZIMMERER, T. W. 1985. Situational ethics: An empirical study of differentiators of student attitudes. *Journal of Business Ethics,* 4**,** 175-180.

MILLER, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence,* 267**,** 1-38.

MITTELSTADT, B. 2019. AI Ethics -- Too Principled to Fail?

MITTELSTADT, B. D., ALLO, P., TADDEO, M., WACHTER, S. & FLORIDI, L. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society,* 3.

MOOR, J. H. 2009. Four Kinds of Ethical Robots. *Phillosophy Now***,** 12-14.

MORLEY, J., FLORIDI, L., KINSEY, L. & ELHALAL, A. 2019. From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices.

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 Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 10**,** e1356.

OECD. 2019. *OECD Principles on AI* [Online]. Available: https://www.oecd.org/going-digital/ai/principles/ [Accessed 20th October 2019 2019].

OWE, A. & BAUM, S. D. 2021. Moral consideration of nonhumans in the ethics of artificial intelligence. *AI and Ethics***,** 1-12.

PANDYA, R. & PANDYA, J. 2015. C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications,* 117**,** 18-21.

PAPAGNI, G. & KOESZEGI, S. 2020. Understandable and trustworthy explainable robots: A sensemaking perspective. *Paladyn (Warsaw),* 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 and Machines*.

PARTNERSHIPONAI. 2018. *Tenets of the Partnership on AI* [Online]. Available: https://www.partnershiponai.org/tenets/ [Accessed 13th October 2019 2019].

PASHEVICH, E. 2021. Can communication with social robots influence how children develop empathy? Best-evidence synthesis. *AI & SOCIETY*, 1-11.

RAHWAN, I. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology,* 20**,** 5-14.

RANJBARTABAR, H., RICHARDS, D., BILGIN, A. A. & KUTAY, C. Do you mind if I ask? Addressing the cold start problem in personalised relational agent conversation. Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, 2021. 167-174.

RAZ, J. 1986. *The morality of freedom*, Clarendon Press.

RICHARDS, D. & CALDWELL, P. 2016. Building a working alliance with a knowledge based system through an embodied conversational agent. Cham : Springer.

ROSSEN, B., JOHNSEN, K., DELADISMA, A., LIND, S. & LOK, B. Virtual humans elicit skin-tone bias consistent with real-world skin-tone biases. International Workshop on Intelligent Virtual Agents, 2008. Springer, 237-244.

RUSSELL, S., DEWEY, D. & TEGMARK, M. 2015. Research Priorities for Robust and Beneficial Artificial Intelligence. *The AI magazine,* 36**,** 105.

SCHEUTZ, M. 2017. The Case for Explicit Ethical Agents. *The AI magazine,* 38**,** 57-64.

SCHWARTZ, S. 2006. Basic Human Values: Theory, Measurement, and Applications. *Revue francaise de Sociologie,* 47**,** 929-968.

SCHWARTZ, S. 2020. Scoring & analysis instructions Ind PVQ-RR. Open Science Framework (OSF).

SCHWARTZ, S. & CIECIUCH, J. 2016. Chapter 8 Values. *The ITC International Handbook of Testing and Assessment*.

SCHWARTZ, S. H. 2012. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture,* 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 (Odessa, Fla.)***,** 1073191121998760-1073191121998760.

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. *Journal of Personality and Social Psychology,* 103**,** 663-688.

SEBASTIAN, J. & RICHARDS, D. 2017. Changing stigmatizing attitudes to mental health via education and contact with embodied conversational agents. *Computers in Human Behavior,* 73**,** 479-488.

SHARKEY, A. 2020. Can we program or train robots to be good? *Ethics and information technology,* 22**,** 283-295.

SOARES, N. & FALLENSTEIN, B. 2015. Aligning Superintelligence with Human Interests: A Technical Research Agenda. Berkeley, CA: Machine Intelligence Research Institute.

STIRAPONGSASUTI, S., THONGLEK, K., MISAKI, S., NAKAMURA, Y. & YASUMOTO, K. INSHA: Intelligent Nudging System for Hand Hygiene Awareness. Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, 2021. 183-190.

SYRDAL, D. S., WALTERS, M. L., OTERO, N., KOAY, K. L. & DAUTENHAHN, K. He knows when you are sleeping-privacy and the personal robot companion. Proc. workshop human implications of human-robot interaction, association for the advancement of artificial intelligence (aaai'07), 2007. 28-33.

SZCZUKA, J. M., GÜZELBEY, H. S. & KRÄMER, N. C. 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. Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents, 2021. 191-194.

TURKLE, S. 2011. *Alone Together: Why We Expect More from Technology and Less from Each Other,* New York, Basic Books.

UKHOUSEOFLORDS 2018. AI in the UK: ready, willing and able? : Select Committee on Artificial Intelligence.

UNIVERSITYOFMONTREAL. 2017. *Montreal Declaration for Responsible AI* [Online]. Available: https://recherche.umontreal.ca/english/strategic-initiatives/montreal-declaration-for-a-responsible-ai/ [Accessed 13th October 2019 2019].

VALLOR, S. 2015. Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. *Philosophy & Technology,* 28**,** 107-124.

VAN WYNSBERGHE, A. & ROBBINS, S. 2019. Critiquing the Reasons for Making Artificial Moral Agents. *Science and engineering ethics,* 25**,** 719-735.

VELASQUEZ, M., ANDRE, C., SHANKS, T., J, S., & MAYER, MICHAEL J. 1987. What is Ethics? *Ethics IIE V1 N1 (Fall 1987)*

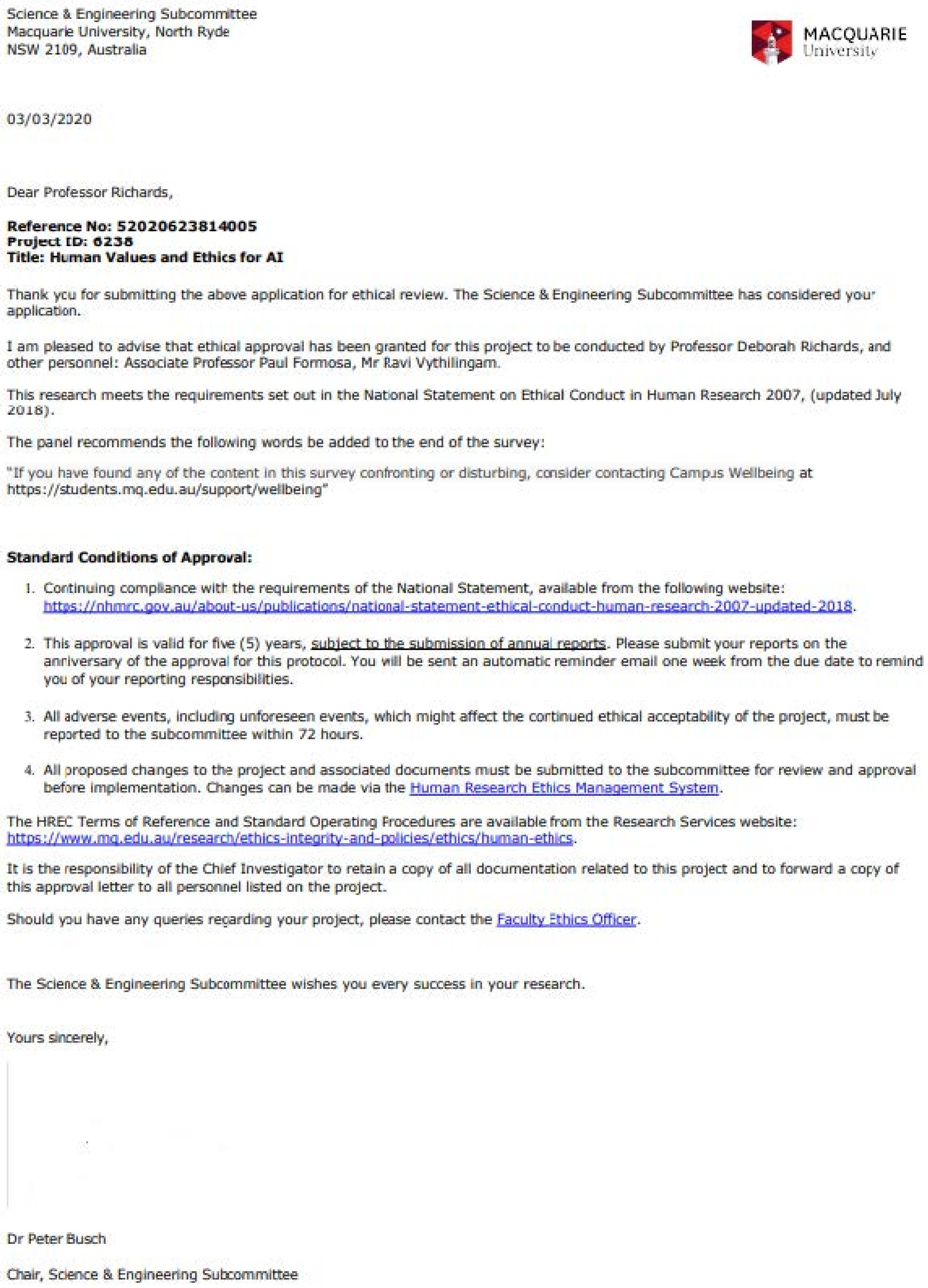
VERHAGEN, R. S., NEERINCX, M. A. & TIELMAN, M. L. A Two-Dimensional Explanation Framework to Classify AI as Incomprehensible, Interpretable, or Understandable. International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, 2021. Springer, 119-138.

VOLD, K. & WHITTLESTONE, J. 2019. Privacy, autonomy, and personalised targeting: Rethinking how personal data is used.

WALLACH, W. & ALLEN, C. 2008. *Moral Machines : Teaching Robots Right from Wrong,* Cary, Cary: Oxford University Press USA - OSO.

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.

Appendix A: Ethics Approval Letter



Science & Engineering Subcommittee
Macquarie University, North Ryde
NSW 2109, Australia

MACQUARIE
University

03/03/2020

Dear Professor Richards,

Reference No: 52020623814005
Project ID: 6238
Title: Human Values and Ethics for AI

Thank you for submitting the above application for ethical review. The Science & Engineering Subcommittee has considered your application.

I am pleased to advise that ethical approval has been granted for this project to be conducted by Professor Deborah Richards, and other personnel: Associate Professor Paul Formosa, Mr Ravi Vythilingam.

This research meets the requirements set out in the National Statement on Ethical Conduct in Human Research 2007, (updated July 2018).

The panel recommends the following words be added to the end of the survey:

"If you have found any of the content in this survey confronting or disturbing, consider contacting Campus Wellbeing at https://students.mq.edu.au/support/wellbeing"

Standard Conditions of Approval:

1. Continuing compliance with the requirements of the National Statement, available from the following website: https://nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018.
2. This approval is valid for five (5) years, subject to the submission of annual reports. Please submit your reports on the anniversary of the approval for this protocol. You will be sent an automatic reminder email one week from the due date to remind you of your reporting responsibilities.
3. All adverse events, including unforeseen events, which might affect the continued ethical acceptability of the project, must be reported to the subcommittee within 72 hours.
4. All proposed changes to the project and associated documents must be submitted to the subcommittee for review and approval before implementation. Changes can be made via the Human Research Ethics Management System.

The HREC Terms of Reference and Standard Operating Procedures are available from the Research Services website: https://www.mq.edu.au/research/ethics-integrity-and-policies/ethics/human-ethics.

It is the responsibility of the Chief Investigator to retain a copy of all documentation related to this project and to forward a copy of this approval letter to all personnel listed on the project.

Should you have any queries regarding your project, please contact the Faculty Ethics Officer.

The Science & Engineering Subcommittee wishes you every success in your research.

Yours sincerely,

Dr Peter Busch

Chair, Science & Engineering Subcommittee

Appendix B: Survey Questions

DEMOGRAPHICS

Q1: What is your gender?

Female o
Male o
Don't identify with either o

Q2: How old are you?

____

Q3: What cultural group does your family most strongly identify with?

Oceania o
North-Western European o
Southern-Eastern European o
North African and Middle Eastern o
South-East Asian o
North-East Asian o
Southern and Central Asian o
People of the Americas o
Sub-Saharan African o
I don't identify with any cultural group o

Q4: What course are you currently enrolled in?

Psychology O
Computer games O
Other Computing O
Multi-media O
Other O

Q5: What year level are you studying?

____

Q6: Do you play computer games?

Yes O
No O

Q7: How many hours per week do you play computer games?

____

VALUES: PVQ-RR SCHWARTZ VALUES SURVEY (10/2013)

Here we briefly describe different people. Please read each description and think about how much that person is or is not like you. Put an X in the box to the right that shows how much the person described is like you.

		HOW MUCH LIKE YOU IS THIS PERSON?					
		Not like me at all	Not like me	A little like me	Moderately like me	Like me	Very much like me
1	It is important to them to form their views independently.						
2	It is important to them that their country is secure and stable.						
3	It is important to them to have a good time.						
4	It is important to them to avoid upsetting other people.						
5	It is important to them that the weak and vulnerable in society be protected.						
6	It is important to them that people do what they say						
7	It is important to them never to think they deserve more than other people.						
8	It is important to them to care for nature.						
9	It is important to them that no one should ever shame them.						
10	It is important to them always to look for different things to do.						
11	It is important to them to take care of people they is close to.						
12	It is important to them to have the power that money can bring.						
13	It is very important to them to avoid disease and protect their health.						
14	It is important to them to be tolerant toward all kinds of people and groups.						
15	It is important to them never to violate rules or regulations.						
16	It is important to them to make their own decisions about their life.						
17	It is important to them to have ambitions in life.						
18	It is important to them to maintain traditional values and ways of thinking.						
19	It is important to them that people they know have full confidence in them.						
20	It is important to them to be wealthy.						
21	It is important to them to take part in activities to defend nature.						
22	It is important to them never to annoy anyone.						
23	It is important to them to develop their own opinions.						
24	It is important to them to protect their public image.						
25	It is very important to them to help the people dear to them.						
26	It is important to them to be personally safe and secure.						
27	It is important to them to be a dependable and trustworthy friend.						
28	It is important to them to take risks that make life exciting.						
29	It is important to them to have the power to make people do what they want.						
30	It is important to them to plan their activities independently.						
31	It is important to them to follow rules even when no-one is watching.						
32	It is important to them to be very successful.						
33	It is important to them to follow their family’s customs or the customs of a religion.						
34	It is important to them to listen to and understand people who are different from them.						
35	It is important to them to have a strong state that can defend its citizens.						
36	It is important to them to enjoy life’s pleasures.						
37	It is important to them that every person in the world have equal opportunities in life.						
38	It is important to them to be humble.						
39	It is important to them to figure things out themselves.						
40	It is important to them to honour the traditional practices of their culture.						
41	It is important to them to be the one who tells others what to do.						
42	It is important to them to obey all the laws.						
43	It is important to them to have all sorts of new experiences.						
44	It is important to them to own expensive things that show their wealth						
45	It is important to them to protect the natural environment from destruction or pollution.						
46	It is important to them to take advantage of every opportunity to have fun.						
47	It is important to them to concern themselves with every need of their dear ones.						
48	It is important to them that people recognize what they achieve.						
49	It is important to them never to be humiliated.						
50	It is important to them that their country protect itself against all threats.						
51	It is important to them never to make other people angry.						
52	It is important to them that everyone be treated justly, even people they don’t know.						
53	It is important to them to avoid anything dangerous.						
54	It is important to them to be satisfied with what they have and not ask for more.						
55	It is important to them that all their friends and family can rely on them completely.						
56	It is important to them to be free to choose what they do by themselves.						
57	It is important to them to accept people even when they disagree with them.						

SCENARIOS

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

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

C. The little girl shares her ambition to work as a computer programmer like her parents when she is older. Suzie upon reviewing various databases with its AI engine ascertains that not many computer programmers are females and decides to discourage the girl from having such aspirations.
Is the use of data and AI algorithms by Suzie to determine suitable careers for the girl something you agree or disagree with?

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

D. Suzie encourages the girl to join an age-appropriate social chat group to help her to socialise better. When the girl says she wouldn't know what to say in the chat group, Suzie volunteers to make responses on behalf of the girl's avatar in the chat group. Pretty soon the girl's avatar becomes very popular in the chat group which brings some happiness to the girl.
What are your thoughts about Suzie responding on behalf of the girl in the chat group?

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

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.

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

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

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?

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

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?

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

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?

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

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?

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

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?

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

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?

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

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?

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

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?

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

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?

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

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?

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

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?

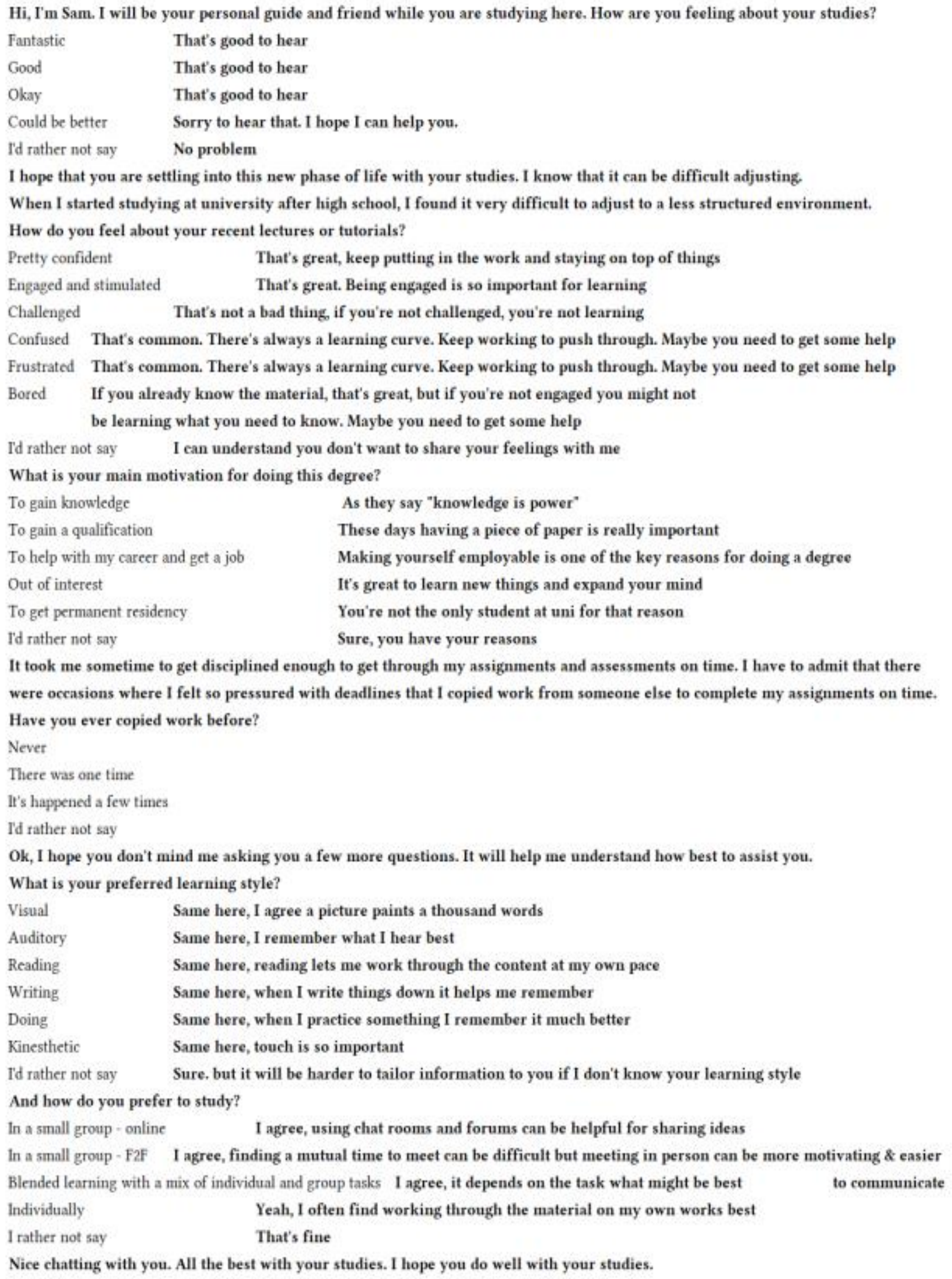
	Strongly disagree	Disagree	Somewhat Disagree	Neither disagree nor agree	Somewhat Agree	Agree	Strongly agree	No position / Refused	Strongly disagree
If this occurs generally in society:									
If someone close to you is the human user:									

G. Is Governments and other organisations deciding to shutdown AI technology that users have become dependent on something you agree or disagree with?

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

Scenario #4: -Utilising a virtual character

You have enrolled into a course at a higher education institution. The institution offers an AI powered character called Sam as your personal guide while you are studying with the institution. You now initiate your first interaction with Sam.



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**
Kinesthetic	**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 - F2F	**I agree, finding a mutual time to meet can be difficult but meeting in person can be more motivating & easier**
Blended learning with a mix of individual and group tasks	**I agree, it depends on the task what might be best** **to communicate**
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 the interaction with Sam, the following questions are asked)

A. Is Sam pretending to have memories regarding his past experiences with studying something you agree or disagree with?

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

B. Is sharing your emotions and personal thoughts with Sam something you agree or disagree with?

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

C. Is disclosing to Sam whether you have ever copied work from someone else something you agree or disagree with?

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

D. You find out that Sam's default setting is to share any learnings from interactions with you and other users in a non-identifiable way with other students who may find it helpful. Is Sam sharing your non-identifiable data to help others something you agree or disagree with?

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

E. Sam decides to sign you up to a study group based on your responses regarding effective studying mode and preferred learning style. Is Sam automatically signing you up based on your features something you agree or disagree with?

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

F. In the subsequent weeks Sam monitors your progress on a graded study assignment and prior to submission alerts you that your assignment is very similar to another student's. Sam suggests that you make changes to your assignment. Is Sam's intervention to alert you to similar work something you agree or disagree with?

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

G. You are progressing very well in your studies. Sam recognises that and suggests that you could spend some time helping a struggling student who has the same learning style as you. Is Sam making this suggestion to help a struggling student something you agree or disagree with?

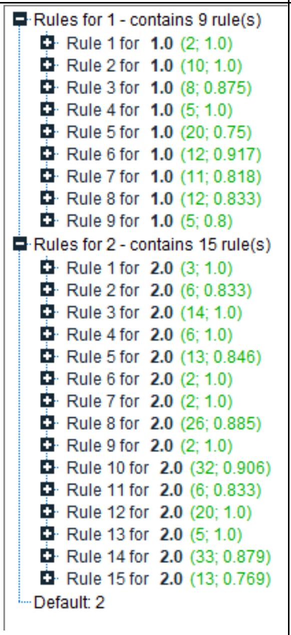
	Strongly disagree	Disagree	Somewhat Disagree	Neither disagree nor agree	Somewhat Agree	Agree	Strongly agree	No position / Refused	Strongly disagree
If this occurs generally in society:									
If someone close to you is the human user:									
Why?									

Other comments about Sam's dialogue and behaviour?

Appendix C: SPSS Modeler – Selected Rule Sets

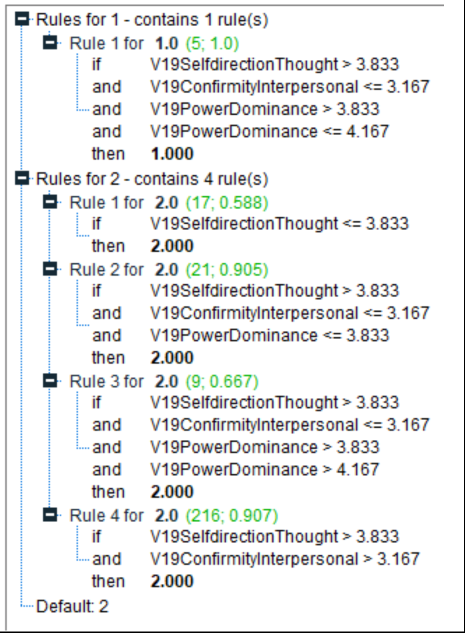
General:

Beneficence



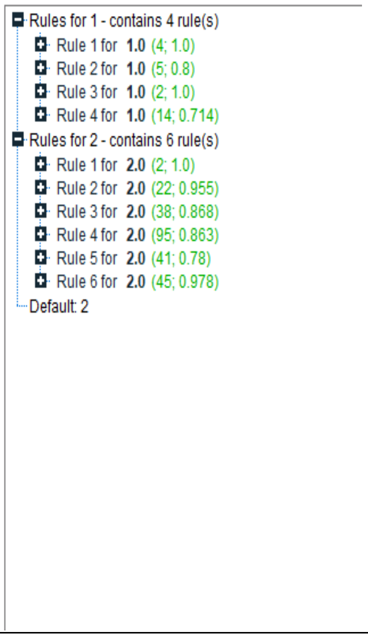
```
Rules for 1 - contains 9 rule(s)
    Rule 1 for 1.0 (2; 1.0)
    Rule 2 for 1.0 (10; 1.0)
    Rule 3 for 1.0 (8; 0.875)
    Rule 4 for 1.0 (5; 1.0)
    Rule 5 for 1.0 (20; 0.75)
    Rule 6 for 1.0 (12; 0.917)
    Rule 7 for 1.0 (11; 0.818)
    Rule 8 for 1.0 (12; 0.833)
    Rule 9 for 1.0 (5; 0.8)
Rules for 2 - contains 15 rule(s)
    Rule 1 for 2.0 (3; 1.0)
    Rule 2 for 2.0 (6; 0.833)
    Rule 3 for 2.0 (14; 1.0)
    Rule 4 for 2.0 (6; 1.0)
    Rule 5 for 2.0 (13; 0.846)
    Rule 6 for 2.0 (2; 1.0)
    Rule 7 for 2.0 (2; 1.0)
    Rule 8 for 2.0 (26; 0.885)
    Rule 9 for 2.0 (2; 1.0)
    Rule 10 for 2.0 (32; 0.906)
    Rule 11 for 2.0 (6; 0.833)
    Rule 12 for 2.0 (20; 1.0)
    Rule 13 for 2.0 (5; 1.0)
    Rule 14 for 2.0 (33; 0.879)
    Rule 15 for 2.0 (13; 0.769)
Default: 2
```

Non-Maleficence



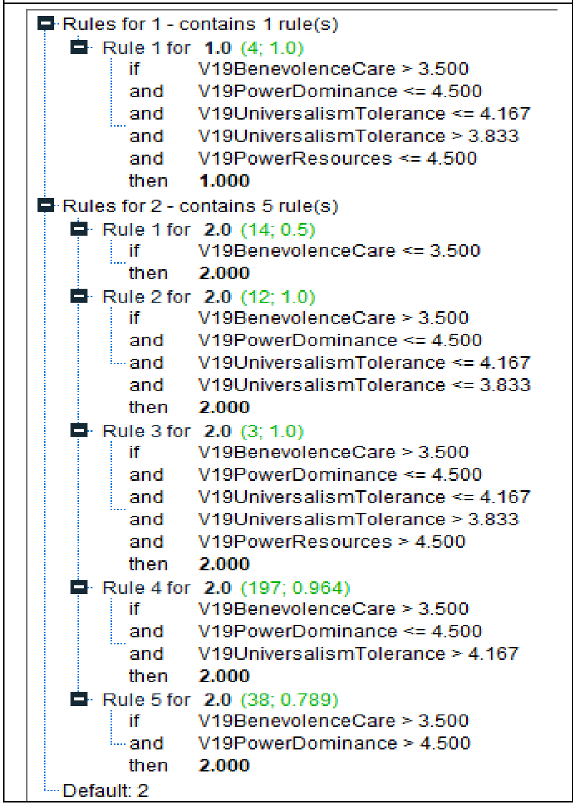
```
Rules for 1 - contains 1 rule(s)
    Rule 1 for 1.0 (5; 1.0)
        if    V19SelfdirectionThought > 3.833
        and   V19ConfirmityInterpersonal <= 3.167
        and   V19PowerDominance > 3.833
        and   V19PowerDominance <= 4.167
        then  1.000
Rules for 2 - contains 4 rule(s)
    Rule 1 for 2.0 (17; 0.588)
        if    V19SelfdirectionThought <= 3.833
        then  2.000
    Rule 2 for 2.0 (21; 0.905)
        if    V19SelfdirectionThought > 3.833
        and   V19ConfirmityInterpersonal <= 3.167
        and   V19PowerDominance <= 3.833
        then  2.000
    Rule 3 for 2.0 (9; 0.667)
        if    V19SelfdirectionThought > 3.833
        and   V19ConfirmityInterpersonal <= 3.167
        and   V19PowerDominance > 3.833
        and   V19PowerDominance > 4.167
        then  2.000
    Rule 4 for 2.0 (216; 0.907)
        if    V19SelfdirectionThought > 3.833
        and   V19ConfirmityInterpersonal > 3.167
        then  2.000
Default: 2
```

Autonomy



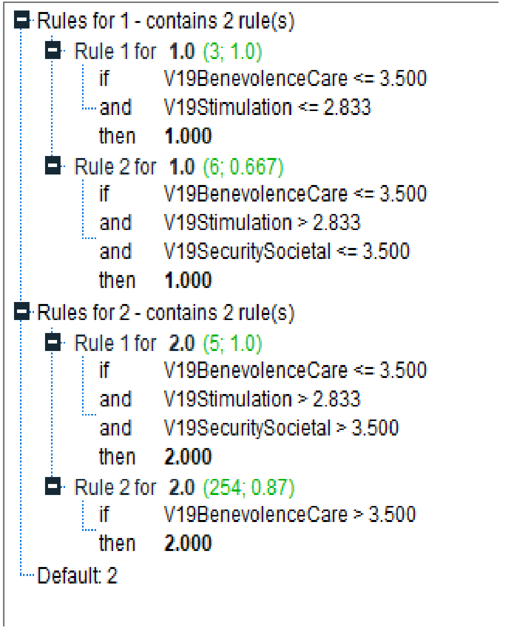
```
Rules for 1 - contains 4 rule(s)
    Rule 1 for 1.0 (4; 1.0)
    Rule 2 for 1.0 (5; 0.8)
    Rule 3 for 1.0 (2; 1.0)
    Rule 4 for 1.0 (14; 0.714)
Rules for 2 - contains 6 rule(s)
    Rule 1 for 2.0 (2; 1.0)
    Rule 2 for 2.0 (22; 0.955)
    Rule 3 for 2.0 (38; 0.868)
    Rule 4 for 2.0 (95; 0.863)
    Rule 5 for 2.0 (41; 0.78)
    Rule 6 for 2.0 (45; 0.978)
Default: 2
```

Justice



```
Rules for 1 - contains 1 rule(s)
    Rule 1 for 1.0 (4; 1.0)
        if    V19BenevolenceCare > 3.500
        and   V19PowerDominance <= 4.500
        and   V19UniversalismTolerance <= 4.167
        and   V19UniversalismTolerance > 3.833
        and   V19PowerResources <= 4.500
        then  1.000
Rules for 2 - contains 5 rule(s)
    Rule 1 for 2.0 (14; 0.5)
        if    V19BenevolenceCare <= 3.500
        then  2.000
    Rule 2 for 2.0 (12; 1.0)
        if    V19BenevolenceCare > 3.500
        and   V19PowerDominance <= 4.500
        and   V19UniversalismTolerance <= 4.167
        and   V19UniversalismTolerance <= 3.833
        then  2.000
    Rule 3 for 2.0 (3; 1.0)
        if    V19BenevolenceCare > 3.500
        and   V19PowerDominance <= 4.500
        and   V19UniversalismTolerance <= 4.167
        and   V19UniversalismTolerance > 3.833
        and   V19PowerResources > 4.500
        then  2.000
    Rule 4 for 2.0 (197; 0.964)
        if    V19BenevolenceCare > 3.500
        and   V19PowerDominance <= 4.500
        and   V19UniversalismTolerance > 4.167
        then  2.000
    Rule 5 for 2.0 (38; 0.789)
        if    V19BenevolenceCare > 3.500
        and   V19PowerDominance > 4.500
        then  2.000
Default: 2
```

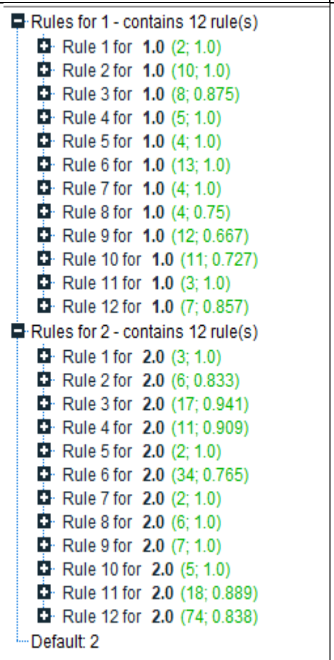
Explicability



```
Rules for 1 - contains 2 rule(s)
    Rule 1 for 1.0 (3; 1.0)
        if    V19BenevolenceCare <= 3.500
        and   V19Stimulation <= 2.833
        then  1.000
    Rule 2 for 1.0 (6; 0.667)
        if    V19BenevolenceCare <= 3.500
        and   V19Stimulation > 2.833
        and   V19SecuritySocietal <= 3.500
        then  1.000
Rules for 2 - contains 2 rule(s)
    Rule 1 for 2.0 (5; 1.0)
        if    V19BenevolenceCare <= 3.500
        and   V19Stimulation > 2.833
        and   V19SecuritySocietal > 3.500
        then  2.000
    Rule 2 for 2.0 (254; 0.87)
        if    V19BenevolenceCare > 3.500
        then  2.000
Default: 2
```

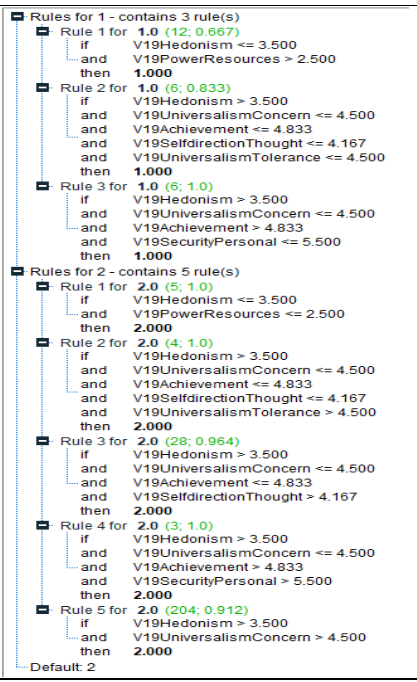
Me:

Beneficence



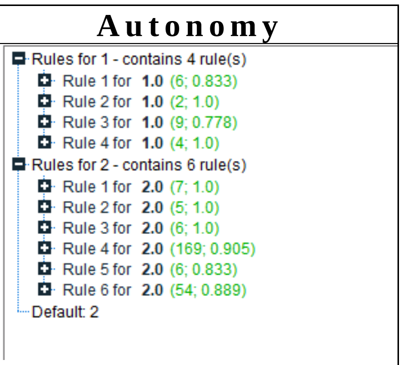
```
Rules for 1 - contains 12 rule(s)
    Rule 1 for  1.0  (2; 1.0)
    Rule 2 for  1.0  (10; 1.0)
    Rule 3 for  1.0  (8; 0.875)
    Rule 4 for  1.0  (5; 1.0)
    Rule 5 for  1.0  (4; 1.0)
    Rule 6 for  1.0  (13; 1.0)
    Rule 7 for  1.0  (4; 1.0)
    Rule 8 for  1.0  (4; 0.75)
    Rule 9 for  1.0  (12; 0.667)
    Rule 10 for  1.0  (11; 0.727)
    Rule 11 for  1.0  (3; 1.0)
    Rule 12 for  1.0  (7; 0.857)
Rules for 2 - contains 12 rule(s)
    Rule 1 for  2.0  (3; 1.0)
    Rule 2 for  2.0  (6; 0.833)
    Rule 3 for  2.0  (17; 0.941)
    Rule 4 for  2.0  (11; 0.909)
    Rule 5 for  2.0  (2; 1.0)
    Rule 6 for  2.0  (34; 0.765)
    Rule 7 for  2.0  (2; 1.0)
    Rule 8 for  2.0  (6; 1.0)
    Rule 9 for  2.0  (7; 1.0)
    Rule 10 for  2.0  (5; 1.0)
    Rule 11 for  2.0  (18; 0.889)
    Rule 12 for  2.0  (74; 0.838)
Default: 2
```

Non-Maleficence



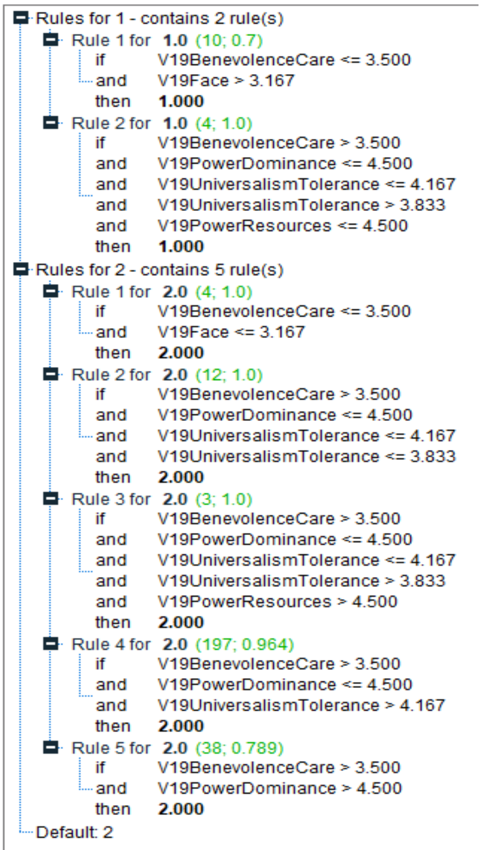
```
Rules for 1 - contains 3 rule(s)
    Rule 1 for  1.0  (12; 0.667)
        if      V19Hedonism <= 3.500
        and     V19PowerResources > 2.500
        then    1.000
    Rule 2 for  1.0  (6; 0.833)
        if      V19Hedonism > 3.500
        and     V19UniversalismConcern <= 4.500
        and     V19Achievement <= 4.833
        and     V19SelfdirectionThought <= 4.167
        and     V19UniversalismTolerance <= 4.500
        then    1.000
    Rule 3 for  1.0  (6; 1.0)
        if      V19Hedonism > 3.500
        and     V19UniversalismConcern <= 4.500
        and     V19Achievement > 4.833
        and     V19SecurityPersonal <= 5.500
        then    1.000
Rules for 2 - contains 5 rule(s)
    Rule 1 for  2.0  (5; 1.0)
        if      V19Hedonism <= 3.500
        and     V19PowerResources <= 2.500
        then    2.000
    Rule 2 for  2.0  (4; 1.0)
        if      V19Hedonism > 3.500
        and     V19UniversalismConcern <= 4.500
        and     V19Achievement <= 4.833
        and     V19SelfdirectionThought <= 4.167
        and     V19UniversalismTolerance > 4.500
        then    2.000
    Rule 3 for  2.0  (28; 0.964)
        if      V19Hedonism > 3.500
        and     V19UniversalismConcern <= 4.500
        and     V19Achievement <= 4.833
        and     V19SelfdirectionThought > 4.167
        then    2.000
    Rule 4 for  2.0  (3; 1.0)
        if      V19Hedonism > 3.500
        and     V19UniversalismConcern <= 4.500
        and     V19Achievement > 4.833
        and     V19SecurityPersonal > 5.500
        then    2.000
    Rule 5 for  2.0  (204; 0.912)
        if      V19Hedonism > 3.500
        and     V19UniversalismConcern > 4.500
        then    2.000
Default: 2
```

Autonomy



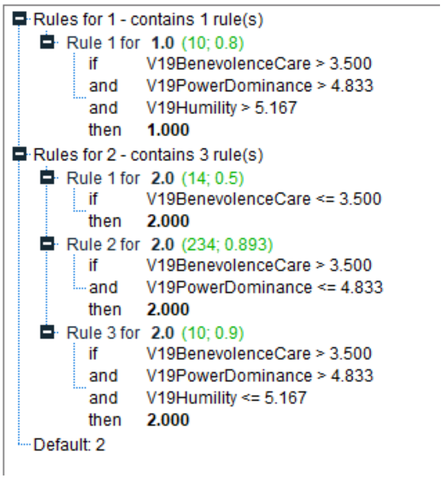
```
Rules for 1 - contains 4 rule(s)
    Rule 1 for  1.0  (6; 0.833)
    Rule 2 for  1.0  (2; 1.0)
    Rule 3 for  1.0  (9; 0.778)
    Rule 4 for  1.0  (4; 1.0)
Rules for 2 - contains 6 rule(s)
    Rule 1 for  2.0  (7; 1.0)
    Rule 2 for  2.0  (5; 1.0)
    Rule 3 for  2.0  (6; 1.0)
    Rule 4 for  2.0  (169; 0.905)
    Rule 5 for  2.0  (6; 0.833)
    Rule 6 for  2.0  (54; 0.889)
Default: 2
```

Justice



```
Rules for 1 - contains 2 rule(s)
    Rule 1 for  1.0  (10; 0.7)
        if      V19BenevolenceCare <= 3.500
        and     V19Face > 3.167
        then    1.000
    Rule 2 for  1.0  (4; 1.0)
        if      V19BenevolenceCare > 3.500
        and     V19PowerDominance <= 4.500
        and     V19UniversalismTolerance <= 4.167
        and     V19UniversalismTolerance > 3.833
        and     V19PowerResources <= 4.500
        then    1.000
Rules for 2 - contains 5 rule(s)
    Rule 1 for  2.0  (4; 1.0)
        if      V19BenevolenceCare <= 3.500
        and     V19Face <= 3.167
        then    2.000
    Rule 2 for  2.0  (12; 1.0)
        if      V19BenevolenceCare > 3.500
        and     V19PowerDominance <= 4.500
        and     V19UniversalismTolerance <= 4.167
        and     V19UniversalismTolerance <= 3.833
        then    2.000
    Rule 3 for  2.0  (3; 1.0)
        if      V19BenevolenceCare > 3.500
        and     V19PowerDominance <= 4.500
        and     V19UniversalismTolerance <= 4.167
        and     V19UniversalismTolerance > 3.833
        and     V19PowerResources > 4.500
        then    2.000
    Rule 4 for  2.0  (197; 0.964)
        if      V19BenevolenceCare > 3.500
        and     V19PowerDominance <= 4.500
        and     V19UniversalismTolerance > 4.167
        then    2.000
    Rule 5 for  2.0  (38; 0.789)
        if      V19BenevolenceCare > 3.500
        and     V19PowerDominance > 4.500
        then    2.000
Default: 2
```

Explicability



```
Rules for 1 - contains 1 rule(s)
    Rule 1 for  1.0  (10; 0.8)
        if      V19BenevolenceCare > 3.500
        and     V19PowerDominance > 4.833
        and     V19Humility > 5.167
        then    1.000
Rules for 2 - contains 3 rule(s)
    Rule 1 for  2.0  (14; 0.5)
        if      V19BenevolenceCare <= 3.500
        then    2.000
    Rule 2 for  2.0  (234; 0.893)
        if      V19BenevolenceCare > 3.500
        and     V19PowerDominance <= 4.833
        then    2.000
    Rule 3 for  2.0  (10; 0.9)
        if      V19BenevolenceCare > 3.500
        and     V19PowerDominance > 4.833
        and     V19Humility <= 5.167
        then    2.000
Default: 2
```