TUM School of Management
Technische Universität München

From Pixels to Principles:

Enhancing Interactional Technology Ethics through Psychological Insights

Auxane N. G. Boch

Complete reprint of the dissertation approved by the TUM School of Management of the Technical University of Munich for the award of the

Doktorin der Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.).

Chair: Prof. Dr. Jutta Roosen

Examiner:
1. Prof. Dr. Christoph Lütge
2. Prof. Dr. Jochen Hartmann

The dissertation was submitted to the Technical University of Munich on 21.03.2025 and accepted by the TUM School of Management on 15.06.2025.

"I am not one of those who fear the future.

The younger generations sometimes surprise us because they differ from us;

we ourselves raised them differently from the way we were raised.

But this youth is courageous, capable of enthusiasm and sacrifice, just like

the others. Let us trust them to preserve the supreme value of life."

- Simone Veil to the French Assemblée Nationale, 26 novembre 1974.

Abstract

What are the psychological impacts of interactional technologies like large language models, social robots, and video games, and how can these insights inform ethical frameworks for their design and deployment? As interactional systems become more integrated into daily life, their potential to shape cognition, behaviour, and well-being raises ethical concerns. This dissertation explores the intersection of psychology and technology ethics, examining how psychological principles can, in practice, enhance ethical guidelines for interactive and interactional technologies. Through four contributions, this dissertation investigates different aspects of psychological impacts and their ethical implications. The first paper addresses ethical dilemmas in using care robots in healthcare, particularly focusing on the emotional and cognitive dependencies and biases they may foster. The second paper explores the ethical challenges in human-robot interaction, highlighting how psychological insights into human culture, empathy and attachment can inform the design of social robots. The third paper examines the ethical landscape of video games, analysing how game mechanics influence cognitive and social skills, mental health, and moral decision-making. The fourth paper uses narrative-driven video games to investigate how interactions with android characters influence moral judgments and attitudes toward AI systems. These contributions build a comprehensive ethical framework integrating psychological research into existing ethical guidelines. The findings underscore the

importance of well-being, autonomy, and cultural sensitivity in developing interactional technologies. This thesis advocates for a nuanced approach to ethics, where the psychological effects of technology are systematically considered to ensure responsible innovation. After presenting the framework and results of the four papers, the concluding section reflects on the broader implications for technology ethics. It highlights the need for ongoing interdisciplinary collaboration between psychology and ethics.

Acknowledgements

First of all, I sincerely thank my supervisor, Prof. Christoph Lütge, for giving me an ideal environment to grow as a researcher and person, supporting and supervising me with wisdom, kindness and diligence, and trusting me to go about my ideas freely. Merci milles fois! Furthermore, I would like to extend this gratitude to Prof. Jochen Hartmann and Prof. Jutta Roosen for agreeing to be my second supervisor and chairwoman of the examination. As in every good adventure, it takes a village. To Jaimee Stuart, Liam Cross and Gray Atherton, thank you for being great advisors, examples, mentors, and friends! To Franziska Poszler, thank you for always being open to discussing, helping me think, and being a great example to follow. To Raphael Max, thank you for debating with me, agreeing to disagree, and teaching me things, always with patience. To Seamus Ryan, thank you for being my PhD Buddy, an undying supporter, a great reviewer, and a hilarious friend with whom to vent. To Bethany Thomas, thank you for being an amazing teammate, an inspiration, and a great friend! An enormous thanks to each of my co-authors on this thesis. And there would be many more to thank, such as the Caitlin Corrigan and the IEAI team, the Think Tank Team, my Immersive Realities Working group co-leads, and the Women in AI German team for giving me a space to always go further and for a positive and endearing environment to have crazy ideas and realise them in. But also, to my friends here in Munich, Elizabeth, Mohammed, Philip, Brittany and to my friends abroad with whom we grew throughout the

years, sometimes at a distance but never apart, Marie, Alia, Cristalline, Perle, Elsa, Dr. Clément Caiazzo, Marion, Valentine, Clara, Juliette, Justine, and Anais, thank you for believing this was going to happen when I didn't, and for always being only a text or a phone call away. Finally, my biggest gratitude goes to my parents, who supported and believed in me throughout all of it and without whom this would not have been possible. This success I share with you. In memory of my Yaya, who keep inspiring me from where she is.

Table of Contents

"Zealous conviction is a dangerous substitute for an open mind."

Elizabeth Loftus in *The Reality of Repressed Memories* (1993)

1 Introduction

The psychological impacts of technology have become a significant topic in public discourse, potentially leading to tumult for the general population and, ensuingly, moral panics. Moral panics are scenarios in which an entity "emerges to become defined as a threat to societal values and interests" (Cohen, 1973). These scenarios often appear in media and public discourse, where certain acts are depicted as deviant. A prominent example of moral panic is the concern surrounding violent video games. Kneer and Ward (2020) discussed how video games have been cited as an essential cause of mass shootings, such as exemplified by the Columbine High School massacre on April 20, 1999, where the perpetrators were known to play games like *Doom* and *Quake* (Campbel, 2018). Interestingly, Campbel (2018) reports that only 4 of the 33 active school shooters studied between 1980 and 2014 were fans of video games. Just as violent video games have sparked debates on psychological impacts, the emergence of AI systems similarly invites moral panic, especially when psychological harm is insufficiently understood or defined. In this line, another example is the British opposition to sex robots, led by movements such as *the Campaign Against Sex Robots*[1] and robotics researcher Dr Kathleen Richardson. Richardson (2016) posits that new technology, such as sex robots, supports and expands

[1] https://campaignagainstsexrobots.org/

the sex industry, citing the internet's role in the industry's growth while also perpetuating the potential stigma of women being seen as objects to be used due to the human ability of anthropomorphisation of robots. On the other hand, some scholars counterargue, endorsing sex robots as a potential legal replacement for human sex workers (Danaher, 2014) and as a solution to other social problems (Eggleeton, 2019). As generative AI technologies advance, they have sparked similar concerns over psychological impacts, paralleling earlier fears associated with violent video games and sex robots. Indeed, a more recent example concerns the discourse on generative AI as a technology able to contribute to suicidal behaviour through disturbing or manipulative content (White, 2023).

These concerns underscore the disruptive nature of technological developments, which challenge and displace established societal norms and routines (Chandler, 1995; Novicevic et al., 2009). Notably, these disruptions can influence policy discussions, leading to the creation of regulatory bodies and policies aimed at addressing societal issues. For example, the Pan European Game Information (PEGI) and the Entertainment Software Rating Board (ESRB) are rating systems for video games developed in response to public concerns about their potential impact on children and adolescents. On the other hand, in response to data privacy and technology ethics concerns, regulations like the General Data Protection Regulation (GDPR) in the European Union (EU) have been introduced, highlighting the impact of social discourse on legislative action (White, 2023).

Following a similar trend, the psychological impact of AI is increasingly recognised in international texts and ethical frameworks, reflecting a growing awareness of its crucial role in understanding technological impacts. These impacts may include altered emotional well-being, social dynamics, cognitive processing, or developmental trajectories, increasingly linked to interactive and interactional technologies. UNESCO's (2021) "Recommendation on the Ethics of Artificial Intelligence" highlights concerns about AI

systems challenging human experience and agency, urging states to investigate the psychological and cognitive effects on children and adolescents. The WHO's (2024) "Ethics and Governance of Artificial Intelligence for Health" emphasises avoiding AI applications that cause preventable psychological harm. The European regulation laying down harmonised rules on artificial intelligence (EU AI Act, 2024) also addresses the psychological risks of AI systems by mentioning their potential for subconsciously manipulating behaviour. However, those texts lack precise definitions of psychological harm, particularly in how psychological impacts are measured and addressed, complicating their integration into actionable policy frameworks (Leiser, 2024).

Therefore, there is a clear call for a better understanding of the psychological impacts of technologies and their integration into the ethics and policy frameworks. This thesis offers a first step towards addressing this issue by proposing methodological approaches to integrating psychology research outcomes into technology ethics and policy discourse, focusing on interactive and interactional systems, together classified here as interactional technologies.

Interactive systems, such as video games and immersive realities, engage users by enabling them to manipulate and interact with digital environments in real-time. These tools are central to creating extended reality experiences (Coronado et al., 2023). In contrast, interactional systems, including chatbots, large language models (LLMs), and social robots, are designed to facilitate communication between humans and machines through, for instance, natural language processing and social interactions. These tools simulate animal or human-like conversations and interactions, enhancing user experience (Rajesh et al., 2023). Understanding the differences between these technologies is crucial, as they each engage users' psychological processes in unique ways - whether through immersive

environments or social interactions - posing distinct ethical challenges. These technologies can also converge, for instance, in video games like *Red Dead Redemption 2* with AI-powered non-playable characters.

Those interactional technologies have the potential to significantly impact human social behaviour, cognition, and overall well-being through, on one side, their interactive and immersive qualities and, on the other hand, their interactional and social features. The scope of this thesis is thus to bridge the gap between psychological research and the ethics of interactional technologies, offering a reproducible methodology for psychology research to inform and enhance existing ethical frameworks and discussions.

The structure of this thesis is designed to explore the integration of psychological insights into the ethics of interactional technologies. Following the introduction, the Literature Review will provide an overview of existing ethical frameworks, focusing on how psychology can inform and enhance ethical considerations in developing and deploying interactive and interactional technologies. The Academic Contributions section will then present original research demonstrating the practical application of these insights across various technological contexts. The Discussion will then synthesise these findings into a cohesive framework for incorporating psychological perspectives into technology ethics, addressing both the strengths and limitations of this approach and suggesting directions for future research in this interdisciplinary field. Finally, the Conclusion will summarise the key contributions of the thesis.

"In life, nothing is to be feared, everything is to be understood.

Now is the time to understand more so that we may fear less."

Maria Skłodowska-Curie[2]

2 What is the Relevance of Psychology for Interactional Ethics

Integrating psychological insights into ethical frameworks is essential in an era of technology that rapidly reshapes and creates new ways for humans to interact. Psychology offers valuable contributions to technology ethics by helping us understand human cognition, behaviour, and well-being, ensuring that these technologies enhance human flourishing, respect autonomy, and align with human-centred values. To illustrate where psychology can inform existing ethics research and discussions, this chapter reviews key ethical frameworks for AI and interactional technologies, emphasising how psychological perspectives can enrich these frameworks.

2.1 In Frameworks for Technology Ethics

As AI and autonomous systems evolve, various ethical guidelines and frameworks have been developed internationally to ensure that these technologies uphold human values and societal well-being. This section reviews prominent frameworks, identifying where psychological considerations can be better integrated.

[2] in Centenary Lectures (International Atomic Energy Agency, 1968), p. 163

One of the most influential frameworks comes from the Organisation for Economic Cooperation and Development (OECD), which established guiding principles emphasising that AI should benefit humanity as a whole. Central to the OECD's framework (2024) is the focus on human-centred values such as fairness, inclusivity, and the protection of human rights. Additionally, the OECD framework stresses transparency and explainability in AI systems, ensuring that people can understand how decisions are made. However, the framework could benefit from a deeper exploration of psychological impacts, especially concerning how AI systems influence individual cognitive processes and emotional responses, which are critical to understanding fairness and trust.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE SA, 2024) also emphasises human well-being, accountability, and transparency. The initiative's focus on ethical responsibility aligns well with the psychological principles of agency and decision-making. Yet, the framework could further address how AI systems shape autonomy at the cognitive and emotional levels, particularly how algorithmic systems may manipulate or influence users' decision-making processes. By integrating psychological research on human autonomy and manipulation, this framework could offer more concrete strategies for safeguarding user autonomy.

AI4People, a European initiative, offers another ethical framework grounded in five core principles: beneficence, non-maleficence, autonomy, justice, and explicability (Floridi et al., 2018). These principles align well with psychological concepts such as promoting well-being, ensuring fairness, and respecting human autonomy. For example, psychological research on well-being can inform how AI might enhance or undermine users' emotional and social experiences. However, the principle of explicability would benefit from integrating psychological insights into how individuals understand and interpret AI decisions—addressing potential gaps in cognitive comprehension and trust.

Finally, the European Commission's Ethics Guidelines for Trustworthy AI (HLEG, 2019), developed by the High-Level Expert Group on AI, introduces seven key requirements for AI systems: human agency and oversight, technical robustness, transparency, and accountability. Notably, the European Commission introduced the Assessment List for Trustworthy Artificial Intelligence (Publications Office of the European Union, 2020), which helps organisations assess their systems' ethical alignment. While the framework promotes ethical design, it could be enriched by exploring how AI systems impact psychological well-being, especially in interactional technologies like chatbots and social robots, which can profoundly affect users' social and emotional experiences.

These frameworks collectively underscore the importance of ethical AI development. Still, they often lack detailed considerations of psychological impacts, particularly how AI systems affect users' cognitive, emotional, and social well-being. The following sections will explore how key psychological insights could strengthen these frameworks.

2.1.1 Well-Being

Well-being is a multifaceted concept encompassing emotional, social, and psychological dimensions. From a psychological perspective, well-being includes both hedonic (pleasure-oriented) and eudaimonic (purpose-oriented) aspects, which are crucial for understanding how technologies impact human happiness and functioning (Upadhyay & Arya, 2015). For instance, AI-driven companionship tools can enhance well-being by providing personalised support and answering the human need for company. Still, there are concerns about AI exacerbating social isolation or eroding meaningful human relationships by potentially replacing them (Capraro et al., 2024).

In the context of interactional technologies—such as AI-powered virtual assistants or social robots—well-being is shaped by how these systems facilitate or hinder human connection. Psychology can inform the design of these technologies to ensure they foster positive social interactions and mitigate the risk of isolation. Additionally, developmental psychology highlights the importance of secure relationships and positive environments during childhood and adolescence (Gómez-López et al., 2019), suggesting that AI systems targeting these age groups must carefully consider their long-term impact on social and emotional development.

2.1.2 Autonomy and Decision Making

Similarly, the notion of autonomy, particularly in decision-making, is central to discussions on AI ethics. Autonomy is the capacity to make informed, uncoerced decisions that align with an individual's interests and values. It involves self-regulation, where actions are guided by internal values and beliefs rather than external pressures (Wichmann, 2011). Autonomy should not be confused with independence; instead, it refers to aligning one's actions with internal values, even when influenced by external factors. Decision-making, meanwhile, is the cognitive process of choosing a course of action from multiple alternatives. When decision-making is autonomous, it is free from undue external influences and grounded in reflective, rational deliberation (Felsen & Reiner, 2011). This process balances rational thought and personal values, with genuine autonomy achieved when decisions are rational and free from covert external influences.

In the context of AI, autonomy refers to individuals' capacity to make informed and independent decisions - a capacity that AI systems can both support and undermine (Poszler & Lange, 2024). While AI-driven decision-making tools can empower users by providing tailored recommendations, they also raise concerns about the potential

manipulation of choices and the erosion of human agency (Pàez-Gallego et al., 2020). For example, AI-driven systems that recommend content or products based on user behaviour can influence decisions in ways that may conflict with users' long-term goals or well-being. Psychological insights into decision-making processes can help design systems that empower users while safeguarding their ability to make reflective, independent choices. Achieving a balance between leveraging AI for enhanced decision-making and maintaining user autonomy remains a critical focus of ongoing research.

2.1.3 Understandability and Transparency

Regarding the understanding of technology, psychology provides valuable insights into defining appropriate levels of transparency or explicability for diverse user groups. Understanding abilities are examined through the cognitive, emotional, and social aspects that contribute to human performance. For instance, building mental models that align with natural human information processing can enhance comprehension of complex AI systems (Mittelstadt et al., 2018). This involves creating explanations that are accurate and accessible to non-experts, such as using simplified models that approximate AI decision-making processes to help users predict outcomes and understand potential failures. Additionally, explanations must be context-sensitive and personalised to the audience, as users vary in expertise and cognitive resources (Mishra & Rzeszotarski, 2021).

Interactional systems, such as conversational AI or social robots, often require users to trust their decisions and behaviour. However, transparency alone is not enough; psychological research suggests that explanations must be designed to align with users' cognitive capacities. Techniques like interactive explanations, where users can explore how AI systems make decisions, may improve understanding and reduce cognitive overload (Chia, 2023; Hoffman et al., 2022). Cognitive load management is critical, as overwhelming

users with complex information can erode trust and hinder comprehension (Volz et al., 2018).

2.1.4 Value Definition and Cultural Sensitivity

Values are deeply embedded cognitive structures that shape human behaviour and decision-making (Oyserman, 2015). Understanding how different cultures prioritise certain values over others is crucial for designing ethical technologies. For instance, psychological insights into cultural differences can inform value-sensitive design (VSD), ensuring that AI systems align with the cultural values of the populations they serve (Schwartz & Bilsky, 1987).

Moral and cultural psychology provide methods to identify and operationalise values within technological systems to design interactional technologies. Techniques like VSD can help bridge the gap between abstract ethical principles and practical design decisions, ensuring that technologies respect and reflect diverse user values (Friedman et al., 2020). Furthermore, psychology helps predict how users will engage with AI based on their value systems, ensuring ethical alignment and enhancing user satisfaction (Cohen et al., 2021).

While the frameworks presented and discussed here were meant to encompass AI systems, their applicability extends beyond AI, encompassing various technologies, such as video games and immersive worlds, which currently lack formal ethical guidelines. We can thus consider those to apply to general interactional technologies.

While ethical frameworks provide high-level guidance, the practical application of these principles depends on how we design human-technology interactions. Fields like Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI) offer valuable

avenues for embedding psychological principles into the design of interactional technologies.

2.2 In Applied Interactional Technology Ethics

Ethics by Design, as outlined by Brey and Dainow (2023), presents a holistic framework that integrates ethical principles directly into technological systems' design and development processes. Initially conceived in the context of artificial intelligence (AI), this approach has since broadened its scope to encompass a wide range of technologies, including video games and immersive environments. The core idea behind Ethics by Design is that technology can be inherently ethical if ethical considerations guide its development. By embedding ethical principles and practices during the design phase, it becomes possible to create technologies that are functional and aligned with societal values and norms, such as the safeguard of human overall well-being, autonomy, and dignity. This is thus a somewhat impact-based approach aimed at avoiding adverse outcomes and fostering technologies' positive effects. Regarding interactive and interactional systems, two research fields specialise in improving the design and outcomes of innovations, namely, HCI and HRI.

2.2.1 Human-Computer Interaction

Human-Computer Interaction (HCI) lies at the crossroads of computer science, technology, and cognitive psychology, where systems design is informed by how humans process information externally (Jiao, 2022). While HCI focuses on the immediate interaction between users and technology, it also delves into the psychological aspects of user behaviour to improve interfaces and user experiences (Baharum et al., 2023). Psychology, thus, already plays a critical role in HCI by providing a lens to understand users, tasks, and computers, enabling a deeper analysis of user behaviour and facilitating better design (Olson & Olson,

2003; Liu, 2020). However, the current focus of HCI primarily centres on optimising interactions without always considering the long-term societal impact (Munteanu et al., 2015).

To bridge this gap and align with the ethical frameworks calling for a broader societal perspective, psychology can offer valuable insights. By incorporating knowledge from social psychology, HCI can better address ethical challenges such as consent in research and technology design (Munteanu et al., 2015; Sun et al., 2022). Understanding user demands through psychological modelling can lead to more socially responsible interactive systems (Sun et al., 2022). Moreover, the integration of cognitive psychology principles in artificial intelligence (AI) systems within HCI can enhance the development of technologies that not only mimic human cognition but also adapt to user needs effectively (Prasad & Kalavakolanu, 2023; Jain & Alam, 2022). This convergence of psychology and AI is pivotal in creating intelligent systems prioritising user well-being and long-term societal impact (Prasad & Kalavakolanu, 2023; Jain & Alam, 2022).

2.2.2 Human-Robot Interaction

In Human-Robot Interaction (HRI), researchers draw inspiration from diverse fields like psychology, philosophy of mind, and neuroscience to imbue robots with cognitive abilities for collaborative actions. This interdisciplinary approach, rooted in social robotics, delves into replicating human psychological mechanisms essential for shared activities. Concepts like the theory of mind, emotional recognition, and human-aware navigation form the bedrock of these endeavours, aiming to bridge the gap between humans and robots (Belhassein et al., 2022; Giger et al., 2019; Thomaz et al., 2016).

The infusion of psychology into HRI is pivotal, enriching the understanding of human-robot interactions by incorporating cognitive, emotional, and social dimensions

(Stock-Homburg, 2021). For instance, emotional recognition and response mechanisms are central to this synergy, exploring how humans perceive and react to artificial emotions evoked by robots, underscoring the potential significance of emotional intelligence in robots (Stock-Homburg, 2021). Integrating social and cognitive psychology theories fortifies HRI research, offering insights into how humans perceive robots as social entities or even outgroup members, dictating their responses and interactions (Irfan et al., 2018).

However, the current landscape of HRI, in line with HCI, tends to prioritise technological design and short-term outcomes over a holistic understanding of the human aspect, both at an individual and societal level. Psychology, with its study of human ecological workings, can serve as a bridge to address this gap. By integrating psychological perspectives that consider humans in their entirety - both as individuals and within broader societal contexts - HRI and HCI can transcend their current focus on technology-driven solutions to embrace a more comprehensive and enduring approach to human-robot interactions.

In sum, while ethical frameworks provide essential foundations for the responsible development of AI and interactional technologies, their efficacy is limited without a deeper understanding of the psychological impacts on users. By integrating psychological principles into high-level ethical guidelines and practical design approaches in HCI and HRI, we can ensure that these technologies align with societal values and genuinely enhance human well-being and autonomy. This approach will facilitate a deeper understanding of technologies' short-term and long-term impacts on individuals and ensure that the integration of psychology is methodical and targeted.

Having established the importance of integrating psychology into technology ethics, particularly within the context of HCI and HRI, the following section presents a structured

framework that highlights which fields of psychology can contribute to this interdisciplinary effort. This framework synthesises insights from both psychology and ethics, offering a clear roadmap for systematically integrating psychological principles into interactional technologies' ethical evaluation and design.

2.3 A Framework for Psychology Research Integration in Interactional Technology Ethics

Building on the limitations of existing ethical frameworks, we propose a structured approach to integrating psychological insights across different domains of interactional technology design. Figure 1 presents a framework categorising key psychological disciplines and their contributions to technology ethics. This framework is not static but evolves as psychological research advances, ensuring that technology ethics remain responsive to new challenges.

The figure categorises the relevant fields into four groups: Applied Psychology Fields, Core Psychological Disciplines, Developmental and Cultural Psychology Fields, and Technology Ethics. Each group contributes uniquely to the ethical evaluation and design of interactional technologies. It seems important here to highlight that if these fields can be identified individually, they interact and learn from each other. Thus, sometimes, the distinction will not be as evident.

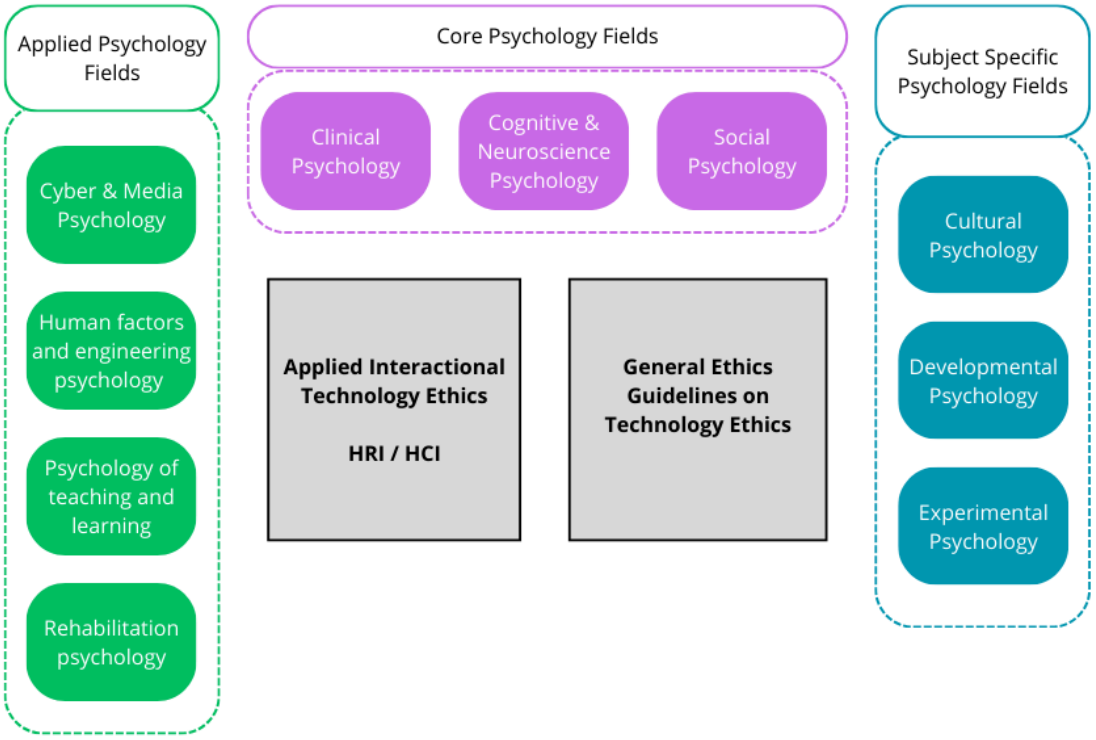

Figure 1. *A Framework for Psychology Research Integration in Interactional Technology Ethics*

2.3.1 Technology Ethics (Grey Group):

This group centres on the ethical implications of applying psychological principles to technology. *Applied Interactional Technology Ethics (HRI/HCI)* focuses on the ethical challenges of designing and using interactional technology in ways that respect human dignity and rights, particularly in human-computer and human-robot interactions. *General Ethics Guidelines on Technology Ethics* provide overarching principles that ensure technology is developed and used in ways aligned with societal values.

2.3.2 Core Psychological Disciplines (Purple Group):

These foundational areas of psychology, including *Clinical Psychology*, *Cognitive & Neuroscience Psychology*, and *Social Psychology*, provide the essential understanding of

human behaviour and mental processes. *Clinical Psychology* focuses on diagnosing and treating mental health issues, ensuring that technologies do not create or exacerbate mental health problems (Kazdin, 2017). It provides insights into how technology can be used ethically to support mental health and well-being, such as through therapeutic applications or by avoiding features that could contribute to anxiety, stress, or other problems. *Cognitive & Neuroscience Psychology* informs the design of technologies that align with human cognitive processes, such as memory and problem-solving (Gazzaniga et al., 2018). It contributes to interactional technology ethics by ensuring that technologies are designed to complement human cognition, avoid cognitive overload, and support social and metacognitive skill development. *Social Psychology* contributes to understanding the social implications of technology, including how it influences group dynamics, social behaviour, and interpersonal relationships (Myers & Twenge, 2019). It plays a critical role in interactional technology ethics by addressing issues such as social isolation, anti-social behaviours, and the ethical implications of AI in social contexts, ensuring that technology enhances rather than undermines social well-being.

2.3.3 Subject Specific Psychology Fields (Blue Group):

This group explores the influence of culture and human development on behaviour, integrating these perspectives into the ethical considerations of technology. *Cultural Psychology* examines how cultural contexts are shaped by human experiences, ensuring that technologies are culturally sensitive (Heine, 2020). Interactional technology ethics ensure that technologies are culturally sensitive and do not impose values misaligned with users' cultural backgrounds. It contributes to creating ethical guidelines that respect cultural diversity and promote inclusivity in technology design. *Developmental Psychology* provides insights into how psychological development across the lifespan affects interactions with

technology, guiding the design of age-appropriate technologies (Daum & Manfredi, 2022). *Experimental Psychology* uses scientific methods to study fundamental aspects of human behaviour, such as sensation and perception, which are critical for evaluating the impact of new technologies (Kantowitz et al., 2014). In interactional technology ethics, it provides the empirical evidence needed to evaluate the impact of new technologies, ensuring that ethical guidelines are based on robust scientific data.

2.3.4 Applied Psychology Fields (Green Group):

This group includes fields focused on applying psychological principles to address real-world problems. *Cyber & Media Psychology* examines the psychological effects of digital media and online behaviour, providing insights crucial for ethically managing digital technologies, communication and online interactions (Riva et al., 2016). For example, it helps evaluate the potential for digital technologies to cause harm or foster well-being, guiding the design of ethical digital environments that respect users' psychological needs. *Human Factors and Engineering Psychology.* This field ensures that interactional technologies, particularly in HCI and HRI, are designed with the user's capabilities and limitations in mind (Boehm-Davis et al., 2015). It contributes to interactional technology ethics by emphasising the importance of safety, usability, and user satisfaction. This field informs the ethical design of intuitive and safe systems, minimising the risk of harm and maximising the effectiveness of the interaction between humans and technology. The *Psychology of Teaching and Learning* informs the development of educational technologies by exploring cognitive processes involved in learning and effective instructional strategies (Bransford et al., 2000). In the context of interactional technology ethics, it ensures that educational technologies are designed to enhance learning outcomes without introducing biases or inequities. It also guides the ethical use of technology in educational settings,

ensuring that it supports rather than hinders learning processes. *Rehabilitation Psychology* contributes to interactional technology ethics by ensuring that technologies are designed to support individuals with disabilities or chronic health conditions (Dunn, 2019). It emphasises the need for ethical design principles that respect the autonomy and dignity of users, helping to create technologies that are empowering and accessible to all individuals, regardless of their physical or mental capabilities.

While the framework categorises psychology fields, it is important to note the dynamic interaction between these disciplines. Insights from developmental psychology may inform HCI design principles, while cultural psychology can guide the ethical evaluation of AI systems deployed in diverse cultural contexts. This interdisciplinary collaboration allows for a more nuanced understanding of how technologies affect human behaviour and well-being, ensuring that ethical frameworks are comprehensive and adaptable to emerging technological challenges.

Building on this integrative framework, the next section will outline the academic contributions of this thesis, demonstrating how psychological principles can be systematically integrated into interactional technology ethics. The following Discussion will introduce the methodologies employed to incorporate these contributions into existing ethical discourse, providing concrete steps for advancing the ethical development of interactional systems.

“I was taught that the way of progress was neither swift nor easy.”

Maria Skłodowska-Curie in *Pierre Curie: With Autobiographical Notes by Marie Curie* (1963)

3 Academic Contributions

The doctoral thesis is based on the following four submissions to international journals and book chapters authored by Auxane Boch and colleagues. More information on the individual contribution of Auxane Boch and her co-authors can be retrieved in the appendix of this dissertation.

The following academic contributions form the backbone of this thesis, offering diverse case studies that explore the integration of psychology in resolving ethical dilemmas across different domains, from healthcare and social robotics to video game ethics. Each contribution highlights a specific intersection between psychological insights and ethical frameworks, providing practical approaches to ensuring ethically sound and psychologically beneficial technologies. These contributions discuss how:

- Psychology can inform the resolution of dilematic situations in a given ethical frame for a specific use case.
- Psychology can inform the design of specific technologies from an ethics-by-design perspective, building on existing ethical frameworks.
- A psychological evaluation framework can arise from different psychological perspectives and impacts.

- Psychology knowledge can be used to enhance critical thinking about relationships to technology for the broader population.

Essay 1. Beyond the Metal Flesh: Understanding the Intersection between Bio- and AI Ethics for Robotics in Healthcare

Reference: Boch, A., Ryan, S., Kriebitz, A., Amugongo, L. M., & Lütge, C. (2023). Beyond the Metal Flesh: Understanding the Intersection between Bio- and AI Ethics for Robotics in Healthcare. *Robotics, 12*(4), 110. https://doi.org/10.3390/robotics12040110

This paper advances the understanding of how psychology can contribute to resolving dilemmatic situations within a specific ethical framework, particularly in the context of care robots (CRs) in healthcare. Specifically, it highlights how psychological insights into human behaviour, decision-making, and emotional responses can help resolve conflicts between competing ethical principles, such as autonomy versus dependency or patient centricity versus profit centricity. For example, the discussion on the potential emotional dependency of patients on CRs and the risks associated with this draws on psychological understanding of attachment and autonomy. This psychological insight is crucial for resolving the dilemma of whether and how to deploy CRs in ways that enhance patient well-being without compromising their autonomy. By integrating these psychological considerations into the ethical framework, the paper provides a structured approach to navigating and resolving ethical dilemmas in real-world applications of CRs.

The paper also contributes to the idea that psychology can inform the design of technologies from an ethics-by-design perspective. It emphasises the need to incorporate psychological principles into the design process of CRs to ensure that these technologies align with ethical standards while also being effective and user-friendly. For instance, the

paper discusses how understanding psychological concepts like autonomy, emotional attachment, and trust can guide the development of CRs that respect patients' decision-making capabilities and promote their well-being. This includes designing CRs that are transparent, easy to understand, and adaptable to the psychological needs of different users. The paper's recommendations for creating CRs that can effectively interact with patients, provide support without fostering unhealthy dependencies, and maintain patient autonomy are grounded in psychological insights. These recommendations ensure that the technology is functional and ethically aligned with the principles of beneficence, non-maleficence, and justice.

In addition to these contributions, the paper also adds to the broader thesis by proposing practical guidelines for the ethical development and deployment of CRs. It suggests methodologies for involving stakeholders in the design process, ensuring that the voices of patients, caregivers, and healthcare professionals are heard and integrated into the technology's development. The paper's focus on iterative and context-specific ethical evaluation further enhances the thesis by advocating for continuous adaptation and improvement of ethical frameworks as CR technology evolves and becomes more widely adopted.

This paper contributes to the thesis by demonstrating how psychology can inform ethical decision-making and the design of AI-driven technologies, specifically in the sensitive and complex healthcare field. It provides a solid foundation for developing CRs that are not only technically proficient and ethically sound, ensuring that the benefits of this technology are maximised while minimising potential harm.

Essay 2. Human-Robot Dynamics: A Psychological Insight into the Ethics of Social Robotics

Boch, A., & Thomas, B. R. (2025). Human-robot dynamics: a psychological insight into the ethics of social robotics. *International Journal of Ethics and Systems, 41*(1), 101-141. DOI 10.1108/IJOES-01-2024-0034.

This paper comprehensively explores how psychological principles can significantly contribute to social robots' ethical design and deployment.

First, the paper delves into critical ethical dilemmas such as deception, dehumanisation, and violence in human-robot interactions. It leverages psychological theories and empirical findings to provide nuanced insights into how these dilemmas can be addressed. For instance, the discussion on anthropomorphism and its effects on human empathy toward robots directly applies psychological understanding to mitigate ethical concerns related to deception and emotional attachment. By framing robots in ways that align with human psychological tendencies, the paper suggests strategies to avoid ethical pitfalls, such as excessive emotional attachment or replacing human relationships, thereby resolving these dilemmas within an ethical framework.

Second, the paper contributes to the ethics-by-design approach by recommending specific design elements for social robots that consider psychological factors. For example, it emphasises the importance of engagement factors, such as the robot’s ability to display emotions and provide social support, which are crucial for fostering positive human-robot interactions. These recommendations are rooted in psychological theories about human behaviour, such as the need for social connections and the impact of warmth and competence in relationships. Additionally, the paper emphasises the importance of cultural

sensitivity, user engagement, and personalisation. It advocates for designing robots that respect cultural differences, manage user expectations, and offer personalised interactions to enhance user satisfaction and trust. By integrating these psychological insights, the paper offers a blueprint for designing robots that meet functional requirements and adhere to ethical principles.

Overall, the paper enriches the thesis by demonstrating how psychological theories and empirical research can be directly applied to address ethical concerns in social robotics. It underscores the necessity of integrating psychological insights into the design process to ensure that social robots are developed ethically, respecting human values and fostering positive interactions. This comprehensive approach enhances the design of social robots and contributes to the broader discourse on ethical AI, ensuring that technological advancements align with human well-being and societal norms.

Essay 3. Introduction to Video Games Ethics

Boch, A., Thomas, B., Wanick, V., & Clifford, C. (2024). Introduction to video games ethics. In *The Elgar Companion to Applied AI Ethics* (pp. 313-343). Edward Elgar Publishing. https://doi.org/10.4337/9781803928241.00021

This book chapter contributes to the broader discussion of how psychology can inform the resolution of dilemmas within ethical frameworks, particularly in the context of video games. It highlights the intersection of psychology and ethics by examining video games' complex and multifaceted impacts on mental health and well-being, cognitive skills, social behaviours, and the broader gaming culture. It gives the blueprint for an initial methodology for psychology to be included in ethical discussions. The chapter illustrates how psychological insights can be leveraged to understand the effects of video game mechanics

on cognitive and social skills, such as problem-solving, attention, and empathy, which can be translated into other areas of life, including education and social interactions From an ethics-by-design perspective, the chapter underscores the importance of integrating psychological knowledge into the design of video games to mitigate potential negative impacts, such as addiction, manipulation through dark patterns, and the reinforcement of harmful stereotypes. Developers can create games that entertain and foster positive behaviours and social norms by understanding players' cognitive and emotional responses. The chapter also emphasises the need for a psychological evaluation framework that considers the long-term and cumulative effects of video games on well-being, which can help inform ethical guidelines and best practices in game design.

Moreover, the chapter highlights the role of psychology in enhancing critical thinking about individuals' relationships with technology, particularly in gaming culture. It discusses how the representation of gender, culture, and identity in video games can influence players' attitudes and behaviours and how psychological research can inform efforts to create more inclusive and positive gaming environments.

In summary, this chapter contributes to the thesis by showing how psychology can inform both the resolution of ethical dilemmas and the design of technology, specifically video games, through an ethics-by-design approach. It also adds to the discussion by proposing that a psychological evaluation framework is essential for understanding the broader impacts of video games and by advocating for the use of psychological insights to enhance critical thinking about technology's role in society.

Essay 4. Playing with Morality: Investigating the Potential of Narrative Games on Human-AI Interactions

Boch, A., Jackson, B., McDonnell, D., Atherton, G., Belyk, M., Cross, L. (*Submitted, Under Review).*

The last paper aims to bridge the gap between how psychology can help inform and design more ethical interactive and interactional technology, and using technology to foster ethical conversations in the broader population.

These two study experiments explored how narrative video games can influence human perceptions and attitudes toward AI, particularly in the context of moral decision-making and trust. The methodology involved two experimental studies where participants played selected chapters from the game "Detroit: Become Human," a narrative-driven game that places players in the roles of AI characters facing ethical dilemmas. The first study measured participants' attitudes toward AI before and after gameplay using surveys focused on trust, AI roles in society, and the potential for AI consciousness. The second study extended this investigation by incorporating a virtual reality interaction with an AI agent, allowing researchers to observe explicit attitudes and implicit behaviours. The main outcome revealed that engaging with AI characters in morally charged scenarios led to stronger emotional connections with AI, more positive attitudes towards AI as a potential companion, and a greater willingness to attribute moral responsibility to AI agents. This suggests that narrative games can be a powerful tool for shaping public perceptions of AI and fostering more nuanced, informed discussions about AI's role in society.

Therefore, this last paper demonstrates how psychological insights - such as how people respond to moral dilemmas in immersive settings - can be leveraged to foster ethical discussions about human relationships with AI. By examining how narrative games can change perceptions and attitudes towards AI, the paper highlights the role of psychological knowledge in enhancing critical thinking about technology. The findings suggest that immersive gaming experiences can encourage individuals to reflect on their relationship with AI, potentially leading to more informed and critical perspectives on the role of AI in society. This aligns with the broader goal of using psychology to foster a more thoughtful and responsible approach to technology adoption among the general public.

“All sorts of things can happen when you’re open to new ideas and playing around with things."

Oral history interview with Stephanie L. Kwolek (1998)

4 Discussion: A Roadmap for Integrating Psychological Insights into Interactional Technology Ethics

Returning to the initial question of how psychological insights can be systematically integrated into interactional technology ethics, this section provides a roadmap for addressing the key challenges identified in the thesis. Drawing on deductive and inductive approaches, the proposed framework synthesises the psychological contributions from academic studies and lays out a path for future ethical evaluation.

4.1 Deductive Approach: Applying Ethical Frameworks

In our paper "Beyond the Metal Flesh: Understanding the Intersection between Bio- and AI Ethics for Robotics in Healthcare" (Boch et al., 2023), the deductive approach is evident as we systematically apply established ethical principles to the context of care robots (CRs) in healthcare. Here, Clinical Psychology played a critical role in understanding how CRs could potentially impact mental health, addressing concerns about emotional dependency and autonomy. The study drew on core ethical principles such as beneficence and

non-maleficence, deductively applying them to evaluate the ethical implications of CRs on patient well-being. Cognitive Psychology also contributed by assessing how patients' cognitive processes, such as decision-making capabilities, could be supported or undermined by CRs. This paper demonstrates how psychological insights can be deductively applied to evaluate and resolve ethical dilemmas in technology design, ensuring that CRs are both ethically sound and beneficial to users.

4.2 Inductive Approach: Deriving Ethical Guidelines from Psychological Impacts

In contrast, the inductive approach is prominently featured in our paper "Human-Robot Dynamics: A Psychological Insight into the Ethics of Social Robotics" (Boch & Thomas, 2025). This paper investigates the real-world psychological impacts of social robots, using empirical findings to inform and derive ethical guidelines. Social and Cognitive Psychology were crucial here, as they provided insights into how anthropomorphism and human-robot interactions could affect human empathy and social behaviours. By observing the psychological effects of social robots on individuals, the study inductively developed ethical considerations regarding issues like deception and emotional attachment. Similarly, Cultural Psychology informed the study by exploring how different cultural contexts could influence perceptions of and interactions with social robots, leading to culturally sensitive ethical guidelines. This paper highlights how empirical data from psychological studies can be used inductively to build and refine ethical frameworks for emerging technologies.

4.3 Bridging Deduction and Induction: A Comprehensive Ethical Framework

Our chapter "Introduction to Video Games Ethics" (Boch et al., 2024) bridges both deductive and inductive methods by first establishing a broad ethical framework for video games based on Cognitive & Neuroscience Psychology, Clinical Psychology and Social Psychology. This deductive application ensured that video game design aligned with ethical principles related to cognition, behaviour, and social interaction. Subsequently, the chapter incorporated inductive reasoning by examining the psychological impacts of video game mechanics on players' cognitive skills, social behaviours, mental health and wellbeing. The findings from these observations were then used to refine and expand the ethical framework, addressing specific issues such as addiction, manipulation through dark patterns, and the reinforcement of harmful stereotypes. The final inductive insights inform the ethical design of video games, ensuring they foster positive behaviours and social norms.

4.4 Integrative Approach: Enhancing Critical Thinking and Public Discourse

Finally, in our study "Playing with Morality: Investigating the Potential of Narrative Games on Human-AI Interactions" (Boch et al., *under review*), we employed both deduction and induction to explore how narrative-driven games can shape public perceptions and ethical discussions around AI. The study started with a deductive framework, utilising Cyber and Media Psychology to ensure that the narrative games were appropriate. As the study progressed, empirical findings leading to inductive insights were gathered by observing how players' attitudes toward AI evolve through their interactions with the game. Experimental

Psychology played a key role here, as it provides the methodological tools to measure these changes in perception and behaviour, leading to new ethical considerations about the interplay of games and AI in society. This integrative approach shows how psychological research can both inform and be informed by ethical discussions, enhancing critical thinking and public discourse on the ethical implications of AI technologies.

4.5 Limitations

While this thesis has demonstrated the critical role of psychology in shaping technology ethics, it is important to acknowledge its limitations. The interdisciplinary nature of the work presents inherent challenges, particularly in navigating the differences in methodologies, theoretical frameworks, and terminologies across fields like psychology, ethics, and technology design. These complexities may sometimes lead to oversimplification when integrating concepts from different disciplines into a cohesive framework.

Moreover, the ethical frameworks discussed in this thesis are largely rooted in Western perspectives, particularly European contexts, which may limit the applicability of the findings to non-Western or more diverse global contexts. Local cultural values, social norms, and political environments often influence ethical considerations surrounding technology development and deployment. Therefore, future research must broaden its scope to include diverse cultural perspectives to ensure that the ethical frameworks and psychological insights are adaptable to different regions and populations.

Another limitation lies in the focus on specific types of interactional technologies - namely, healthcare robots, social robots, and video games. While these examples provide valuable case studies, they represent only a fraction of the rapidly expanding landscape of AI-driven technologies. As AI continues to evolve, so too must the frameworks and methodologies used to evaluate its ethical and psychological impacts.

Furthermore, while the thesis highlights the importance of integrating psychological insights into the design and ethical evaluation of technology, it does not fully address the need for standardised tools and metrics to measure the psychological impacts of these technologies, but calls for it. Developing reliable and scalable psychometric tools that can assess the short-, medium-, and long-term effects of interactional technologies on users will be essential for advancing the field. These tools would enable a more objective and quantifiable assessment of how technologies affect mental health, well-being, cognitive function, and social behaviour, ensuring that ethical frameworks are grounded in empirical evidence.

4.6 Outlook

The field of interactional technology ethics can greatly benefit from expanding its interdisciplinary collaborations. Future research should involve ethicists, psychologists, and technologists and draw insights from fields like anthropology, sociology, and political science to build a more holistic understanding of technology's societal impact. In particular, involving psychologists in the early stages of technology development, alongside designers and engineers, will ensure that ethical considerations are embedded from the outset.

Global cooperation is also essential. Given the diverse ways in which different societies interact with technology, fostering cross-cultural dialogue on technology ethics will help create more inclusive, adaptable frameworks. Such dialogues will ensure that technological advancements respect not only individual well-being but also broader cultural and societal values.

In conclusion, while this thesis has laid a strong foundation for integrating psychology into interactional technology ethics, it also highlights the need for further refinement and expansion. By addressing these limitations and pursuing more inclusive,

interdisciplinary, and culturally sensitive research, we can ensure that future technological innovations are both ethically sound and aligned with the psychological well-being of users across the globe.

"I am one of those who think like Nobel, that humanity will draw more good than evil from new discoveries."

Maria Skłodowska-Curie[3]

5 Conclusion

This thesis has sought to contribute to our understanding of the psychological impacts of technology and how these insights can inform ethical frameworks for developing interactional systems. As technologies like AI, social robots, and immersive realities continue to evolve, their effects on human cognition, behaviour, and well-being become increasingly complex and, in many cases, uncertain. By integrating psychological insights into these technologies' ethical design and evaluation, we can ensure, to the best of our abilities, that they are developed in ways that enhance, rather than undermine, human flourishing.

This work has demonstrated psychology's critical role in shaping technology ethics, particularly in addressing moral dilemmas, fostering ethical design, and promoting critical thinking about our relationship with technology. Through interdisciplinary collaboration and the application of empirical psychological research, we can build more ethically sound technologies that align with societal values and promote the well-being of diverse populations.

While integrating psychology into technology ethics is ongoing, this thesis has laid a foundation for future research and practice. By continuing to refine these methodologies,

[3] as quoted in *White Coat Tales : Medicine's Heroes, Heritage and Misadventures* (2007) by Robert B. Taylor (p.141)

expanding their cultural scope, and encouraging interdisciplinary dialogue, we can navigate the complex interplay between humans and technology, ultimately creating innovations that align with human ethical values and psychological needs.

References

Baharum, A., Rahim, E., Ismail, R., Noor, N., Daruis, D., Deris, F., & Samat, J. (2023). Unveiling the psychology of human-computer interaction: Bridging the gap for future advancements. *2023 International Conference on Platform Technology and Service (PlatCon)*, 36-38. https://doi.org/10.1109/PlatCon60102.2023.10255194.

Belhassein, K., Fernández-Castro, V., Mayima, A., Clodic, A., Pacherie, E., Guidetti, M., Alami, R., & Cochet, H. (2022). Addressing joint action challenges in HRI: Insights from psychology and philosophy. *Acta Psychologica*, 222, 103476. https://doi.org/10.1016/j.actpsy.2021.103476.

Boch, A., Ryan, S., Kriebitz, A., Amugongo, L. M., & Lütge, C. (2023). Beyond the metal flesh: Understanding the intersection between bio- and AI ethics for robotics in healthcare. *Robotics, 12*(4), 110. https://doi.org/10.3390/robotics12040110.

Boch, A., Thomas, B., Wanick, V., & Clifford, C. (2024). Introduction to video games ethics. In *Edward Elgar Publishing eBooks* (pp. 313–343). https://doi.org/10.4337/9781803928241.00021

Boch, A., & Thomas, B. R. (2025). Human-robot dynamics: a psychological insight into the ethics of social robotics. *International Journal of Ethics and Systems, 41*(1), 101-141.

Boehm-Davis, D. A., Durso, F. T., & Lee, J. D. (2015). *APA handbook of human systems integration* (pp. xxix-625). American Psychological Association.

Bransford, J., Brophy, S., & Williams, S. (2000). When computer technologies meet the learning sciences: Issues and opportunities. *Journal of Applied Developmental Psychology, 21*(1), 59-84. https://doi.org/10.1016/S0193-3973(99)00051-2

Brey, P., & Dainow, B. (2023). Ethics by design for artificial intelligence. *AI and Ethics*, 1-13. https://doi.org/10.1007/s43681-023-00253-7

Brühlmann, F. (2021). *Understanding and improving subjective measures in human-computer interaction* (Doctoral dissertation, University of Basel). https://doi.org/10.5451/UNIBAS-007078326

Campbel, C. (2018, March 10). A brief history of blaming video games for mass murder. *Polygon*. https://www.polygon.com/2018/3/10/17101232/a-brief-history-of-video-game-violence-blame

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J., Brañas-Garza, P., Butera, L., Douglas, K. M., Everett, J. a. C., Gigerenzer, G., Greenhow, C., Hashimoto, D. A., Holt-Lunstad, J., Jetten, J., Johnson, S., Kunz, W. H., Longoni, C., . . . Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, 3(6). https://doi.org/10.1093/pnasnexus/pgae191

Casas-Roma, J. (2022). Ethical idealism, technology and practice: A manifesto. *Philosophy & Technology, 35*(3). https://doi.org/10.1007/s13347-022-00575-7

Chandler, D. (1995). *The act of writing: A media theory approach*. University of Wales.

Chia, H. L. B. (2023). The emergence and need for explainable AI. *Advances in Engineering Innovation, 3*(1), 1–4. https://doi.org/10.54254/2977-3903/3/2023023

Cohen, S. (1973). *Moral panic and folk devils*. Paladin Press.

Copp, D. (Ed.). (2005). *The Oxford handbook of ethical theory*. Oxford University Press.

Coronado, E., Itadera, S., & Ramirez-Alpizar, I. G. (2023). Integrating virtual, mixed, and augmented reality to Human–Robot Interaction Applications using game Engines: A Brief review of accessible software tools and frameworks. *Applied Sciences, 13*(3), 1292. https://doi.org/10.3390/app13031292

Danaher, J. (2019). The philosophical case for robot friendship. *Journal of Posthuman Studies, 3*(1), 5-24.

Daum, M. M., & Manfredi, M. (2022). Developmental Psychology. *In Springer international handbooks of education* (pp. 239–272). https://doi.org/10.1007/978-3-030-28745-0_13

Dunn, D. S. (Ed.). (2019). *Understanding the experience of disability: Perspectives from social and rehabilitation psychology.* Oxford University Press.

Eggleton, J. (2019). Comment on 'I, Sex Robot: The health implications of the sex robot industry.' *BMJ Sexual & Reproductive Health, 45*(1), 78–79. https://doi.org/10.1136/bmjsrh-2018-200251

Felsen, G., & Reiner, P. (2011). How the Neuroscience of Decision Making Informs Our Conception of Autonomy. *AJOB Neuroscience*, 2, 14 - 3. https://doi.org/10.1080/21507740.2011.580489.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*, 689-707. https://doi.org/10.1007/s11023-018-9482-5.

Friedman, B., Kahn, P., Borning, A., & Huldtgren, A. (2020). Value Sensitive Design and Information Systems. The Ethics of Information Technologies. https://doi.org/10.1007/978-94-007-7844-3_4.

Gazzaniga, M., Ivry, R., & Mangun, G. (2002). *Cognitive science*. New York: WW Norton.

Giger, J., Piçarra, N., Alves-Oliveira, P., Oliveira, R., & Arriaga, P. (2019). Humanization of robots: Is it really such a good idea? *Human Behavior and Emerging Technologies, 1*(2), 111-123. https://doi.org/10.1002/hbe2.147.

Gómez-López, M., Viejo, C., & Ortega-Ruiz, R. (2019). Psychological Well-Being During Adolescence: Stability and Association With Romantic Relationships. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.01772.

Heine, S. J. (2020). *Cultural psychology: Fourth international student edition*. WW Norton & company.

High Level Expert Group on Artificial Intelligence (HLEG). (2019, April 8). *Ethics guidelines for trustworthy AI*. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hoffman, R., Miller, T., & Clancey, W. (2022). Psychology and AI at a Crossroads: How Might Complex Systems Explain Themselves?. *The American Journal of Psychology*. https://doi.org/10.5406/19398298.135.4.01.

IEEE SA. (2024, August 21). *The IEEE Global Initiative 2.0 on Ethics of Autonomous and Intelligent Systems - IEEE Standards Association.* IEEE Standards Association. https://standards.ieee.org/industry-connections/activities/ieee-global-initiative/

Irfan, B., Kennedy, J., Lemaignan, S., Papadopoulos, F., Senft, E., & Belpaeme, T. (2018). Social psychology and human-robot interaction: An uneasy marriage. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/3173386.3173389.

Jain, S., & Alam, M. (2022). Applications of human-computer interaction in health psychology. *Journal of Digital Art & Humanities*. https://doi.org/10.33847/2712-8148.3.1_5.

Jiao, M. (2022). The use of cognitive psychology-based human-computer interaction tax system in ceramic industry tax collection and management and economic development of Jingdezhen city. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.944924.

Kantowitz, B. H., Roediger III, H. L., & Elmes, D. G. (2014). *Experimental psychology: Nelson Education.*

Kazdin, A. E. (Ed.). (2017). *Encyclopedia of psychology* (Vol. 2). Oxford University Press.

Kneer, J., & Ward, M. R. (2020). With a rebel yell: Video gamers' responses to mass shooting moral panics. *New Media & Society, 23*(3), 497–514. https://doi.org/10.1177/1461444819901138

Leiser, M. (2024). Psychological patterns and Article 5 of the AI Act: AI-Powered Deceptive Design in the system architecture and the user interface [Article]. *Journal of AI Law and Regulation*, 5–23. https://doi.org/10.21552/aire/2024/1/4

Liu, Y. (2020). Human-computer interface design based on design psychology. In *2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)* (pp. 5-9). IEEE. https://doi.org/10.1109/ICHCI51889.2020.00009

Mishra, S., & Rzeszotarski, J. (2021). Crowdsourcing and Evaluating Concept-driven Explanations of Machine Learning Models. *Proceedings of the ACM on Human-Computer Interaction, 5*, 1 - 26. https://doi.org/10.1145/3449213.

Mittelstadt, B., Russell, C., & Wachter, S. (2018). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency.* https://doi.org/10.1145/3287560.3287574.

Munteanu, C., Molyneaux, H., Moncur, W., Romero, M., O'Donnell, S., & Vines, J. (2015). Situational ethics. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 105-114). ACM. https://doi.org/10.1145/2702123.2702481

Myers, D. G., & Twenge, J. M. (2019). *Social psychology* (13th ed.). McGraw-Hill Education.

Novicevic, M., Hayek, M., Buckley, M., & Humphreys, J. (2009). Chandler and Technological Determinism in the Histories of Management. *The journal of applied management and entrepreneurship, 14*, 3.

Olson, G. M., & Olson, J. S. (2003). Human-computer interaction: Psychological aspects of the human use of computing. *Annual review of psychology, 54*(1), 491-516.

Organisation for Economic Cooperation and Development (OECD). (2024). *OECD legal instruments*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Oyserman, D. (2015). Values, Psychology of. *In Elsevier eBooks* (pp. 36–40). https://doi.org/10.1016/b978-0-08-097086-8.24030-0

Páez-Gallego, J., Gallardo-López, J. A., López-Noguero, F., & Rodrigo-Moriche, M. P. (2020). Analysis of the relationship between psychological well-being and decision making in adolescent students. *Frontiers in psychology, 11*, 1195.

Publications Office of the European Union. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office of the EU. https://op.europa.eu/en/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1

Poszler, F., & Lange, B. (2024). The impact of intelligent decision-support systems on humans' ethical decision-making: A systematic literature review and an integrated framework. *Technological Forecasting and Social Change, 204*, 123403. https://doi.org/10.1016/j.techfore.2024.123403

Prasad, K., & Kalavakolanu, S. (2023). The study of cognitive psychology in conjunction with artificial intelligence. *Conhecimento & Diversidade, 15*(36). https://doi.org/10.18316/rcd.v15i36.10788

Rajesh, R., Chinthamu, N., Rani, S., B, M., & Sivaiah, B. (2023). Development of Powered Chatbots for Natural Language Interaction in Metaverse using Deep Learning with

Optimization Techniques. *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS),* 534-539. https://doi.org/10.1109/ICAISS58487.2023.10250650.

Richardson, K. (2016). The asymmetrical "relationship." *ACM SIGCAS Computers and Society, 45*(3), 290–293. https://doi.org/10.1145/2874239.2874281

Riva, G., Wiederhold, B. K., & Cipresso, P. (2016). Psychology of social media: From technology to identity. *The psychology of social networking: Personal experience in online communities, 1*, 4-14.

Schwartz, S. H., & Bilsky, W. (1987). Toward a universal psychological structure of human values. *Journal of personality and social psychology, 53*(3), 550.

Stock-Homburg, R. (2021). Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research. *International Journal of Social Robotics, 14*, 389-411. https://doi.org/10.1007/s12369-021-00778-6.

Sun, T., Luo, L., & Chen, G. (2022). The role of social psychology in human-computer interaction in teaching systems. In *2022 3rd International Conference on Information Science and Education (ICISE-IE)* (pp. 11-14). IEEE. https://doi.org/10.1109/icise-ie58127.2022.00009

Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence ("EU AI Act"). (2024). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

Thomaz, A., Hoffman, G., & Cakmak, M. (2016). Computational human-robot interaction. *Foundations and Trends® in Robotics, 4*(2-3), 105-223. https://doi.org/10.1561/2300000049

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000380455

Upadhyay, U., & Arya, S. (2015). A Critique of Research Studies on Application of Positive Psychology for Augmenting Children's Emotional Wellbeing. *Indian journal of positive psychology, 6*, 417.

Volz, V., Majchrzak, K., & Preuss, M. (2018). A Social Science-based Approach to Explanations for (Game) AI. *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, 1-2. https://doi.org/10.1109/CIG.2018.8490361.

White, G. (2023). Technology use and mental health disorders: Dueling aspects of technology as a problem and a solution for mental health. *Journal of Mental Health Disorders*. https://doi.org/10.33696/mentalhealth.3.014

World Health Organization. (2024). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models.* https://iris.who.int/bitstream/handle/10665/375579/9789240084759-eng.pdf?sequence=1

Wichmann, S. (2011). Self-Determination Theory: The Importance of Autonomy to Well-Being across Cultures.. *The Journal of Humanistic Counseling, 50*, 16-26. https://doi.org/10.1002/J.2161-1939.2011.TB00103.X.

Appendix

A1 Essay 1. Beyond the Metal Flesh: Understanding the Intersection between Bio- and AI Ethics for Robotics in Healthcare

Reference: Boch, A., Ryan, S., Kriebitz, A., Amugongo, L. M., & Lütge, C. (2023). Beyond the Metal Flesh: Understanding the Intersection between Bio- and AI Ethics for Robotics in Healthcare. *Robotics, 12*(4), 110. https://doi.org/10.3390/robotics12040110

Information about the Article

Title	Beyond the Metal Flesh: Understanding the Intersection between Bio- and AI Ethics for Robotics in Healthcare.
Authors	Boch, A., Ryan, S., Kriebitz, A., Amugongo, L. M., & Lütge, C.
Accepted	29.07.2023
Journal	Robotics
Volume	12
Number	4
Year	2023
Pages	110

Copyright information

robotics

MDPI

Article

Beyond the Metal Flesh: Understanding the Intersection between Bio- and AI Ethics for Robotics in Healthcare

Auxane Boch [1,*], Seamus Ryan [2], Alexander Kriebitz [3], Lameck Mbangula Amugongo [1] and Christoph Lütge [1,3]

[1] Institute for Ethics in AI, Technical University of Munich, 80333 Munich, Germany; lameckmbangula.amugongo@tum.de (L.M.A.); luetge@tum.de (C.L.)
[2] School of Computer Science and Statistics, Trinity College Dublin, D02PN40 Dublin, Ireland; ryans58@tcd.ie
[3] Peter Löscher Chair of Business Ethics, Technical University of Munich, 80333 Munich, Germany; a.kriebitz@tum.de
* Correspondence: auxane.boch@tum.de

Abstract: As we look towards the future of healthcare, integrating Care Robots (CRs) into health systems is a practical approach to address challenges such as an ageing population and caregiver shortages. However, ethical discussions about the impact of CRs on patients, caregivers, healthcare systems, and society are crucial. This normative research seeks to define an integrative and comprehensive ethical framework for CRs, encompassing a wide range of AI-related issues in healthcare. To build the framework, we combine principles of beneficence, non-maleficence, autonomy, justice, and explainability by integrating the AI4People framework for a Good AI Society and the traditional bioethics perspective. Using the integrated framework, we conduct an ethical assessment of CRs. Next, we identify three key ethical trade-offs and propose remediation strategies for the technology. Finally, we offer design recommendations for responsible development and usage of CRs. In conclusion, our research highlights the critical need for sector-specific ethical discussions in healthcare to fully grasp the potential implications of integrating AI technology.

Keywords: care robots; bioethics; AI ethics; healthcare

Citation: Boch, A.; Ryan, S.; Kriebitz, A.; Amugongo, L.M.; Lütge, C. Beyond the Metal Flesh: Understanding the Intersection between Bio- and AI Ethics for Robotics in Healthcare. *Robotics* **2023**, *12*, 110. https://doi.org/10.3390/robotics12040110

Academic Editor: Sylwia Łukasik

Received: 3 June 2023
Revised: 27 July 2023
Accepted: 29 July 2023
Published: 1 August 2023

1. Introduction

Social robots (SRs) are defined by Fox and Gambino [1] as human-made artificial intelligence (AI) technologies, presented in a digital or physical form, with some degree of human or animal-like attributes. According to a recent review study, there are five main areas where SR technology could potentially be adopted: companionship, healthcare, education, social definition, and social impact [2]. The authors detail that the expected qualities of SRs lie in their abilities to make decisions, have conversations, and react to social cues. Interestingly, when it comes to decision-making, SRs do not seem expected to be moral, but to make the most efficient decision regardless of social implications [3]. Currently, the research surrounding SRs as reported by Lambert et al. [2], focuses mainly on personalisation and social awareness of the tool with the aim to create adaptable social agents with abilities to recognise social cues and mimic emotions.

With the advancement of AI, built using techniques such as machine learning (ML), robots are increasingly being adopted in the healthcare sector [4]. While their main use is reported to be in surgery and rehabilitation units, other areas of deployment includes assistive care with dementia patients [5]. This type of SR application is called care robots (CRs). CRs exhibit conventional communication skills in their abilities to comprehend natural language, display emotions, as well as mimic conversation and understanding social cues [2]. In healthcare, CRs aim to monitor patients' well-being, assist with difficult tasks and proactively avert potential health deterioration [6]. In their work, Lambert et al. [2] highlight the main areas of applications to be in assisted living, monitoring of physical and

robotics

MDPI

Article

Beyond the Metal Flesh: Understanding the Intersection between Bio- and AI Ethics for Robotics in Healthcare

Auxane Boch [1,*], Seamus Ryan [2], Alexander Kriebitz [3], Lameck Mbangula Amugongo [1] and Christoph Lütge [1,3]

[1] Institute for Ethics in AI, Technical University of Munich, 80333 Munich, Germany; lameckmbangula.amugongo@tum.de (L.M.A.); luetge@tum.de (C.L.)
[2] School of Computer Science and Statistics, Trinity College Dublin, D02PN40 Dublin, Ireland; ryans58@tcd.ie
[3] Peter Löscher Chair of Business Ethics, Technical University of Munich, 80333 Munich, Germany; a.kriebitz@tum.de
* Correspondence: auxane.boch@tum.de

Abstract: As we look towards the future of healthcare, integrating Care Robots (CRs) into health systems is a practical approach to address challenges such as an ageing population and caregiver shortages. However, ethical discussions about the impact of CRs on patients, caregivers, healthcare systems, and society are crucial. This normative research seeks to define an integrative and comprehensive ethical framework for CRs, encompassing a wide range of AI-related issues in healthcare. To build the framework, we combine principles of beneficence, non-maleficence, autonomy, justice, and explainability by integrating the AI4People framework for a Good AI Society and the traditional bioethics perspective. Using the integrated framework, we conduct an ethical assessment of CRs. Next, we identify three key ethical trade-offs and propose remediation strategies for the technology. Finally, we offer design recommendations for responsible development and usage of CRs. In conclusion, our research highlights the critical need for sector-specific ethical discussions in healthcare to fully grasp the potential implications of integrating AI technology.

Keywords: care robots; bioethics; AI ethics; healthcare

Citation: Boch, A.; Ryan, S.; Kriebitz, A.; Amugongo, L.M.; Lütge, C. Beyond the Metal Flesh: Understanding the Intersection between Bio- and AI Ethics for Robotics in Healthcare. *Robotics* **2023**, *12*, 110. https://doi.org/10.3390/robotics12040110

Academic Editor: Sylwia Łukasik

Received: 3 June 2023
Revised: 27 July 2023
Accepted: 29 July 2023
Published: 1 August 2023

1. Introduction

Social robots (SRs) are defined by Fox and Gambino [1] as human-made artificial intelligence (AI) technologies, presented in a digital or physical form, with some degree of human or animal-like attributes. According to a recent review study, there are five main areas where SR technology could potentially be adopted: companionship, healthcare, education, social definition, and social impact [2]. The authors detail that the expected qualities of SRs lie in their abilities to make decisions, have conversations, and react to social cues. Interestingly, when it comes to decision-making, SRs do not seem expected to be moral, but to make the most efficient decision regardless of social implications [3]. Currently, the research surrounding SRs as reported by Lambert et al. [2], focuses mainly on personalisation and social awareness of the tool with the aim to create adaptable social agents with abilities to recognise social cues and mimic emotions.

With the advancement of AI, built using techniques such as machine learning (ML), robots are increasingly being adopted in the healthcare sector [4]. While their main use is reported to be in surgery and rehabilitation units, other areas of deployment includes assistive care with dementia patients [5]. This type of SR application is called care robots (CRs). CRs exhibit conventional communication skills in their abilities to comprehend natural language, display emotions, as well as mimic conversation and understanding social cues [2]. In healthcare, CRs aim to monitor patients' well-being, assist with difficult tasks and proactively avert potential health deterioration [6]. In their work, Lambert et al. [2] highlight the main areas of applications to be in assisted living, monitoring of physical and

mental well-being, and enhancer for social learning experience for patients with autistic spectrum disorders [2]. While CRs refer to all types of assistive care robotics, in the context of study we use the term "care robots" to refer to a specific type of SR created to assist or, in certain cases, replace human caregivers when providing care to vulnerable populations [7]. This type of SR is also referred to as a "socially assistive robot" in the literature [8]. CRs are believed to have the potential to benefit the health care systems around the world by helping to fulfil an increasing care demand and improving the quality of care services provided [9]. While this type of technology comes with potential positive improvements for prospective patients and carers, it also sparks ethical concerns when considering the patients' best interests [10,11]. For instance, patients presenting with physical, cognitive, or emotional impairments due to their condition are considered as a vulnerable population and could see benefits from using CRs in their activities of daily living (e.g., grooming, feeding, moving). On the other hand, discussions have to be had when considering their rights, and understanding of the implications in regards to, e.g., privacy and autonomy when adopting such technology. The most common ethical considerations discussed in the literature to date relate to deception, independence, and informed consent as it relates to data governance and privacy, as well as autonomy [8,10,12–14]. Furthermore, current frameworks do not explicitly integrate both perspectives of AI and bioethics for practical implementation for CRs [15–17].

To facilitate the actionability of ethics for AI in healthcare, and especially for CRs, we propose to bridge this gap. This normative research thus aims to define an integrative and comprehensive ethical framework for CRs, to enable an ethical analysis that encompasses a wide range of issues relevant to AI in health care, and, based on the framework, to present guidelines for the development and use of CRs.

2. Methodology

Recognising the unique nature of the medical sector and its strong tradition of ethics centred on human values, the integration proposed aims to comprehensively address the ethical implications of technology in healthcare, to finally reach comprehensive guidelines for the precise case of CRs [8]. The normative approach taken is defined as "A theoretical, prescriptive approach (...) that has the aim of appraising or establishing the values and norms that best fit the overall needs and expectations of society" [18]. Moreover, we build on past work such as Van de Ven's [19], Edgett's [20], and others [21–24] to build ethical frameworks and recommendations based on inductive conceptual discussions supported by empirical arguments present in the literature.

2.1. Reconciling Both Perspectives

The study advocates for integrating the perspectives of bioethics and AI ethics to propose a sector-specific approach to ethical discussions in the healthcare domain. This reconciliation will be carried out by conceptually discussing the integration of both AI and bioethics principles from Jones' [25] framework, and the AI4People framework for a Good AI Society [26].

This choice of initial material is motivated by the AI4People framework's similarity in terminology to the field of bioethics [26], considering the bioethics field is the area of applied ethics most resembling digital ethics through its ecosystem approach of patients, agents, and environment. For bioethics itself, the normative framework we build on is the one presented by Jones, which integrates different views to reach a consensus on the principles for the ethics of the field [25]. The integration of both frameworks encompasses five principles: beneficence, non-maleficence, autonomy, justice, and explainability. These principles provide a foundation for ethical guidelines and are derived from both bioethical and AI ethical perspectives.

2.2. Ethical Assessment of CRs

We will then propose a "proof of concept" application of the framework for the use case of CRs through the discussion of an ethical assessment principle by principle. This proof of concept is necessary in the case of normative approach to ensure the relevance of our recommendations, and legitimise our approach such as suggested by Väyrynen [27]. As per Edgett's [20] methodology, we will support our arguments with the existing literature to provide a sufficient background from which we will then be able to draw recommendations, or guidelines, for an ethical development and deployment of CRs.

2.3. Trade-Offs Deliberation

Further, we will discuss three ethical deliberations ensuing from the use of CRs in the general population through the lance of our proposed integrated framework, and, once again, in presenting arguments proposed in the literature. The purpose is to justify the relevance and applicability of the proposed integrated framework by analysing the ethical implications and providing insights on ethical decision-making. The three proposed trade-offs discussions relate to patient centricity versus profit centricity, autonomy versus dependence, and data privacy versus efficiency.

2.4. Practical Recommendations for the Integration of Ethics in CRs Lifecycle

The final part of the paper focuses on delivering practical recommendations for the integration of ethics in the lifecycle of CRs. These recommendations are derived from our conceptual and practical discussions regarding CRs building on the literature presented in our arguments, the ethical assessment and analysis of three trade-offs. They aim to guide the ethical development, implementation, and use of CRs in healthcare settings.

3. Integrating Bio- and AI Ethics

In the next section, we take a brief look at what the two areas of literature say on the ethical implications of using technologies and care methods in health.

3.1. AI Ethics

The discussion surrounding the ethics of AI has been focusing on the ethical consequences of the technology, particularly with regards to normative principles such as human autonomy, human rights, non-discrimination, and privacy, as AI could have significant impacts on these concepts [28,29]. One key concern originates in the black box nature of some AI algorithms, rendering it difficult to understand how these algorithms make decisions, due to the complexity of the model, for example, deep learning-based models [30]. Furthermore, the literature has pointed out that AI solutions might amplify existent patterns of discrimination, owing to the standardising effect AI solutions unfold when put onto the market.

The Western approaches used to gauge the ethical effects of AI are based on considerations pertaining to virtue ethics, deontology, and utilitarianism [31]. The utilitarian approach considers an act as moral if, compared to possible alternatives, it provides a better outcome to a greater number of persons. It can thus be understood as consequential. On the other hand, the deontological theory judges actions over consequences. Thus, no matter how morally good or bad the implications of a behaviour, or of a decision, some choices are morally forbidden. Based on the insights of this discourse, scholars and policymakers have articulated ethical frameworks [26], applications of existing normative frames such as Human Rights [32], soft laws [33], sector-specific standards and legal approaches such as the "European Union Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts" [34] to mitigate risks posed by the under-, over-, and misuse of the nascent technology.

Most of these frameworks have been addressing rather abstract features of AI. Owing to the plethora of AI use cases in human resources, finance, and health, and their different

normative implications, breaking down general moral implications of developing and deploying AI presents a key challenge within the field of AI ethics. Meanwhile, this research gap has been partly addressed by more recent scholarly contributions [35,36]. In particular, autonomous driving has been discussed intensively, particularly due to dilemmas occurring in unavoidable crash situations [37]. Furthermore, the literature has examined AI used in human resources, facial recognition technologies and finance [38,39]. However, one quintessential sector needs to be studied further as its integration of AI technology grows, namely AI in the healthcare sector and its ethical limitations. A recent paper proposed relational ethics to rethink and ground AI ethics in healthcare [40]. The paper further highlighted "the need for non-Western ethical approaches to be utilised in AI ethics more broadly".

3.2. Bioethics

The health sector is dominated by a specific moral tradition dating back to antiquity, such as depicted in the well-known Hippocratic oath (400 B.C.). The founding document of the discipline of bioethics is the oath of Hippocrates that calls on practitioners in health to "help patients" (beneficence), to "do no harm" (non maleficence), and practice medical confidentiality [41]. The notion of patient centricity seeks to secure that treatment methods and clinical practices are in the interest of the patient, especially in situations characterised by conflicts of interest between the best of the patient and the personal interests of the practitioner [42]. Patient autonomy constitutes another key value of medicine, implying that the will of the patients is decisive for the course of action adopted by practitioners [43]. Furthermore, the notion of justice suggests not discriminating between patients due to personal and individual characteristics [44,45]. These values has been encapsulated in the bioethics principles: beneficence, non-maleficence, autonomy, and justice [25]. These four principles are action-guiding for the clinical treatment of patients and aim to prevent breaches of ethical standards or acts of omission [46]. Nevertheless, there are still situations where the normative implications arising from this set of principles remain unclear. In dilemmatic situations, practitioners might have to decide between different pillars of bioethics, for example in triage situations or in situations when the will of the patient remains unclear.

3.3. The Quest for Reconciling Both Perspectives

In this paper, we reconcile both ethical approaches using CRs as a use case. The discourse surrounding SRs, including CRs, goes beyond the realm of traditional bioethics and is embedded within a broader academic, technological, and legal discourse concerning the advancement and implementation of artificial intelligence in our societies [47,48]. The use of CRs constitutes one of the applied cases in which AI and bioethics have to meet to consider fully important morals arising from the use of the technology. We thus argue the need for a framework that integrates and reconciles both perspectives, as has been conducted in other applications of AI in health [49].

The medical sector, which deals with personal and essential human issues such as survival, health, and well-being, has a long tradition of ethics that reflects its human-centric nature. We thus argue that to fully comprehend the ethical implications of integrating AI into the medical sector, it is necessary to respect and integrate the traditional vision of bioethics into technological ethics discussions. This approach enables a comprehensive understanding of the implications and ensures that the view fits the context. On the other hand, the AI ethics perspective looks into the societal and systemic impacts of AI implementations. This integration thus allows for a necessary multi-level view, from micro to macro, of the possible impacts of AI integration in the healthcare sector [8].

The AI4People framework for a good AI society [26] is an example of a framework that builds on bioethics terminology to address problematic AI ethics issues: beneficence, non-maleficence, justice, and autonomy. The framework has been developed based on four bioethics principles terms, while acknowledging the difference in interpretation as it

pertains to the bioethics perspective. We here propose a short introduction to each of the four common terminologies as defined by both perspectives, as well as a reconciliation of their interpretation to create an integrative framework [26,50].

3.3.1. Beneficence

The bioethics principle of beneficence encompasses the obligation to contribute to a person's welfare, entailing the responsibility to take actions that promote the well-being of others [25]. When designing interventions and provisions, the primary goal should be to directly benefit the patient. This approach emphasises actively engaging in activities that positively impact another individual's welfare, rather than simply refraining from causing harm. It necessitates proactive measures aimed at providing assistance and support.

On the other hand, the AI ethics perspective defines the principle as the need for creating AI prominently benefiting the well-being of people and the planet though positive economical impact, sustainability promotion, and safeguarding, as well as the empowerment of populations [26].

By acknowledging the common goal of promoting well-being and considering the specific contexts of bioethics and AI ethics, a reconciliation can be achieved. This entails designing interventions and provisions that actively contribute to the welfare of individuals while ensuring AI technologies have a positive impact on society, the environment, and human empowerment.

3.3.2. Non-Maleficence

Non-maleficence is defined in bioethics as the obligation not to inflict harm on other persons, referring to the responsibility of individuals to refrain from causing injury or negative consequences to others [25]. It emphasises the importance of avoiding or minimising harm to the best of one's ability. This principle serves as a foundational belief in the mission statements of medical professionals, often exemplified by the Hippocratic Oath.

On the other hand, the AI ethics perspective offers non-maleficence as a principle that highlights the importance of avoiding harm or negative consequences when developing and utilising AI technologies [26]. While the goal is to create AI systems that have positive impacts, it is crucial to be cautious about potential risks and misuse, such as issues related to personal privacy, security, and accountability.

The reconciliation of perspectives on non-maleficence involves a shared commitment to avoiding harm and negative consequences. In bioethics, it pertains to the responsibility of individuals, particularly medical professionals, to minimise harm and injury to others. In AI ethics, it extends to the development and use of AI technologies, emphasising the need to prevent harm and ensure positive impacts. By recognising this common commitment and applying the principle in both contexts, a reconciliation can be achieved, prioritising the minimisation of harm in healthcare and responsible AI development and deployment.

3.3.3. Autonomy

In bioethics, autonomy is defined as the respect for persons, entailing the recognition of the inherent worth and dignity of individuals, emphasising that humans should be treated as ends in themselves rather than mere instruments or tools [25]. This principle encompasses the fundamental right to autonomy, including the freedom to make decisions about one's own body and personal choices.

The AI ethics perspective understands autonomy as finding a balance between the decision-making power retained by individuals and that which is delegated to artificial agents [26].

The reconciliation of perspectives on autonomy involves recognising the common foundation of respecting individuals as ends in themselves. In bioethics, autonomy emphasises the inherent worth and dignity of humans, acknowledging their right to make decisions about their own bodies and personal choices. From the AI ethics perspective, autonomy involves striking a balance between individual decision-making and delegation

to artificial agents. By acknowledging the importance of individual agency and finding a harmonious equilibrium between human and machine decision-making, a reconciliation can be achieved, honouring both the principles of respect for persons and the dynamic relationship between humans and AI.

3.3.4. Justice

In bioethics, justice refers to the fair and equitable distribution of both health outcomes and health care services. It necessitates the careful consideration of prioritisation and rationing [25]. Allocating resources in a just manner does not have a single universal approach, as different systems employ multiple prioritisation strategies in combination to strive for a fair distribution.

In AI ethics, justice encompasses the fair and equitable distribution of AI's decision-making power and its consequences, considering the societal disparities in autonomy. It emphasises the promotion of equity, elimination of discrimination, shared benefit, shared prosperity, and equal access to AI's good doings [26]. Justice also addresses concerns related to biased data sets, defending solidarity in systems like social insurance and healthcare, and rectifying past wrongs through AI, such as eliminating unfair discrimination and promoting diversity.

The reconciliation of both definitions lies in the shared value of fair distribution, whether in health outcomes and healthcare services or AI's decision-making power. Both emphasise equitable distribution, considering societal disparities, avoiding discrimination, and employing various strategies for fairness. They strive for equity, equal access to benefits, and address issues of bias and solidarity. By recognising this common thread, adapting principles to specific contexts, and upholding fairness and equity, reconciliation is possible.

3.3.5. Explainability

In addition, the AI ethics framework expands to include the dimension of explainability, which addresses the black box character of AI solutions [26]. If chosen to be followed, the principle of explainability requires developers of AI solutions to, as an example, explain the general rationale and methodology behind an AI solution, the data used to train the model, and the data governance decisions and actions surrounding them. This principle is particularly relevant in the healthcare setting, where patients have a legitimate reason to know how AI has detected a specific illness and on what factors a health-relevant recommendation has been based [51].

3.3.6. The Reconciliation

In conclusion, reconciling the perspectives of bioethics and AI ethics requires recognising the common goals and values shared by both approaches. The integration of these perspectives becomes crucial when considering specific use cases such as CRs and the broader deployment of artificial intelligence in society. By combining the traditional vision of bioethics, which focuses on individual well-being and human-centric ethics, with the systemic and societal considerations of AI ethics, a comprehensive understanding of the ethical implications can be achieved. We argue this approach to be forceful proposition for future ethical consideration of AI in the medical sector.

4. Ethical Assessment

In the following, we develop a non-comprehensive ethical assessment for CRs based on the proposed integrative framework.

4.1. Beneficence

Beneficence implies that a technology or treatment method is conducive to conditions or situations perceived as desirable such as the promotion of well-being, economic gains or considerations relating to environmental sustainability [26]. However, when looking closer at the matter, bioethics and AI ethics offer different perspectives on the principle

of beneficence, or at least tend to emphasise different aspects. From the perspective of AI ethics, beneficence is often interpreted in the sense of wider gains for humankind and associated with frameworks such as the United Nations Sustainable Development Goals (UN SDGs), sustainability or human rights [32]. On the other hand, the bioethical interpretation of beneficence revolves around the patient in terms of the reduction of physical pain, higher life satisfaction, or the general improvement of the physical or mental condition of the patient [52,53]. Thus, the implementation of CRs should here be understood as needing to be beneficent for the patient ecosystem, and society as a whole (AI ethics perspective), but also with a strong enhancement of care offered to the individual patients (bioethics perspective).

A major concern as it relates to the future of care is the ageing of society. Indeed, as life expectancy increases and fertility rates fall worldwide, aged care services are under increasing strain. Globally, the number of older people 60 years and older is expected to double to 2.1 billion by 2050. In the same period, the number of people older than 80 years or older is expected to triple to 426 million [54]. This point is not just a concern for the individual patient, but also for the economic system and other stakeholders in the health sector such as caregivers. As it happens in parallel to the expectation of a workforce crisis linked to the global shortage, ageing, and burnout among physicians, and to the coming increasing demand for chronic care, the implementation of technological solutions is required to face the incoming crisis healthcare systems will soon face [55]. This situation can lead to a change in the way people are cared for and increase the demand for CRs [56]. Such solutions could be deployed to help with the high demand for care, ensuring that patients are timely attended to, reducing the pressure on healthcare facilities and services [57]. Moreover, without quick and implementable solutions to reduce the pressure on health systems and practitioners, the cost of healthcare will increase. This will require a paradigm shift in how care is delivered. Thus, the implementation of CRs fulfilling tasks in a similar way to, or supporting human carers by, saving scarce resources such as time and allowing for a better arrangement of the labour force in care, while increasing productivity, and quality of service for the well-being of all [4].

On the patient side, a successful and beneficial example of CRs can be found in the implementation of Paro [58] and NAO [59]. In this sense, patient centricity suggests streamlining the development and deployment of CRs to improve access to individualised care. Nevertheless, the precondition is here that CRs present an improvement of the status quo from the perspective of the individual patient.

The quintessential challenge of CRs lies therefore in the following. Owing to the immense pressure to adapt traditional caregiving to ageing societies, the deployment of CRs needs to prioritise patients over other stakeholders. This is suggested by the traditional interpretation of bioethics with its focus on patient centricity. Gains in other areas such as working conditions for caregivers are relevant too and are likely to enhance the treatment of patients. Nevertheless, improvements for other stakeholders or the realisation of other considerations such as economic or financial ones through the introduction of CRs must not lead to a deterioration of patient care. In other words, beneficence understood from the wider AI ethics perspective must not hollow out the established principle of patient centricity.

4.2. Non-Maleficence

Non-maleficence implies that a course of action or technology deployed should not create harm or risks for human beings [26]. Again, we can observe that bioethics and AI ethics create different implications here for the deployment of technology.

Bioethics specifies non-maleficence as the prevention of risks and harms to patients. This definition covers not only physical harm but also traumatic experiences that might be created by a specific treatment method or therapy. In the context of CRs, negative impacts or primarily are discussed in the context of psychological effects created by the loss of human interaction. Moreover, malfunctioning CRs might unfold adverse effects

on vulnerable groups. One example would be individuals with dementia that are more likely to be affected by deception [60]. While this argument has been discussed widely in traditional care, earlier studies have shown that the elderly in specific are often subject to cyber criminality [61], a phenomenon that could be exacerbated by the increased use of CRs.

The AI ethics discourse also highlights the relevance of the personal data of patients. AI4People has mainly defined non-maleficence in the AI context as data privacy concerns [26]. This is relevant especially when considering the relevance of health data, but also the close communication between CRs and patients. According to this argument, people might not be aware of the volume of data that robots are collecting, where that data is uploaded, or how it will be used. This lack of awareness would inhibit giving informed permission [62]. While laws like the European General Data Security Regulation (GDPR) [63] provide a few layers of protection in the European environment, these types of laws and regulations have their limitations. The novelty of social machines is in their ability to sense, process, and record the entirety of a patient's environment, as well as their augmentation of daily medical or non-medical routines [62,64]. When thinking of having a home CRs, the main goals could be to ensure that a patient takes their prescriptions on time, while constantly monitoring the patient's position in space to inform emergency assistance in case of a fall. The patient may comprehend what the robot performs, but it does not mean they are aware of the continual data collecting required for the robot to function well. Therefore, it is necessary to specify how data from CRs will be gathered and processed, how much of these data should be retained or uploaded to the cloud, how to obtain consent for doing so, and how to stop unauthorised external actors from obtaining personal specific information. The development of specific regulations must be taken into consideration as a result of new technology advancements that enable robots to acquire more data about their surroundings than ever before. In Europe, the GDPR [63] could be expanded, taking into account the characteristics of ML algorithms needed in social privacy [65]. However, the specific use case of CRs will remain problematic owing to the amount of data collected. Developers and deployers of AI solutions will therefore need to take care of the management and governance of data.

The AI ethics perspective uses a wider definition of non-maleficence, which also includes other stakeholder interests. As we consider possible harm to society, the replacement of carers by robots comes as a strong concern in the literature [66], bringing scepticism towards robots and their deployment. If we not prepared for a change in the ecosystem, consequences such as unemployment, and thus a possible decrease in the quality of life for the target population, are potential negative second and third order effects of CRs adoption. But a major point to consider is the political impact such deployment could have on societies and globalisation. It is thought that robots replacing the low-skilled workforce could strengthen populism and anti-immigration sentiments [67].

Thus, the "right way" to implement CRs in healthcare systems have to be considered on a global scale, but also on an individual scale to prevent foreseeable harm to individual and societies. Solutions might involve strong strategies in the re-orientation of other tasks for persons seeing their work done by robots, but also the education and monitoring of populations on cybersecurity, data collection, and artificial intelligence, as well as the requirement for stronger regulations around data governance and reliability of CRs abilities and cybersecurity architecture.

4.3. Autonomy

The third principle of autonomy calls for the consideration of individuals' right to make enlightened decisions [26]. In bioethics, this concept focuses on the right to make informed decisions for one's medical care. When considering this definition alone, the black box issues for complex AI systems seem to be a strong case against the use of AI in healthcare as a whole and will be discussed in the explainability part of this paper. In

AI ethics, autonomy focuses on the right to decide what decision power to give to the AI system.

In the case of vulnerable populations, it may happen that the patient is not in a condition to decide on their own care settings, for example, if no physicians are available due to a shortage of time and practitioners and the only time-sensitive solution is the use of CR to support the needed care [55]. In this context, the willingness of the patient's care ecosystem to delegate decision-making tasks to AI systems for the purpose of efficiency could go against the principle of autonomy. A balance needs to be achieved to protect the individual choice of the patient to delegate care tasks concerning their own health and well-being to AI systems. The human should always have the possibility of reversing, not implementing, stopping, and starting all decisions made by the CR.

A long-term consideration concerns the potential creation of functional emotional dependency on CRs over time for the patients [13]. Considering that CRs here are a specific type of SRs, their main features and goals are to build, develop, and maintain relationships over time with people. Through physical behaviours or spoken communication, an SR can express social and emotional cues. These queues may cause attachment to develop towards such machines [68,69]. Furthermore, Boch et al. [70] argue that, the more a robot is perceived by the users as autonomous and emotional, the more their attachment towards it seems to grow [71–73]. Interestingly, attachment towards CRs does not always happen. If it is present at a high level though, it stays consistant throughout time [74]. The ensuing feelings developing towards the CRs might lead to the creation of a relationship perceived as structured, real, and evolving throughout time between the user and the CRs. This relationship is a unidirectional one, defined in other areas of science as para-social relationships [75,76]. Such relationships, experienced as genuine, might raise opportunities for better care, by, for example, enhancing users' will to listen to the robots care instructions [77,78] and possible risks regarding emotional distress if CRs are taken away [79]. It is to be considered that in this context, patients with mental health and ability impairments, as well as children, are to be seen as more vulnerable than other populations to this possibility, but all populations can develop this type of relationship [58,59]. The creation of a para-social relationship can also be linked to the notion of emotional trust [70,80]. This irrational sense of trust poses a risk when issues of responsibility arise [62]. Indeed, the consequences of such a relationship could be simple enjoyment of CRs's company and go as far as an emotional dependency. In this case, the autonomy of the patient could be at risk, as its decision-making process could be impacted by its affection for the robot, and thus create a situation in which the human loses part of its freedom of choice, especially as it relates to the robot itself. For example, as a consequence of attachment, patients could refuse to part from the CR, regardless of the beneficence it brings to their health, or the threat it can be to their privacy.

Following up on this first worry, the European Parliament expressed its concern in 2017 that robots deployed in care settings may dehumanise the action of providing care by limiting human–human interaction [81]. In order to determine whether a care service is successful and can be categorised as "preserving the meaning", key social characteristics that relate to it, including friendliness, should be thoroughly defined [82]. On the other hand, the use of CRs for specific tasks could support carers in reducing their workload and allowing them to spend more time on the human side of their relationship with patients, counterbalancing this concern and reducing the stress put on such workers [83]. As a counterpoint to this argument, the implementation of CRs could support caregivers in taking over some of the laborious work, thus giving them back some autonomy; a frequent issue when carers attempt to prioritise their daily activities [84].

Finally, the use of CRs in everyday life can help compensate for functional losses and promote everyday skills, supporting or restoring the independence—in this context understood as the ability to achieve tasks without the need for help from another human—of individuals [85]. If this point is short, it is a major one as it pertains to the daily autonomy of a human. The implications on an individual's life can thus be incomparable in allowing

them to, e.g., live at home for a longer time and avoid the crowding of nursing homes for an elderly population requiring only support and monitoring.

Thus, CRs bring possible threats to patients' autonomy, and concerns can be raised in regard to their use on an everyday basis, but their implementation could be highly beneficial for individuals' independence and bring strong support to healthcare workers. In detail, trade-offs need to be considered on a case-by-case strategy to bring more beneficence to patients and systems.

4.4. Justice

Justice, as a bioethics principle, relates widely to resource allocation, including the availability of novel and experimental therapies, the highest possible treatment quality, and ordinary healthcare [44]. In the case of CRs, it could translate into their availability and accessibility to the global population, as well as their adptability to different cultures and needs. Adding layers to this definition, AI ethics calls for the avoidance of discrimination by AI systems based on individual and personal characteristics (e.g., gender, age, ethnicity), and the creation of shared benefits on a global and individual level, while preventing the creation of new harm, or the enhancement of existing ones [26].

Starting with the bioethics perspective, it can be stated that equal access to CRs technology is unlikely to be met on a worldwide scale as SRs are entering different societies at various rates. As presented in Boch et al. [70], the fastest growing markets for SRs in the coming years (between 2021 and 2026) will be led by economically developed countries; the USA and countries in Oceania and East Asia project the highest, with Europe and Canada projecting a medium growth rate [86]. This leaves a big part of the world out of the equation; South America, the Middle East, and Africa, are all predicted to have a low rate of growth as it relates to robotic technologies and the costs associated with the development and thus use of the technology [86]. Significant disparities and inequalities can be observed in terms of regional inclusion and participation in discussions surrounding the development, design, and implementation of CRs [70]. These considerations gain importance when acknowledging the existing divide between the Global North and Global South, which already influences the development of numerous AI-driven technologies available in today's market, as discussed in the literature but also brought to light in the 2023 World Economic Forum annual meeting [87,88]. A specific example of such bias impact on accessibility is facial recognition features. Algorithms allowing for such technology allow, in CRs, for a higher level of human–machine interactions. In the case of vulnerable minorities, already facing structural bias in society, the error rate in those systems disproportionately affects them [89].

Growing this argument into the AI ethics perspective of justice, the risk of entrenching existing inequalities due to the lack of geographical and other background diversity in the creation and development team could participate in enhancing bias towards specific populations and thus generate additional issues in the case of deployment in countries or regions that are initially non-targeted [90]. Moreover, when looking at the specific implementation of AI systems in the healthcare sector during the pandemic, screaming examples of the accentuation of discrimination against specific groups have been noted even in geographical target populations [91,92]. As it relates to CRs, those concerns are highly relevant; if a CR makes health related decisions for patients towards whom algorithmic bias plays, the well-being and even the life of the patient could be at risk.

Thus, the accessibility to the beneficence of CRs technology on a global scale is not a given, and neither is its equality in performance with every type of target population. When narrowing down the scope to the individual level, the reduction of algorithmic bias against interpersonal characteristics of vulnerable populations has to be a paramount point of concern for the developers, and care ecosystem of patients to avoid harm at all costs.

4.5. Explainability

Explainability is the one principle pertaining to only AI ethics, enabling other principles to co-exist, and allowing for accountability [26]. It is a complex but foundational requirement in all AI-enabled healthcare technology [93].

The challenge of explainability exists at two separate but interconnected levels when discussing AI-enabled robotics. At one level there exists technical explainability, the ability to explain and understand the mathematical weighting and prioritisation that the underpinning advanced statistical models used to create the AI [94]. As an example, the visual system of a SR may take in and process a 360-degree field around it but may only make decisions using a small subset of this visual information. In an environment where decisions may need to be audited or reviewed, then what exact information is used in the decision-making process needs to be clearly explainable. The feasibility of this type of explainability depends on the type of statistical model and the technical decisions made during the development of AI-enabled robotics.

This challenge is further confounded when considering the types of explainability available. Some models are "White-box", where the exact approach is identifiable and clear and all of the reasoning involved can be reviewed [95]. In situations like this, an auditor would be able to give a deterministic and affirmative explanation as to why the robot did what it did. This sits in contrast to "Black-box" models where, if the decision can be explained at all, it is intuited via a set of external analysis tools [96]. In this situation, an auditor may only be able to give a probable, non-deterministic answer as to why a decision was made.

Which level of explainability required in a healthcare environment has not yet been regulated explicitly but adjacent laws such as GDPR [63] expect data processing to be handled in a transparent and explainable manner [97]. As CRs begin to see adoption there, the expected level of explainability will need to be planned for in advance and needs to be part of the requirements at the earliest stage of conception.

At the next level, there exists socio-technical explainability, the need to understand the context in which the systems are used and on which levels they affect our everyday life [98,99].

Going back to the bioethics perspective on autonomy, patients (when cognitively and physically able) have the right to make their own decisions regarding their medical care with all the information presented to them. In the current state of AI, the use of SR in healthcare cannot ensure full transparency in the decision making process of complex systems such as CRs. A strong need for "White Box" implementation and an understandable explanation for the patients' abilities should be at the heart of CR development to allow for full autonomy of the patient in understanding all technical aspects pertaining to using such tools in their medical routine.

On the other hand, users should be informed in a transparent way about the benefits and harms that might emerge from the interactions with CRs [70]. When considering accountability as a necessary point of ethical technology, explainability of the tool and processes around it are at the centre [100]. Providing a comprehensive view of potential adverse outcomes, including physical and psychological harm that may arise from the utilisation and engagement with CRs, is of the utmost importance. This necessitates a clear delineation of the circumstances in which problems might occur. For instance, the consequences of establishing a para-social relationship between users and CRs are currently uncertain. This highlights the importance of conducting further empirical research and establishing accountability mechanisms before the widespread adoption of social robots in personal care contexts. Some argue that genuine friendships can develop between humans and robots, and thus CRs [101], while others point out the issue of deception inherent in such a connection [34,102]. Thus, the potential for para-social relationships to bloom between the patient and the CRs is real and comes with both positive and negative possible consequences. For instance, users may unknowingly trust and confide in CRs based on their evolving relationship, thereby sharing more personal information and data [103].

This highlights the need to provide understandable explanations regarding the actual functioning and data usage of CRs, and enabling users to understand the implications of their exchange with CRs regardless of their perceived relationship.

In addition, when considering the context of elderly care, trade-offs may arise as the efficacy or beneficence of CRs may affect users' autonomy. Striking a balance between a more autonomous social robot and cultural considerations is thus crucial [104]. Indeed, users' cultural environments might moderate their comfort levels in delegating tasks to CRs. Therefore, different societal contexts may require specific sets of values to be embedded in their robotic systems, and the set of values needs to be presented clearly to the patient prior for their understanding and approval of the use of the CRs.

Thus, from a technical and socio-technical perspective, explainability is of paramount importance for CRs implementation. Their use for healthcare purposes in addition to their social purpose creates a particularly challenging context when it comes to the need for transparency and understandability.

5. Ethical Trade-Offs Deliberations

In this section, we will now discuss the three main ethical trade-offs arising from the use of CRs in healthcare through the scope of our integrative framework. We will first consider patient centricity as an ethics of care requirement, versus profit centricity, which is a business requirement. We will then discuss patients' autonomy versus dependency, and how to balance the risks. Finally, the question of data privacy rights versus efficiency of the robot will be addressed.

5.1. Patient Centricity vs. Profit Centricity

The use of CRs in healthcare presents a complex ethical trade-off between patient centricity and profit centricity. Patient centricity stems primarily from the bioethics perspective of beneficence, while profit centricity can be argued as belonging to the economical "do good" vision of the principle through the AI ethics lens.

Patient centricity is considered to be more than just an ethical requirement but a social responsibility in the healthcare sector globally. Russo's [105] work on the topic partly builds on Werhane's [106] theory emphasising the importance of maximising the treatment and well-being of designated populations while also respecting the rules of the game, such as operating within a free market. However, informative asymmetries may prevent the optimal situation in which all patients involved are satisfied, highlighting the need for regulatory agencies to put a frame on the market. Borgonovi's theory [107], on the other hand, focuses on the healthcare organisation's ability to carry out its function in the best way possible, which requires efficiency, shared definition of health strategies, and environmental protection. Finally, Emanuel and Emanuel [23] propose a theory based on 'New Contractualism', which emphasises the equality of fundamental rights of all parties involved and is based on a concept of justification rooted in a social contract between stakeholders. This theory attempts to balance the economic interests of shareholders, the social impact of meeting patient needs, and the expectations of healthcare organisations, while ensuring accountability. Russo [105] thus identifies three main dimensions that are important for the social responsibility of healthcare organisations: maximising the treatment and well-being of designated populations, carrying out the healthcare organisation's function in the best way possible, and providing a justification and being held responsible for actions by another party. Finding the trade-off between all parties to obtain to a balance between profit and patient centricity is thus an ongoing discussion in the healthcare sector at the global level. Interestingly, Collins [108] found that healthcare managers tend to prioritise patient care over profit maximisation. However, the pressure healthcare managers face to produce higher results with fewer resources may inadvertently test their moral fortitude and social consciousness. Future healthcare managers may strongly focus on patient care but may still require guidance to ensure ethical and socially responsible decision-making. Thus, providing the best care for patients is at the centre of current behaviours and objectives

from a personal and organisational social responsibility perspective in healthcare, with a layer of accountability that healthcare providers have to answer to, while considering the profit of an organisation. These principles align with the principle of patient centricity requirements while integrating the profit an organisation needs to ensure survival. Those issues can be translated to the product level, such as for CRs.

CRs can promote patient centricity by providing personalised care and support that meets the unique needs and preferences of each individual, while improving the effectiveness of care [109]. Their use enhances the patient experience, improves health outcomes, and empowers patients to take an active role in their care [110]. However, the overuse of such technologies could lead to the deficiency in individualised care by humans, which might in turn reduce the quality of care received [110]. From a personal perspective, CRs can also prioritise the economic interests of the patient and of the healthcare system through the reduction of costs and increasing efficiency of care. Indeed, providing personalised care that meets the unique needs of each individual can reduce the likelihood of adverse events, hospital readmissions, and unnecessary procedures, all of which can result in significant economic costs [111]. From a group perspective, CRs can reduce the cost of care for entire systems by increasing efficiency and reducing the need for human labour. For example, robot-based systems for telemedicine have economic value and can potentially provide proper and timely medical care to patients in medically underdeveloped regions [112].

However, it is argued that the deployment of robotics technology in the field of care is increasingly focused on standardisation and selection into economic and marketable care measures, which can unintentionally produce CRs that rely less on traditional, care-intrinsic knowledge [111]. For instance, referred to as the "Silver Economy" by the European Commission [113], increasing numbers of elderly people will create a new "silver" market of consumers, and are being targeted as a new consumer group for assistive technologies.

In summary, the use of CRs in healthcare presents a complex ethical trade-off between patient centricity and profit orientation. Healthcare organisations have a social responsibility to maximise the treatment and well-being of designated populations, carry out their functions in the best way possible, and be held accountable for their actions by stakeholders. When looking at the specific technology of CRs, it is recognisable that they can promote patient centricity by providing personalised care and support, improving the patient experience, and reducing economic costs. However, the overuse of CRs and their focus on standardisation and marketability could result in the deficiency of care-intrinsic knowledge and compromise the safety and well-being of individuals. Hence, the deployment of CRs must be guided by ethical and socially responsible decision-making to balance patient needs, economic interests, and social expectations.

5.2. Autonomy vs. Dependence

Here, we deep dive into the trade-off CRs present between promoting autonomy and dependence to the tool. This trade-off pertains to the Autonomy principle as seen by both ethics as it relates to the autonomy of an individual, and their use of a technology.

We here define autonomy as the independent living ability of an individual [114], and the capabilities related to everyday life regardless of the physical impairment or ageing faced by a patient, hoping for the technology to reduce avoidable hospitalisation or institutionalised care [111]. Autonomy is moreover viewed as the individual's right to make informed decisions, based on the cultural ideal that agents are independent, rational, and self-interested. In order to be autonomous, people must be able to freely make decisions that are not influenced by coercion and that reflect their own thought processes [115]. Finally, we argue that the care ethics perspective of keeping the patient at the centre of all concerns also enhances the need for available and adequate care-giving services to fulfil patients' everyday care autonomy.

Taking into account the definition of autonomy that we have presented, the implementation of CRs may potentially enhance patients' autonomy by satisfying their requirements in three key areas: (1) the continuous demand for care, which surpasses the capacity of

human caregivers; (2) the prevalence of patient abuse in the care of others, leading to diminished autonomy and dignity; and (3) the inadequacy of current practices to meet the expected level of care, resulting in patients' compromised autonomy, dignity, and health [116]. Moreover, CRs have the potential to promote autonomy by enabling individuals to maintain their independence, agency, and control over their healthcare decisions, also due to the availability of options in regards to services, which can lead to improved quality of life and better health outcomes [117].

On the other hand, the efficacy and resilience of CRs in fulfilling its intended duties are likely to evoke heightened levels of trust from patients [118]. Trust can thus support the use of such technology, but in some cases, result in overuse of and over reliance on the CRs, which might end up creating dependency. In this case, patients would tend towards dependence on the robot rather than independence thanks to the robot, leading to a loss of agency and possible social isolation.

An instance in which individuals may experience a reduction in their ability to self-govern may arise when they follow the recommendations made by robots. Studies have indicated that people tend to be more compliant when instructed by robots, as compared to when given instructions by other humans [119]. While this feature may be advantageous in aiding patients with autism or those undergoing challenging behavioural modifications, there is a valid concern that people may be unduly influenced or coerced into carrying out actions that they would not otherwise, due to the novelty of the technology or the absence of companions to discuss alternate courses of action with [12]. In this case, social isolation is thus a risk and a factor. Furthermore, the literature suggests that there is a concern regarding the adoption of CRs for elderly patients, as it may exacerbate their sense of loneliness. It is vital for healthcare providers to acknowledge the dignity and independence of older adults, as well as their right to participate in social and cultural activities when introducing new technologies into their care [120].

Thus, a trade-off needs to be found, and might have to be case by case. In regards to care ethics, we understand that agency and autonomy of the patients go hand in hand when considering decisions of care services they want and agree to access.

Thus, when considering the use of CRs, the patient's consent—or that of their family if the patient is not able to give enlightened consent—is necessary. The use of CRs should also be stoppable and retractable without any conflict at any point [121]. Moreover, this entails that detailed information regarding the use of CRs are to be given to the patient to allow for a full understanding of the limitations and risks of the technology before making their decisions. Interestingly, in some cases, patients might be open to limitations in their own autonomy if it is necessary to ensure their safety in the long term [122]. Thus, prioritising the robot's role in promoting safety for patients seems the better road to reach an agreement and keeping the patients' interests at the centre. Therefore, the terms of robot use should be discussed and agreed upon in advance between all the involved stakeholders [122]. Finally, the most important point is to use CRs in a transparent manner to allow for the patient's autonomy to be intact. To ensure that patient autonomy is respected, guidance should be developed on how to implement applications, including when and how consent should be obtained, and how to handle matters related to vulnerability, manipulation, coercion, and privacy. Building on the Clinical Trial Regulations on informed consent proposed by the European Patient Forum [123], we emphasise here the need for detailed and clear discussions with the patient regarding their right to privacy, an adapted presentation of the limitations and risks associated with the proposed CRs solution with confirmation of understanding on the patient's part through questions or tests, but also the integration of a dynamic consent process through which the patient can access the information they require ongoingly regarding the tool. The capacity of the patient to receive and understand such information is moderated by but is not limited to their cognitive abilities, their spoken language, and other diversity characteristics to take into consideration.

In summary, the use of CRs presents a trade-off between promoting autonomy and dependence on the tool. CRs have the potential to enhance patient autonomy by providing

continuous care, preventing abuse, and offering options for improved quality of life and better health outcomes. However, over-reliance on CRs may result in dependency and loss of agency, leading to social isolation and a reduction in the ability to self-govern. To ensure patient autonomy, stakeholders need to find a balance that prioritises the patient's safety while respecting their right to make informed decisions. In other words, the use of CRs should be transparent, involve patient consent, and be accompanied by guidelines that address issues related to specific situations. Ultimately, the goal should be to keep the patient at the centre of all concerns and to provide adequate and available caregiving services that fulfil patients' everyday care autonomy.

5.3. Data Privacy vs. Efficiency

One of the major ethical concerns that arises with the use of CRs is the trade-off between data privacy and the efficiency of CRs in their tasks. This trade-off belongs to the principle of non-maleficence as perceived by both perspectives when it comes to ensuring the privacy of a user, but also touches on the principle of beneficence when it comes to "do good" and efficiency. Efficiency in this context refers to maximising CRs' accuracy and effectiveness in delivering care reliably. The collection and analysis of large amounts of patient data are crucial for training machine learning algorithms that can adapt to the specific needs of individual patients. Moreover, to interact naturally with humans, social robots rely on sophisticated algorithms and collect large amounts of data both about users and their environment [124]. For example, current generations of SR are equipped with sensors such as cameras, and GPS sensors [125]. Additionally, CRs have the ability to establish internet connections and transmit data collected by their integrated cameras and microphones to remote servers [121]. The collection of personal data enabled by the variety of sensors on CRs is necessary for their proper functioning. However, the use of sensors such as cameras and microphones can infringe on the privacy of patients. Similarly, the transmission of personal data via the internet risk potential exposure of personal data through hacking and cyber attack. Here, data privacy should be discussed to understand the acceptable balance between data privacy and the efficiency of the robot.

Data privacy is a fundamental right guaranteed by Article 8 of the European Convention on Human Rights and the Universal Declaration of Human Rights [126]. Privacy concerns have been raised by users, patients, and caregivers when considering the use of CRs [127,128]. Lutz et al. [124] define privacy in social robots, which is thus applicable to CRs as defined in this paper, in four categories: (1) informational privacy: quantity and confidentiality of data, potential security breaches, the ability for third-party access, connectivity to cloud services, the opaqueness of data collection processes, and a lack of understanding on the part of users; (2) psychological privacy: psychological dependence, diminished self-reflection and human autonomy, chilling effects as a result of feeling surveilled, and specific concerns for vulnerable user groups such as children; (3) social privacy: the social connection established between the robot and the user, accompanied by feelings of fondness and trust, which may result in the disclosure of confidential information; (4) physical privacy: capacity to access areas that are private, or that users may not be able to access, and the discomfort of being too close.

Thus, social robots, and thus CRs, present unique privacy risks aligned with their collecting of sensitive information. Moreover, the possibility for users to create emotional bonds with their robots and interact with them in more open and intimate ways leads to increased privacy risks. Finally, owners tend to forget that data collection is ongoing while interacting with their robots [121]. Informed consent is therefore crucial for users as they are at risk of not being (made) aware of the variety of data collected while using CRs.

To address privacy risks, Lutz et al. [124] recommend the use of privacy by design and privacy as contextual integrity frameworks. Interestingly, a small amount of training data seems to be necessary for the high accuracy of the system in similar cases [129], and human-generated feedback could facilitate personalisation [130]. Moreover, the reuse of existing data sets seems to be a valid solution for re-collection of data in specific situations

to learn end-to-end skills [131], meaning that the collection of large amounts of data from the patient might not be necessary for every situation. The concept of data minimisation is here relevant to implement, understanding which type of data are needed for efficiency, and making sure to focus collection on those information solely. Signalling data collection could be a viable alternative, building on the concept of explainability [124]. In this case, the users might also receive an explanation and thus understand more clearly which types of data is being collected, and be in control of what they are willing to share or not, and with whom. In the context of CRs, the questions also relate to sharing medical information with family and care staff, in addition to providers and third parties. Having an honest map of which data will be used by the system to provide better care is also of paramount importance as the trade-off with efficiency is clear. The patients should thus know what will or will not help their tool better its services. In addition, ensuring the understanding of all stakeholders involved in the CRs ecosystem (e.g., patients and their direct family ecosystem, doctors and nurses) regarding the robot's data collection and governance is crucial. In this context, finding the proper balance with transparency also means understanding the target population and adapting the discourse to the patient's abilities. Only by doing so can the full consent of using the tool given.

Another potential solution to privacy is federated learning, a privacy-preserving ML technique that involves sharing the model instead of the data [132]. This implies that a base model will be sent to CRs, learn from data, and make inferences locally without sending user data to a central server. While this approach addresses concerns about data sharing, it introduces new technical limitations related to the accuracy, transparency, and security of models developed using federated learning. Therefore, caution should be exercised when considering the adoption of solutions from domains where federated learning is not a standard practice into the healthcare field, as such solutions may not be feasible within a federated environment.

In conclusion, the use of CRs presents ethical concerns related to data privacy and the efficiency of CRs in delivering reliable care. While the collection and analysis of large amounts of patient data are crucial for training machine learning algorithms and improving CRs' accuracy and effectiveness, it also raises concerns about data privacy. The collection of (personal) data allowed by the variety of sensors present on CRs is quite important and necessary for their proper functioning. However, issues regarding data privacy need to be discussed to understand the acceptable balance between data privacy and the efficiency of the robot. To address privacy risks, CRs must be designed and implemented in a way that complies with data protection regulations such as the General Data Protection Regulation (GDPR) to protect patient privacy. Technological solutions such as anonymisation, encryption, and design that signals data collection are also recommended [124]. In addition, finding the proper balance in regards to transparency and explainability means understanding the target population and adapting the discourse in regards to data governance to the patient's abilities [133].

6. Discussion

In this paper, we reconcile the theoretical frameworks of bioethics and AI ethics to create an integrated ethical framework to guide the design, deployment and use of CRs. As a proof of concept, we explored ethical trade-offs accounting for multiple ethical perspectives. Furthermore, we provide practical resolution and recommendations, including finding the adequate discourse for each patient and stakeholders to understand the technology they are using and what it entails, the establishment of regulations and standards, and prioritising the patients' interests.

In the past decade, there has been significant growth in research on the ethics of robotics, especially in the field of healthcare. For instance, Stahl and Coeckelbergh [134] argued that in addition to ethical analysis, traditional technology assessment, and philosophical speculation, it is crucial to incorporate forms of reflection, dialogue, and experimentation that closely align with innovation practices and real-world contexts of the use of

robotics in healthcare. However, there are few studies that are focused on ethics of CRs. Bradwell et al. [135] surveyed 64 adults after they have had interactions with companion robots. Their study, demonstrated disparities between ethical concerns discussed in philosophical literature and the considerations that influence the decision-making process of purchasing a companion robot. These discrepancies, observed between philosophers and end-users involved in the care of older individuals, as well as differences in the methods used to gather information, highlight the need for additional empirical research and discussion. Another study outlines the ethical challenges associated with robotic care assistants and proposes potential strategies for addressing them through their design and use [136]. Finally, a series of frameworks and recommendations have been proposed [15–17]. Those bring a lot to the conversation, but do not comprehensively review the AI ethics and bioethics approach. Rather, they give pointers to methodologies on how to integrate ethics in general in AI in healthcare, and who should be responsible for it. Moreover, existing frameworks are not tailored for the use case of CRs. Our approach is to bridge the existing gap by integrating both AI and bioethics in an ethical framework for CRs, ensuing in the proposition of practical design recommendations.

6.1. Practical Design Constraints for CRs

In terms of implementation, developers of CRs should ensure that the system adheres to AI ethical principles such as beneficence, non-maleficence, autonomy, justice, and explainability as understood in the integration of both AI and bioethics perspectives. This can be achieved by prioritising the well-being and safety of users at the same time as averting potential harm, ensuring users are not coerced into decisions, promoting fairness and equity in healthcare, being transparent and establishing the means to hold developers accountable for the decision processes of SR. It is necessary to collaborate with healthcare providers, users, and AI ethicists to develop a well-rounded robot that prioritises patients' autonomy and well-being.

To ensure the involvement of relevant parties in the early stage of CR development, multiple methodologies can be implemented. First, focus groups can provide a platform for different groups to come together and share perspectives, experiences, and expectations regarding the given technology [137]. This methodology, nevertheless, might come with its own set of challenges when facing vulnerable populations, which might be the case in the context of CRs for autism, or CRs for dementia patients [138]. A second possible methodology to involve different target populations opinions and ideas is to deploy user surveys, interviews, and observation [139]. This set of data collection allows a more personalised understanding of the target users to be obtained. Finally, public consultations or meetings can facilitate a broader engagement with the community at large. This methodology provides open dialogue, and a platform for knowledge sharing, as well as the opportunity to understand a community's values [140]. By employing these approaches, we can establish a collaborative and inclusive environment where stakeholders are actively involved in shaping the development of CRs in healthcare. Their input and insights can guide decision-making, address ethical considerations, and ensure the development of CRs that align with societal needs and values.

From a methodological point of view, developers should adopt ethics by the design approach, which ensures that ethical principles are not just an afterthought but human values are considered from the offset and maintained throughout the AI pipeline. Practically, developers of SR should consider the following:

1. Beneficence: CRs should be designed to promote the well-being and safety of the users. This could involve incorporating features that encourage healthy behaviour or providing personalised medical advice based on the user's health data. It is also important to ensure that the robot does not inadvertently cause harm, such as by providing incorrect medical advice or failing to respond appropriately in an emergency.

2. Non-Maleficence: CRs should be designed to avoid causing harm to the users. This requires careful consideration of the potential risks and benefits of the robot's actions.

For example, if the robot is providing medical advice, it should be based on accurate and up-to-date information and should be tailored to the individual needs of the user. The robot should also be programmed to recognise and respond appropriately to potentially harmful situations, such as detecting physical or mental signs of distress in the user, but also be able to understand a negative feedback from the patient. In other words, CRs should be able to understand the limits of the patient by taking explicit feedback, spoken or perceived.

3. Autonomy: CRs should be designed to respect the autonomy of the users. This means that the robot should not coerce or manipulate the users into making decisions that they do not want to make. Instead, the robot should provide information and support that enables the user to make informed decisions about their health and well-being.

4. Justice: CRs should be designed to promote fairness and equity in healthcare. This could involve incorporating features that address healthcare disparities or providing access to healthcare resources to under-served communities. It is also important to ensure that the robot does not perpetuate or reinforce biases or discrimination in healthcare.

5. Explainability: CRs should be designed to be transparent and accountable in their decision-making processes. This requires that the robot's algorithms and data sources are open and explainable to the users and healthcare providers. The robot should also be programmed to provide clear and understandable explanations for its actions and recommendations.

As acknowledged at the start of this section, there will be specific ethical requirements within the sub-field. This will be domain-, culture-, and potentially user-specific. However, the discussed practical design constraints will need to be conceptual design goals that must be considered as part of the foundational stage for robotics in healthcare. Further ethical considerations will need to build on or augment these constraints.

6.2. Limitations and Outlook

Our study is not without challenges. First, we take a normative approach to address a problem that could be understood as technical. We recognise that our contribution, if not highly technical, is still of importance in the decision making of using, designing, and implementing technologies such as CRs.

Second, this study was not conducted systematically.

Finally, our proposition rests on one case study and thus might not be generalised as is. To overcome this issue, we would offer a few pointers for the adaptation of our framework to different healthcare use cases and contexts. First, cultural sensitivity, through the understanding of cultural diversity and existing normative frameworks is paramount [40]. Second, we would encourage strong stakeholder engagement from diverse backgrounds in the definition and further deliberations attached to the framework application and its use case [141]. Thirdly, an iterative approach to the adaptation of the framework is recommended. This entails ongoing evaluation and refinement of the guidelines based on feedback, empirical evidence, and real-wold implementation. This will allow for its ongoing improvement and adaptation to the specific context [142]. It is important to note that while we provide these general strategies, the precise method of adaptation may vary depending on the specific healthcare context and cultural considerations. Therefore, further research and collaboration with stakeholders will be essential to develop and refine the adaptation process in each unique setting.

Nevertheless, we believe this paper to be a strong contribution to the conversation, and a necessary one to build towards a sector-specific understanding of ethics for the integration of AI systems. We thus believe that future research should first reproduce our integrative approach for different use cases of healthcare, integrating both bio- and AI ethics to evaluate different AI applications and their frame of use.

Second, we would encourage the creation of quantifiable characteristics to evaluate the adherence to ethical principles for the integration of AI in healthcare, building on our proposed integration of perspectives.

We also encourage work towards the standardisation of social AI systems, building on psychology and other human sciences knowledge, depending on the specific context and sector of use. The current frame regarding regulation and standardisation for technology is rapidly evolving, with, amongst other regulatory efforts, the upcoming "Regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (AI Act) and amending certain union legislative act" [123]. This proposed regulation calls for a fundamental rights assessment, and brings with it the possibility of the creation of an EU AI Office to provide guidance and coordination in the implementation of the Act from a legal perspective [143]. On the other hand, regulation is not the only way to go when it comes to ethics. Certifications and standards developed both by governmental and non-governmental organisations can contribute to the accountability mechanisms by introducing specific technical and tangible criteria to reach ethical standards with design technologies. The ongoing work of ISO in developing standards is a notable example, or the existing work of IEEE regarding standardisation on this topic [33,144]. Finally, internal auditing and guidelines can also establish a proactive approach to ensure ethical and legal compliance. Google, for instance, has developed an end-to-end internal auditing framework that offers guidance for responsible implementation [145]. By incorporating these strategies, we can establish robust mechanisms to monitor and ensure compliance with regulations and standards for the ethical use of CRs in healthcare, if adequate standards and regulations are adopted and developed for this specific context. These approaches can promote transparency, accountability, and responsible innovation while safeguarding patient well-being and societal trust. We believe in the need to clearly understand the context (e.g., culture and frame of use) to build and use appropriate technologies.

7. Conclusions

The implementation of CRs in the healthcare sector is seen as a plausible solution to address the impending demographic and workforce challenges. The adoption of CRs, however, gives rise to various ethical concerns and opportunities, which require a holistic analysis from both the AI ethics and bioethics perspectives. Integrating both approaches allows for a multi-level analysis of the situation, where bioethics focuses primarily on the patient care practice, while AI ethics examines the implications of technology for groups and society. By taking this approach, we can ensure that ethical and technical trade-offs are adequately defined to meet performance expectations while safeguarding patients and the healthcare ecosystem they belong to. In discussing those trade-offs, some major points for the reduction of risks are put forth: (1) finding the adequate discourse for each patient and stakeholders to understand the technology they are using and what it entails; (2) the creation of guidance through regulations and standards on the state of the art to follow, accompanied by a clear accountability system for developers, providers, and users; and (3) always keeping the patient's interests at the centre of all deliberations. In addition to ethical considerations, practical design constraints for social robots in healthcare must be taken into account. Developers should adhere to ethical principles such as beneficence, non-maleficence, autonomy, justice, and explainability. Incorporating these principles into the design process and adopting an ethics-by-design approach can help prioritise the well-being and safety of users, avoid harm, respect users' autonomy, promote fairness and equity in healthcare, and ensure transparency and accountability in decision-making processes. It is important to acknowledge that specific ethical requirements may vary depending on the domain, culture, and users involved. However, the discussed practical design constraints serve as foundational goals that must be considered in the development of robotics in healthcare. Further ethical considerations should build upon and augment these constraints, taking into account the specific context and needs of CRs in providing care. Finally, we conclude that a sector-specific approach to ethical discussions is indeed needed to provide a complete understanding of the potential implications of integrating AI technology into healthcare systems.

Author Contributions: Conceptualisation, A.B., S.R., A.K., L.M.A. and C.L.; methodology, A.B., S.R., A.K., L.M.A. and C.L.; project administration, A.B.; supervision, C.L.; validation, A.B., S.R., A.K., L.M.A. and C.L.; visualisation, A.B., S.R., A.K., L.M.A. and C.L.; writing—original draft, A.B., S.R., A.K. and L.M.A.; writing—review and editing, A.B., S.R., A.K., L.M.A. and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: Seamus Ryan's contribution has been supported in part by Science Foundation Ireland under Grant number 18/CRT/6222.

Data Availability Statement: No new data were created for this research.

Acknowledgments: This work was supported by the Institute for Ethics in Artificial Intelligence (IEAI) at the Technical University of Munich.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fox, J.; Gambino, A. Relationship development with humanoid social robots: Applying interpersonal theories to human–robot interaction. *Cyberpsychol. Behav. Soc. Netw.* **2021**, *24*, 294–299. [CrossRef]
2. Lambert, A.; Norouzi, N.; Bruder, G.; Welch, G. A Systematic Review of Ten Years of Research on Human Interaction with Social Robots. *Int. J. Hum.–Computer Interact.* **2020**, *36*, 1804–1817. [CrossRef]
3. Malle, B.F.; Scheutz, M.; Arnold, T.; Voiklis, J.; Cusimano, C. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In Proceedings of the 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Portland, OR, USA, 2–5 March 2015; IEEE: Piscataway, NJ, USA; pp. 117–124.
4. Niemelä, M.; Heikkinen, S.; Koistinen, P.; Laakso, K.; Melkas, H.; Kyrki, V. *Robots and the Future of Welfare Services—A Finnish Roadmap*; Aalto University: Otaniemi, Finland, 2021.
5. Morgan, A.A.; Abdi, J.; Syed, M.A.; Kohen, G.E.; Barlow, P.; Vizcaychipi, M.P. Robots in healthcare: A scoping review. *Curr. Robot. Rep.* **2022**, *3*, 271–280. [CrossRef]
6. Broadbent, E.; Garrett, J.; Jepsen, N.; Ogilvie, V.L.; Ahn, H.S.; Robinson, H.; Peri, K.; Kerse, N.; Rouse, P.; Pillai, A.; et al. Using robots at home to support patients with chronic obstructive pulmonary disease: Pilot randomized controlled trial. *J. Med. Internet Res.* **2018**, *20*, e8640.
7. Vallor, S. Carebots and caregivers: Sustaining the ethical ideal of care in the twenty-first century. In *Machine Ethics and Robot Ethics*; Routledge: Milton Park, UK, 2020; pp. 137–154.
8. Boada, J.P.; Maestre, B.R.; Genís, C.T. The ethical issues of social assistive robotics: A critical literature review. *Technol. Soc.* **2021**, *67*, 101726. [CrossRef]
9. Dawe, J.; Sutherland, C.; Barco, A.; Broadbent, E. Can social robots help children in healthcare contexts? A scoping review. *BMJ Paediatr. Open* **2019**, *3*, e000371. [CrossRef]
10. Wagner, E.; Borycki, E.M. The Use of Robotics in Dementia Care: An Ethical Perspective. In *Informatics and Technology in Clinical Care and Public Health*; IOS Press: Amsterdam, The Netherlands, 2022; pp. 362–366.
11. Riek, L.D. Healthcare robotics. *Commun. ACM* **2017**, *60*, 68–78. [CrossRef]
12. Fiske, A.; Henningsen, P.; Buyx, A. Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *J. Med. Internet Res.* **2019**, *21*, e13216. [CrossRef]
13. de Graaf, M.M.A.; Allouch, S.B.; van Dijk, J.A.G.M. Long-term evaluation of a social robot in real homes. *Interact. Stud. Soc. Behav. Commun. Biol. Artif. Syst.* **2016**, *17*, 461–490. [CrossRef]
14. Fosch-Villaronga, E.; Poulsen, A. Sex care robots. Exploring the potential use of sexual robot technologies for disabled and elder care. *Paladyn J. Behav. Robot.* **2020**, *11*, 1–18. [CrossRef]
15. Vallès-Peris, N.; Domènech, M. Caring in the in-between: A proposal to introduce responsible AI and robotics to healthcare. *AI Soc.* **2021**, *38*, 1685–1695. [CrossRef]
16. McLennan, S.; Fiske, A.; Tigard, D.; Müller, R.; Haddadin, S.; Buyx, A. Embedded ethics: A proposal for integrating ethics into the development of medical AI. *BMC Med. Ethics* **2022**, *23*, 6. [CrossRef] [PubMed]
17. Naik, N.; Hameed, B.; Shetty, D.K.; Swain, D.; Shah, M.; Paul, R.; Aggarwal, K.; Ibrahim, S.; Patil, V.; Smriti, K.; et al. Legal and ethical consideration in artificial intelligence in healthcare: Who takes responsibility? *Front. Surg.* **2022**, *9*, 266. [CrossRef] [PubMed]
18. Normative Approach. 2023. Available online: https://www.oxfordreference.com/display/10.1093/oi/authority.20110803100238783;jsessionid=F2BC2B6AF0277F7B5FCC93F914EC5FC8 (accessed on 30 June 2023).
19. Van de Ven, B. An ethical framework for the marketing of corporate social responsibility. *J. Bus. Ethics* **2008**, *82*, 339–352. [CrossRef]
20. Edgett, R. Toward an ethical framework for advocacy in public relations. *J. Public Relations Res.* **2002**, *14*, 1–26. [CrossRef]
21. King, S.A. Researching Internet communities: Proposed ethical guidelines for the reporting of results. *Inf. Soc.* **1996**, *12*, 119–128. [CrossRef]

22. Borgatti, S.P.; Molina, J.L. Toward ethical guidelines for network research in organizations. *Soc. Netw.* **2005**, *27*, 107–117. [CrossRef]
23. Emanuel, E.J.; Emanuel, L.L. What is accountability in health care? *Ann. Intern. Med.* **1996**, *124*, 229–239. [CrossRef]
24. Kass, N.E. An ethics framework for public health. *Am. J. Public Health* **2001**, *91*, 1776–1782. [CrossRef]
25. Jones, A.H. Literature and medicine: Narrative ethics. *Lancet* **1997**, *349*, 1243–1246. [CrossRef]
26. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef] [PubMed]
27. Väyrynen, P. Normative explanation and justification. *Noûs* **2021**, *55*, 3–22. [CrossRef]
28. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
29. Heilinger, J.C. The ethics of AI ethics. A constructive critique. *Philos. Technol.* **2022**, *35*, 61. [CrossRef]
30. Holzinger, A.; Kieseberg, P.; Weippl, E.; Tjoa, A.M. Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable AI. In Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Hamburg, Germany, 27–30 August 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 1–8.
31. Bauer, W.A. Virtuous vs. utilitarian artificial moral agents. *AI Soc.* **2020**, *35*, 263–271. [CrossRef]
32. Kriebitz, A.; Lütge, C. Artificial intelligence and human rights: A business ethical assessment. *Bus. Hum. Rights J.* **2020**, *5*, 84–104. [CrossRef]
33. *IEEE Std 7010-2020*; IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being. IEEE: Piscataway, NJ, USA, 2020; pp. 1–96. [CrossRef]
34. EU. *Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*; European Parliament, Council of the European Union: Strasbourg, France, 2021.
35. Max, R.; Kriebitz, A.; Von Websky, C. Ethical considerations about the implications of artificial intelligence in finance. *Handb. Ethics Financ.* **2021**, 577–592.
36. Drage, E.; Mackereth, K. Does AI Debias Recruitment? Race, Gender, and AI's "Eradication of Difference". *Philos. Technol.* **2022**, *35*, 1–25. [CrossRef]
37. Bostrom, N.; Yudkowsky, E. The ethics of artificial intelligence. In *Artificial Intelligence Safety and Security*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018; pp. 57–69.
38. Kriebitz, A.; Max, R.; Lütge, C. The German Act on Autonomous Driving: Why ethics still matters. *Philos. Technol.* **2022**, *35*, 1–13. [CrossRef]
39. Bonnefon, J.F.; Shariff, A.; Rahwan, I. The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. *Proc. IEEE* **2019**, *107*, 502–504. [CrossRef]
40. Amugongo, L.M.; Bidwell, N.J.; Corrigan, C.C. Invigorating Ubuntu Ethics in AI for Healthcare: Enabling Equitable Care. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago, IL, USA, 12–15 June 2023; Association for Computing Machinery: New York, NY, USA, 2023; FAccT '23, pp. 583–592. [CrossRef]
41. Jotterand, F. The Hippocratic oath and contemporary medicine: Dialectic between past ideals and present reality? *J. Med. Philos.* **2005**, *30*, 107–128. [CrossRef]
42. Robbins, D.A.; Curro, F.A.; Fox, C.H. Defining patient-centricity: Opportunities, challenges, and implications for clinical care and research. *Ther. Innov. Regul. Sci.* **2013**, *47*, 349–355. [CrossRef] [PubMed]
43. Surbone, A. Telling the truth to patients with cancer: What is the truth? *Lancet Oncol.* **2006**, *7*, 944–950. [CrossRef] [PubMed]
44. Häyry, M. Roles of justice in bioethics. *Roles of Justice in Bioethics. Elements in Bioethics and Neuroethics*; Cambridge University Press: Cambridge, UK, 2022.
45. Takala, T. What is wrong with global bioethics? On the limitations of the four principles approach. *Camb. Q. Healthc. Ethics* **2001**, *10*, 72–77. [CrossRef]
46. Lawrence, D.J. The four principles of biomedical ethics: A foundation for current bioethical debate. *J. Chiropr. Humanit.* **2007**, *14*, 34–40. [CrossRef]
47. Colonna, L. Legal Implications of Using AI as an Exam Invigilator. *Fac. Law Stockh. Univ. Res. Pap.* **2021**, *91*, 13–46. [CrossRef]
48. Hagerty, A.; Rubinov, I. Global AI ethics: A review of the social impacts and ethical implications of artificial intelligence. *arXiv* **2019**, arXiv:1907.07892.
49. Morley, J.; Machado, C.C.; Burr, C.; Cowls, J.; Joshi, I.; Taddeo, M.; Floridi, L. The ethics of AI in health care: A mapping review. *Soc. Sci. Med.* **2020**, *260*, 113172. [CrossRef]
50. Di Nardo, M.; Dalle Ore, A.; Testa, G.; Annich, G.; Piervincenzi, E.; Zampini, G.; Bottari, G.; Cecchetti, C.; Amodeo, A.; Lorusso, R.; et al. Principlism and personalism. Comparing two ethical models applied clinically in neonates undergoing extracorporeal membrane oxygenation support. *Front. Pediatr.* **2019**, *7*, 312. [CrossRef]
51. Sand, M.; Durán, J.M.; Jongsma, K.R. Responsibility beyond design: Physicians' requirements for ethical medical AI. *Bioethics* **2022**, *36*, 162–169. [CrossRef]
52. Varkey, B. Principles of clinical ethics and their application to practice. *Med. Princ. Pract.* **2021**, *30*, 17–28. [CrossRef] [PubMed]
53. Beauchamp, T.L.; McCullough, L.B. Medical ethics: The moral responsibilities of physicians. *Pers. Forum* **1985**, *1*, 46–51.
54. WHO. *Ageing and Health*; WHO: Geneva, Switzerland, 2022.

55. Meskó, B.; Hetényi, G.; Győrffy, Z. Will artificial intelligence solve the human resource crisis in healthcare? *BMC Health Serv. Res.* **2018**, *18*, 545. [CrossRef] [PubMed]
56. Sparrow, R.; Sparrow, L. In the hands of machines? The future of aged care. *Minds Mach.* **2006**, *16*, 141–161. [CrossRef]
57. Robinson, H.; MacDonald, B.; Broadbent, E. The Role of Healthcare Robots for Older People at Home: A Review. *Int. J. Soc. Robot.* **2014**, *6*, 575–591. [CrossRef]
58. Calo, C.J.; Hunt-Bull, N.; Lewis, L.; Metzler, T. Ethical implications of using the paro robot, with a focus on dementia patient care. In Proceedings of the Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 7–11 August 2011.
59. Shamsuddin, S.; Yussof, H.; Ismail, L.; Hanapiah, F.A.; Mohamed, S.; Piah, H.A.; Zahari, N.I. Initial response of autistic children in human-robot interaction therapy with humanoid robot NAO. In Proceedings of the 2012 IEEE 8th International Colloquium on Signal Processing and Its Applications, Malacca, Malaysia, 23–25 March 2012; IEEE: Piscataway, NJ, USA; pp. 188–193.
60. Tan, S.Y.; Taeihagh, A.; Tripathi, A. Tensions and antagonistic interactions of risks and ethics of using robotics and autonomous systems in long-term care. *Technol. Forecast. Soc. Chang.* **2021**, *167*, 120686. [CrossRef]
61. Age, U. *Only the Tip of the Iceberg: Fraud against Older People*; Age UK: London, UK, 2015.
62. Fosch-Villaronga, E.; Lutz, C.; Tamò-Larrieux, A. Gathering Expert Opinions for Social Robots' Ethical, Legal, and Societal Concerns: Findings from Four International Workshops. *Int. J. Soc. Robot.* **2020**, *12*, 441–458. [CrossRef]
63. Commission, E. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*; European Parliament, Council of the European Union: Strasbourg, France, 2016. Available online: https://eur-lex.europa.eu/legalcontent/EN/TXT/PDF/?uri=CELEX:32016R0679 (accessed on 2 June 2023).
64. Denning, T.; Matuszek, C.; Koscher, K.; Smith, J.R.; Kohno, T. A spotlight on security and privacy risks with future household robots: Attacks and lessons. In Proceedings of the 11th International Conference on Ubiquitous Computing, Orlando, FL, USA, 30 September–3 October 2009; pp. 105–114.
65. Müller, V.C. *Ethics of Artificial Intelligence and Robotics*; Stanford Encyclopedia of Philosophy, Stanford University: Stanford, CA, USA, 2020.
66. Ford, M. The rise of the robots: Technology and the threat of mass unemployment. *Int. J. HRD Pract. Policy Res.* **2015**, *1*, 111-112.
67. Frey, C.B.; Berger, T.; Chen, C. Political machinery: Did robots swing the 2016 US presidential election? *Oxf. Rev. Econ. Policy* **2018**, *34*, 418–442. [CrossRef]
68. Darling, K. 'Who's Johnny?'Anthropomorphic framing in human-robot interaction, integration, and policy. In *Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy (March 23, 2015). ROBOT ETHICS*; Oxford University Press: Oxford, UK, 2015; Volume 2.
69. Corretjer, M.G.; Ros, R.; Martin, F.; Miralles, D. The maze of realizing empathy with social robots. In Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Naples, Italy, 31 August–4 September 2020; pp. 1334–1339.
70. Boch, A.; Lucaj, L.; Corrigan, C. A robotic new hope: Opportunities, challenges, and ethical considerations of social robots. *Tech. Univ. Munich* **2020**, *1*, 1–12.
71. Turkle, S. In good company?: On the threshold of robotic companions. In *Close Engagements with Artificial Companions*; John Benjamins: Amsterdam, The Netherlands; Philadelphia, PA, USA, 2010; pp. 3–10.
72. Scheutz, M. The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots. *Robot. Ethics Ethical Soc. Implic. Robot.* **2011**, *1*, 205–221.
73. Darling, K. Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In *Robot Law*; Edward Elgar Publishing: Cheltenham, UK, 2016.
74. Van Maris, A.; Zook, N.; Caleb-Solly, P.; Studley, M.; Winfield, A.; Dogramadzi, S. Designing ethical social robots—A longitudinal field study with older adults. *Front. Robot. AI* **2020**, *7*, 1. [CrossRef] [PubMed]
75. Schiappa, E.; Allen, M.; Gregg, P.B. Parasocial relationships and television: A meta-analysis of the effects. In *Mass Media Effects Research: Advances through Meta-Analysis*; Lawrence Erlbaum Associates Publishers: Mahwah, NJ, USA, 2007; pp. 301–314.
76. Perse, E.M.; Rubin, R.B. Attribution in social and parasocial relationships. *Commun. Res.* **1989**, *16*, 59–77. [CrossRef]
77. Coeckelbergh, M.; Pop, C.; Simut, R.; Peca, A.; Pintea, S.; David, D.; Vanderborght, B. A survey of expectations about the role of robots in robot-assisted therapy for children with ASD: Ethical acceptability, trust, sociability, appearance, and attachment. *Sci. Eng. Ethics* **2016**, *22*, 47–65. [CrossRef] [PubMed]
78. Birnbaum, G.E.; Mizrahi, M.; Hoffman, G.; Reis, H.T.; Finkel, E.J.; Sass, O. What robots can teach us about intimacy: The reassuring effects of robot responsiveness to human disclosure. *Comput. Hum. Behav.* **2016**, *63*, 416–423. [CrossRef]
79. Sharkey, N.; Sharkey, A. The crying shame of robot nannies: An ethical appraisal. *Interact. Stud.* **2010**, *11*, 161–190. [CrossRef]
80. Glikson, E.; Woolley, A.W. Human trust in artificial intelligence: Review of empirical research. *Acad. Manag. Ann.* **2020**, *14*, 627–660. [CrossRef]
81. Directorate-General for Internal Policies, Policy Department, Citizens's Riths and Costitutional Affairs European Civil Law Rules on Robotics. 2016. Available online: https://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf (accessed on 2 June 2023).

82. Decker, M.; Dillmann, R.; Dreier, T.; Fischer, M.; Gutmann, M.; Ott, I.; genannt Döhmann, I.S. Service robotics: Do you know your new companion? Framing an interdisciplinary technology assessment. *Poiesis Prax.* **2011**, *8*, 25–44. [CrossRef]
83. Robert Koch Institute. Gesundheit in Deutschland. 2015. Available online: https://www.gbe-bund.de/pdf/gesber2015.pdf (accessed on 2 June 2023).
84. Jacobs, K.; Kuhlmey, A.; Greß, S.; Klauber, J.; Schwinger, A. *Pflege-Report 2019: Mehr Personal in der Langzeitpflege-Aber Woher*? Springer Nature: Berlin, Germany, 2020.
85. Bendel, O. *Pflegeroboter*; Springer Nature: Berlin, Germany, 2018.
86. Mordor Intellegence. Social Robots Market Size, Share, Growth, Trends: 2022–2027. 2021. Available online: https://www.mordorintelligence.com/industry-reports/social-robots-market (accessed on 2 June 2023).
87. Arun, C. *AI and the Global South: Designing for Other Worlds*; The Oxford Handbook of Ethics of AI; Oxford University Press: Oxford, UK, 2019.
88. The 'AI Divide' between the Global North and the Global South. 2023. Available online: https://www.weforum.org/agenda/2023/01/davos23-ai-divide-global-north-global-south/ (accessed on 30 June 2023).
89. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability and Transparency, PMLR, New York, NY, USA, 23–24 February 2018; pp. 77–91.
90. West, S.M.; Whittaker, M.; Crawford, K. Discriminating systems. *AI Now* **2019**. https://ainowinstitute.org/wp-content/uploads/2023/04/discriminatingsystems.pdf (accessed on 2 June 2023).
91. Leslie, D.; Mazumder, A.; Peppin, A.; Wolters, M.K.; Hagerty, A. Does "AI" stand for augmenting inequality in the era of COVID-19 healthcare? *BMJ* **2021**, *372*, n304. [CrossRef]
92. Delgado, J.; de Manuel, A.; Parra, I.; Moyano, C.; Rueda, J.; Guersenzvaig, A.; Ausin, T.; Cruz, M.; Casacuberta, D.; Puyol, A. Bias in algorithms of AI systems developed for COVID-19: A scoping review. *J. Bioethical Inq.* **2022**, *19*, 407–419. [CrossRef] [PubMed]
93. Pawar, U.; O'Shea, D.; Rea, S.; O'Reilly, R. Explainable AI in Healthcare. In Proceedings of the 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), Dublin, Ireland, 15–19 June 2020; IEEE: Piscataway, NJ, USA; pp. 1–2. [CrossRef]
94. European Parliamentary Research Service. *Understanding Algorithmic Decision-Making: Opportunities and Challenges*; European Parliamentary Research Service: Brussels, Belgium, 2019.
95. Wanner, J.; Herm, L.V.; Heinrich, K.; Janiesch, C.; Zschech, P. White, Grey, Black: Effects of XAI Augmentation on the Confidence in AI-based Decision Support Systems. In Proceedings of the 41st International Conference on Information Systems, ICIS 2020, Making Digital Inclusive: Blending the Locak and the Global, Hyderabad, India, 13–16 December 2020; George, J.F., Paul, S., De', R., Karahanna, E., Sarker, S., Oestreicher-Singer, G., Eds.; Association for Information Systems: Atlanta, GA, USA, 2020.
96. London, A.J. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Cent. Rep.* **2019**, *49*, 15–21. [CrossRef] [PubMed]
97. Brkan, M.; Bonnet, G. Legal and technical feasibility of the GDPR's quest for explanation of algorithmic decisions: Of black boxes, white boxes and Fata Morganas. *Eur. J. Risk Regul.* **2020**, *11*, 18–50. [CrossRef]
98. Ryan, S.; Nurgalieva, L.; Doherty, G. Perceived Fairness Concerns Within Pandemic Response Technology. *Interact. Comput.* **2022**, iwac040. [CrossRef]
99. Nurgalieva, L.; Ryan, S.; Balaskas, A.; Lindqvist, J.; Doherty, G. Public Views on Digital COVID-19 Certificates: A Mixed Methods User Study. In Proceedings of the CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022; ACM: New York, NY, USA; Volume 1, pp. 1–28. [CrossRef]
100. Boch, A.; Hohma, E.; Trauth, R. *Towards an Accountability Framework for AI: Ethical and Legal Considerations*; Institute for Ethics in AI, Technical University of Munich: Munich, Germany, 2022.
101. Danaher, J. The philosophical case for robot friendship. *J. Posthuman Stud.* **2019**, *3*, 5–24. [CrossRef]
102. Nyholm, S.; Frank, L.E. From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible? In *Robot Sex: Social and Ethical Implications*; MIT Press: Cambridge, MA, USA, 2017.
103. Reig, S.; Carter, E.J.; Tan, X.Z.; Steinfeld, A.; Forlizzi, J. Perceptions of Agent Loyalty with Ancillary Users. *Int. J. Soc. Robot.* **2021**, *13*, 2039–2055. [CrossRef]
104. Vanderelst, D.; Willems, J. Can we agree on what robots should be allowed to do? An exercise in rule selection for ethical care robots. *Int. J. Soc. Robot.* **2020**, *12*, 1093–1102. [CrossRef]
105. Russo, F. What is the CSR's Focus in Healthcare? *J. Bus. Ethics* **2016**, *134*, 323–334. [CrossRef]
106. Werhane, P.H. Business ethics, stakeholder theory, and the ethics of healthcare organizations. *Camb. Q. Healthc. Ethics* **2000**, *9*, 169–181. [CrossRef]
107. Borgonovi, E. La responsabilità sociale in medicina. *Mecosan* **2005**, *14*, 3–9.
108. Collins, S.K. Corporate social responsibility and the future health care manager. *Health Care Manag.* **2010**, *29*, 339–345. [CrossRef] [PubMed]
109. Lee, H.; Piao, M.; Lee, J.; Byun, A.; Kim, J. The purpose of bedside robots: Exploring the needs of inpatients and healthcare professionals. *CIN Comput. Inform. Nurs.* **2020**, *38*, 8–17. [CrossRef] [PubMed]
110. Liang, H.F.; Wu, K.M.; Weng, C.H.; Hsieh, H.W. Nurses' views on the potential use of robots in the pediatric unit. *J. Pediatr. Nurs.* **2019**, *47*, e58–e64. [CrossRef] [PubMed]
111. Maibaum, A.; Bischof, A.; Hergesell, J.; Lipp, B. A critique of robotics in health care. *AI Soc.* **2022**, *37*, 1–11. [CrossRef]

112. Jang, S.M.; Lee, K.; Hong, Y.J.; Kim, J.; Kim, S. Economic evaluation of robot-based telemedicine consultation services. *Telemed. e-Health* **2020**, *26*, 1134–1140. [CrossRef] [PubMed]
113. European Commission; Directorate-General for Communications Networks, Content and Technology; Worthington, H.; Simmonds, P.; Farla, K.; Varnai, P. *The Silver Economy: Final Report*; Publications Office: Technopolis Group: Brighton, UK, 2018. [CrossRef]
114. World Health Organisation (WHO). Active Ageing: A Policy Framework. 2014. Available online: https://extranet.who.int/agefriendlyworld/wp-content/uploads/2014/06/WHO-Active-Ageing-Framework.pdf (accessed on 2 June 2023).
115. Killackey, T.; Peter, E.; Maciver, J.; Mohammed, S. Advance care planning with chronically ill patients: A relational autonomy approach. *Nurs. Ethics* **2020**, *27*, 360–371. [CrossRef]
116. van Wynsberghe, A.L. Designing Robots with Care: Creating an Ethical Framework for the Future Design and Implementation of Care Robots. Ph.D. Thesis, University of Twente, Enschede, The Netherlands, 2012.
117. Herstatt, C.; Kohlbacher, F.; Bauer, P. *"Silver" Product Design: Product Innovation for Older People*; Technical Report, Working Paper; Institute for Technology and Innovation Management, Hamburg University of Technology (TUHH): Hamburg, Germany, 2011.
118. Hancock, P.A.; Kessler, T.T.; Kaplan, A.D.; Brill, J.C.; Szalma, J.L. Evolving trust in robots: Specification through sequential and comparative meta-analyses. *Hum. Factors* **2021**, *63*, 1196–1229. [CrossRef]
119. Broadbent, E. Interactions with robots: The truths we reveal about ourselves. *Annu. Rev. Psychol.* **2017**, *68*, 627–652. [CrossRef]
120. Coco, K.; Kangasniemi, M.; Rantanen, T. Care personnel's attitudes and fears toward care robots in elderly care: A comparison of data from the care personnel in Finland and Japan. *J. Nurs. Scholarsh.* **2018**, *50*, 634–644. [CrossRef]
121. De Swarte, T.; Boufous, O.; Escalle, P. Artificial intelligence, ethics and human values: The cases of military drones and companion robots. *Artif. Life Robot.* **2019**, *24*, 291–296. [CrossRef]
122. Jenkins, S.; Draper, H. Care, monitoring, and companionship: Views on care robots from older people and their carers. *Int. J. Soc. Robot.* **2015**, *7*, 673–683. [CrossRef]
123. European Patient Forum. *Clinical Trials Regulation: Informed Consent and Information to Patients*; European Patient Forum: Brussels, Belgium, 2016. Available online: https://www.eupatient.eu/globalassets/policy/clinicaltrials/epf_informed_consent_position_statement_may16.pdf (accessed on 2 June 2023).
124. Lutz, C.; Schöttler, M.; Hoffmann, C.P. The privacy implications of social robots: Scoping review and expert interviews. *Mob. Media Commun.* **2019**, *7*, 412–434. [CrossRef]
125. Abney, K.; Bekey, G.A.; Lin, P. Robots and privacy. In *Robot Ethics: The Ethical and Social Implications of Robotics*; The MIT Press: Cambridge, MA, USA, 2014; pp. 187–201.
126. United Nations. *Universal Declaration of Human Rights*; United Nations, New York, NY, USA, 1948.
127. Pino, M.; Boulay, M.; Jouen, F.; Rigaud, A.S. "Are we ready for robots that care for us?" Attitudes and opinions of older adults toward socially assistive robots. *Front. Aging Neurosci.* **2015**, *7*, 141. [CrossRef]
128. Draper, H.; Sorell, T. Ethical values and social care robots for older people: An international qualitative study. *Ethics Inf. Technol.* **2017**, *19*, 49–68. [CrossRef]
129. Lockhart, J.W.; Weiss, G.M. The benefits of personalized smartphone-based activity recognition models. In Proceedings of the 2014 SIAM International Conference on Data Mining, SIAM, Philadelphia, PA, USA, 24–26 April 2014; pp. 614–622.
130. Tsiakas, K.; Abujelala, M.; Makedon, F. Task engagement as personalization feedback for socially-assistive robots and cognitive training. *Technologies* **2018**, *6*, 49. [CrossRef]
131. Ebert, F.; Yang, Y.; Schmeckpeper, K.; Bucher, B.; Georgakis, G.; Daniilidis, K.; Finn, C.; Levine, S. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv* **2021**, arXiv:2109.13396.
132. Rieke, N.; Hancox, J.; Li, W.; Milletari, F.; Roth, H.R.; Albarqouni, S.; Bakas, S.; Galtier, M.N.; Landman, B.A.; Maier-Hein, K.; et al. The future of digital health with federated learning. *NPJ Digit. Med.* **2020**, *3*, 119. [CrossRef]
133. Markus, A.F.; Kors, J.A.; Rijnbeek, P.R. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inform.* **2021**, *113*, 103655. [CrossRef]
134. Stahl, B.C.; Coeckelbergh, M. Ethics of healthcare robotics: Towards responsible research and innovation. *Robot. Auton. Syst.* **2016**, *86*, 152–161. [CrossRef]
135. Bradwell, H.L.; Winnington, R.; Thill, S.; Jones, R.B. Ethical perceptions towards real-world use of companion robots with older people and people with dementia: Survey opinions among younger adults. *BMC Geriatr.* **2020**, *20*, 244. [CrossRef]
136. Johnston, C. Ethical Design and Use of Robotic Care of the Elderly. *J. Bioethical Inq.* **2022**, *19*, 11–14. [CrossRef] [PubMed]
137. Molewijk, B.; Hem, M.H.; Pedersen, R. Dealing with ethical challenges: A focus group study with professionals in mental health care. *BMC Med. Ethics* **2015**, *16*, 1–12. [CrossRef]
138. Owen, S. The practical, methodological and ethical dilemmas of conducting focus groups with vulnerable clients. *J. Adv. Nurs.* **2001**, *36*, 652–658. [CrossRef] [PubMed]
139. Park, J.; Han, S.H.; Kim, H.K.; Cho, Y.; Park, W. Developing elements of user experience for mobile phones and services: Survey, interview, and observation approaches. *Hum. Factors Ergon. Manuf. Serv. Ind.* **2013**, *23*, 279–293. [CrossRef]
140. Harrison, S.; Mort, M. Which champions, which people? Public and user involvement in health care as a technology of legitimation. *Soc. Policy Adm.* **1998**, *32*, 60–70. [CrossRef]

141. Magelssen, M.; Pedersen, R.; Miljeteig, I.; Ervik, H.; Førde, R. Importance of systematic deliberation and stakeholder presence: A national study of clinical ethics committees. *J. Med. Ethics* **2020**, *46*, 66–70. [CrossRef]
142. Stevenson, F.A.; Gibson, W.; Pelletier, C.; Chrysikou, V.; Park, S. Reconsidering 'ethics' and 'quality'in healthcare research: The case for an iterative ethical paradigm. *BMC Med. Ethics* **2015**, *16*, 1–9. [CrossRef]
143. Releases, P. AI Act: A Step Closer to the First Rules on Artificial Intelligence. 2023. Available online: https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence (accessed on 30 June 2023).
144. ISO. ISO IEC JTC 1 SC 42 Artificial Intelligence. 2023. Available online: https://www.iso.org/committee/6794475.html (accessed on 30 June 2023).
145. Raji, I.D.; Smart, A.; White, R.N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; Barnes, P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 33–44.

A2 Essay 2. Human-Robot Dynamics: A Psychological Insight into the Ethics of Social Robotics

Reference: Boch, A. and Thomas, B.R. (2025). Human-robot dynamics: a psychological insight into the ethics of social robotics. *International Journal of Ethics and Systems,* Vol. 41 No. 1, pp. 101-141. https://doi.org/10.1108/IJOES-01-2024-0034

Information about the Article

Title	Human-robot dynamics: a psychological insight into the ethics of social robotics
Authors	Boch, A. and Thomas, B.R.
Accepted	10 November 2024
Journal	International Journal of Ethics and Systems
Volume	41
Number	1
Year	2025
Pages	101-141

Copyright information

International Journal of Ethics and Systems

Human-robot dynamics: a psychological insight into the ethics of social robotics

Auxane Boch
Institute for Ethics in AI, Technical University of Munich, Munich, Germany, and
Bethany Rhea Thomas
Department of Psychology, Edge Hill University, Ormskirk, UK

Received 31 January 2024
Revised 22 August 2024
3 October 2024
Accepted 10 November 2024

Abstract

Purpose – Social robotics is a rapidly growing application of artificial intelligence (AI) in society, encompassing an expanding range of applications. This paper aims to contribute to the ongoing integration of psychology into social robotics ethics by reviewing current theories and empirical findings related to human–robot interaction (HRI) and addressing critical points of contention within the ethics discourse.

Design/methodology/approach – The authors will explore the factors influencing the acceptance of social robots, explore the development of relationships between humans and robots and delve into three prominent controversies: deception, dehumanisation and violence.

Findings – The authors first propose design factors allowing for a positive interaction with the robot, and further discuss precise dimensions to evaluate when designing a social robot to ensure ethical design technology, building on the four ethical principles for trustworthy AI. The final section of this paper will outline and offer explicit recommendations for future research endeavours.

Originality/value – This paper provides originality and value to the field of social robotics ethics by integrating psychology into the ethical discourse and offering a comprehensive understanding of HRI. It introduces three ethical dimensions and provides recommendations for implementing them, contributing to the development of ethical design in social robots and trustworthy AI.

Keywords Psychology, Ethics, Social robots

Paper type Literature review

1. Introduction

Social robotics is a rapidly growing application of artificial intelligence (AI) in society, encompassing an expanding range of applications. With its inherent social nature, integrating psychology into the multidisciplinary discussion of social robot ethics becomes an evident avenue of research.

Early definitions described social robots as machines capable of acceptable interaction with humans and other robots, effectively conveying intentions and collaborating to achieve goals (Duffy *et al.*, 1999). The applications and goals of social robots are diverse, varying

International Journal of Ethics and Systems
Emerald Publishing Limited
2514-9369
DOI 10.1108/IJOES-01-2024-0034

across specific domains. Extensive literature has explored and categorised these use cases based on their intended purposes.

As examined by Lambert *et al.* (2020) and Naneva *et al.* (2020), companion robots focus on building relationships with owners and facilitating sustained social interaction while providing domestic assistance. These robots align with the goal of fostering social bonds and promoting independence among their end users. Service robots, exemplified by bellhop prototypes in the hospitality industry, as studied by Boch *et al.* (2021) and Pinillos *et al.* (2016), play various roles with a focus on economic productivity objectives. Care robots target patients and health-care providers through tasks that enhance health-care delivery (Naneva *et al.*, 2020) or provide customised assistance to vulnerable groups (Vallor, 2011). Some initiatives prioritise paediatric care, emphasising welfare goals (Naneva *et al.*, 2020). Educational robots, explored by Angel-Fernandez and Vincze (2018) and Naneva *et al.* (2020), enhance learning experiences by supporting instruction and engaging learners, aligning with instructional technology objectives. Sex robots aim to improve intimacy, although this specific domain raises significant ethical complexities (Boch *et al.*, 2021; Fosch-Villaronga and Poulsen, 2020; Richardson, 2016). Furthermore, social interaction and entertainment robots, as examined by Naneva *et al.* (2020) and Lambert *et al.* (2020), provide valuable insights into fundamental human–robot dynamics despite lacking specific purposes.

These diverse applications reflect evolving technological, economic and social priorities. By examining the underlying factors and dynamics of the relationship between humans and social robots from a psychological perspective, we can better understand the tension points in ethical design for fostering positive human–robot interactions (HRI) and relationships.

The ethical design of social robots is a complex and multifaceted endeavour that draws upon insights from various fields, including psychology, human–computer interaction and ethics. In recent years, psychologists have played a crucial role in shaping the ethical design of social robots through empirical research and the development of conceptual frameworks.

Psychology plays a crucial role in the ethical design of social robots by investigating and informing the concept of HRI. Studies have shown that humans tend to anthropomorphise robots, attributing human-like qualities to them, which has implications for designing social robots that evoke emotional attachments (Smith *et al.*, 2021). Moral psychology has also contributed to developing ethical frameworks. For example, Malle's model integrates principles from moral psychology and HRI to guide ethical robot behaviours (Malle, 2016). In addition, psychology helps underpin the psychological mechanisms underlying human behaviour and decision-making, informing the design of social robots. For instance, considering users' cognitive biases and heuristics is essential, as these cognitive processes can shape how individuals perceive and respond to robotic behaviour (Biswas and Murray, 2014).

In conclusion, the ethical design of social robots draws upon a rich body of psychological research and theoretical frameworks. Psychologists have contributed significantly to developing ethical guidelines and design principles for social robots by understanding how humans perceive, interact with and are influenced by robots. As the field continues to evolve, psychologists will undoubtedly play a central role in shaping the future of ethical social robot design.

This paper seeks to contribute to the ongoing integration of psychology into social robotics ethics. We aim to accomplish this by reviewing current theories and empirical findings related to HRI and addressing critical points of contention within the ethics discourse. This paper is structured to provide a comprehensive exploration of the ethical design of social robots through the lens of psychological insights. The discussion begins by

examining the moderating factors that influence the acceptance of social robots, including cultural, appearance, engagement, behavioural and personalisation factors in Section 2. The paper then delves into the development of human–robot relationships, exploring key psychological theories in Section 3. Section 4 addresses potential consequences and ethical controversies associated with social robotics, focusing on deception, manipulation, dehumanisation and violence. Finally, in Section 5 and 6, the paper concludes with practical recommendations for the ethical design of social robots, considering the discussed psychological factors and ethical dimensions, and offers a roadmap for future research in this interdisciplinary field.

2. Moderating factors of acceptance

In this section, we delve into the factors influencing the acceptance of social robots, aiming to understand humans' overall perceptions and receptiveness towards these machines, building on psychology's teachings. Accepting a technology implies having a positive expectation and experience, which aligns with the ethical design objectives. Achieving acceptance involves understanding the various factors influencing individuals' attitudes towards robots and designing robots that meet their expectations and needs. By considering cultural factors, appearance factors, engagement factors, behavioural factors and personalisation, designers can create social robots that foster positive interactions and adhere to ethical principles of promoting user well-being and satisfaction.

2.1 General attitudes towards robots

Early studies provide insights into general public attitudes towards robots. According to Dautenhahn (2007), robots tended to be primarily viewed as valuable tools to assist with household tasks rather than as social companions. This perception of robots as appliances or machines aligned with their initial introduction for industrial and service applications rather than social interaction. However, more recent work has found that attitudes may depend on direct experience with robots. Naneva *et al.* (2020) reviewed survey responses regarding trust, anxiety and willingness to use robots. Overall, participants reported neutral levels of trust and anxiety towards robots – neither explicitly trusting nor distrusting robots and experiencing reasonably moderate anxiety. This suggests an open but cautious baseline attitude. However, interestingly, the review found some influence of demographic factors. Gender appeared to impact trust somewhat, with samples including more female participants reporting higher levels of trust in robots compared to more male-dominated samples. Furthermore, age did not significantly influence reported attitudes contrary to some expectations.

Building on this knowledge, studies have investigated how direct engagement with robots shapes attitudes positively. Li *et al.* (2010) observed strong correlations between interaction factors like likability, trust and satisfaction with a given robot. This implies that fostering active user participation and developing more social interaction tasks could help improve general perceptions and, in-turn attitudes towards robots. Similarly, Leite *et al.* (2013) found that increasing interaction time and frequency through varied applications may heighten user engagement with robots.

In summary, while initial public perceptions of robots aligned with their industrial origins, general baseline attitudes today appear cautiously open but neutral regarding trust and anxiety. Demographic factors like gender may influence views to a degree. Most importantly, direct experience appears instrumental in developing more positive regard, implying that engineering social engagement can help optimise HRI outcomes.

2.2 Cultural factor

Culture warrants in-depth investigation in the realm of technology acceptance (Oyibo and Vassileva, 2020; Fleischmann *et al.*, 2020; Metallo *et al.*, 2022; Boch, 2011). Culture is pivotal in shaping social dynamics and perspectives, as highlighted by Gelfand and Kashima (2016). It significantly influences how individuals interact socially and perceive themselves and others. This becomes particularly relevant when considering the development of interactions and relationships with novel social agents, such as robots, which lack humans' evolutionary familiarity with other humans (Kacancioğlu *et al.*, 2012). In addition, it seems relevant to highlight that cultural values related to robots are not static but subject to change over time, such as in response to disruptive societal events. In the case of the COVID-19 pandemic, Schönmann *et al.* (2024) found that attitudes towards perceived as "sterile" social robots in caregiving contexts shifted positively, demonstrating how moral values and acceptance of robots can evolve even in short time intervals. The authors concluded that attitudes towards care robots could change positively if their use addresses an urgent need. This highlights the importance of considering culture not just as a regional concept but also as a time and context-dependent factor that requires continuous empirical re-evaluation, especially during periods of rapid societal change.

Notably, Eastern and Western philosophies diverge in their worldviews (Lim *et al.*, 2021; Kamide and Mori, 2016). The Western approach seems to emphasise constructing a systematic understanding of phenomena, while the East seem to adopt a more holistic perspective. Some argue that this holistic perspective makes the East more receptive to concepts like animism, which could facilitate the acceptance of robots (MacDorman *et al.*, 2009). However, the increasing trend of globalisation has led to heightened cultural exposure and fusion, potentially diluting traditional values in certain contexts (Lim *et al.*, 2021). Interestingly, in the context of Islamic culture, Alemi *et al.* (2020) emphasise that cultural and religious values play a crucial role in accepting social robots, particularly in educational settings in Iran. The authors argue that to be accepted, robots must align with Islamic teachings and ethical principles, such as modesty and respect for human dignity. The perception of robots as complementary tools rather than replacements for human roles is also significant in shaping acceptance. In addition, traditional educational roles and community consensus are critical factors in whether these technologies will be adopted. This underscores the necessity of culturally sensitive design in technology implementation, particularly in regions where religion profoundly influences cultural norms.

A critical and extensively studied cultural dimension is individualism-collectivism, which examines how individuals define themselves in relation to others, as defined and explored by Markus and Kitayama (1991) and Hofstede (1980). According to this theory, highly individualistic cultures prioritise independence over relationships, whereas collectivistic cultures emphasise interdependence (De Mooij and Hofstede, 2010). This could, thus, align with a stronger adoption of social entities such as social robots, as demonstrated in Marchesi *et al.*'s (2021) study.

Communication styles also vary considering culture, with individualistic cultures favouring explicit communication, while collectivist cultures tend to use more implicit communication (De Mooij and Hofstede, 2010). Differences in communication styles across cultures extend to nonverbal cues, such as gestures, with significant social meaning (Matsumoto, 2006; Burgoon, 1994). Interestingly, Trovato *et al.* (2013) found that individuals could recognise robot emotional displays better when the robots showed facial expressions consistent with their respective cultures' nonverbal cues. This highlights the relevance of adapting visual and nonverbal cues in HRI.

Finally, numerous studies have explored how cultural backgrounds predict preferred attitudes towards robots. Participants from Eastern collectivist cultures, such as Koreans and Chinese, demonstrated greater engagement with service robots than German participants from Western individualistic cultures. Furthermore, the former group found the robots more likeable, trustworthy and satisfactory (Li *et al.*, 2010; Bartneck *et al.*, 2005). It is postulated that cultures emphasising relationships may develop more positive perspectives on social robots. In contrast, individualistic cultures with extensive exposure to industrial robots, like Germany, may prioritise robotic tool use over companionship. To improve acceptance in these cultures, it is crucial to emphasise the practical utility of social robots by focusing on how they can enhance personal productivity and convenience. Highlighting these benefits can increase their acceptance as valuable tools in daily life (Lim *et al.*, 2021). In addition, gradual exposure to robots in social contexts, starting with less intrusive roles, can help users become more comfortable and eventually accept more socially interactive robots, especially when their utility is demonstrated (Ke *et al.*, 2020).

Consequently, a culturally aware interaction design and use case can promote acceptance as people anthropomorphise shared identities, including communication cues (Lim *et al.*, 2021). However, it is important to note that cultural customisation may not always be necessary for general tasks, as humans can interact across cultural differences (Lim *et al.*, 2021). In summary, cultural psychology serves as a fundamental framework for comprehending the intricacies of HRI and acceptance.

2.3 Appearance factors

Social robots have traditionally been broadly categorised based on their physical features as humanoid (i.e. presenting human-like features, such as Nao shown in Figure 1), zoomorphic (i.e. presenting animal-like features, such as Paro shown in Figure 2), or do not fit within one of these categories, such as machine-like robots (Lambert *et al.*, 2020).

Design seems to be important in how people perceive robots and their abilities. A robot that exhibits animal-like behaviour seems to create the illusion that it will be more accommodating to the user's wishes (Lambert *et al.*, 2020). The use of zoomorphic and pet-like robots like Paro and MiRo has been documented in, for instance, the context of entering people's homes and addressing the needs of specific target groups, such as older adults or those with cognitive impairments in care settings (Henschel *et al.*, 2021). They showed great potential for the melioration of social skills and well-being of patients. Robotic pets are usually an appreciated option with elderly populations as they require less care than real animals and avoid issues like allergies and some hygiene considerations in care settings (Hung *et al.*, 2019).

More recently, social robots have been specified as technologies created by humans for interaction that may somewhat physically or behaviourally resemble people, with the goal of HRI mirroring natural human interactions, in other words, anthropomorphic interactions (Fox and Gambino, 2021). With this in mind, many developers design social robots to incorporate human characteristics while carefully avoiding too close an imitation that could cause unease, a phenomenon known as the "Uncanny Valley", initially proposed by Mori in 1970 (Pandey and Gelin, 2018). This theory suggests that as robots become more human-like, they evoke more familiarity and likability until a certain point where the mismatch between their appearance and their behaviour triggers a sense of unease (Ho *et al.*, 2008). This idea aligns with the concept of simulation theory (Turner, 1978), a well-established theory in psychology, proposing we understand the minds of others by simulating their situations and placing ourselves in their shoes. Furthermore, Krach *et al.* (2008) observed a linear relationship between the degree of anthropomorphising and cortical activation in brain

Source: Figure by authors

Figure 1. Nao, an anthropomorphic robot

areas linked to the processing of other minds, also called cognitive empathy. This suggests that when individuals anthropomorphise humanoid robots, their brain areas associated with understanding and processing the mental states of others become more active, supporting the Uncanny Valley theory.

Source: Figure by authors

Figure 2. Paro, Zoomorphic robot representing a white baby seal

Interestingly, anthropomorphism seems to dramatically influence people's responses to robots (Darling *et al.*, 2015). In addition, even though social robots can be intentionally designed as anthropomorphic (Breazeal, 2003; Duffy, 2003), studies have shown that people also tend to anthropomorphise robots with non-human-like designs (Carpenter, 2013; Knight, 2014; Paepcke and Takayama, 2010). Thus, the consequences of anthropomorphism might not occur only in the case of humanoid robots.

When comparing the likeability variation based on design choices, Li *et al.* (2010) note that zoomorphic robots are perceived as more likeable than machine-like robots, while no significant difference in likability seems notable between anthropomorphic and zoomorphic robots. It is worth noting that the anthropomorphic robot used in the experiment differentiated itself from the machine-like robot through recognisable body features such as eyes and a head. This result suggests that the mere presence of humanoid or animal-like features can increase people's familiarity with the robots and consequently enhance likability. As a result, the use application of such technology is to be considered when developing a design; zoomorphic and anthropomorphic robots may be more suitable for entertainment and caring tasks, while machine-like robots are better suited for low-sociability tasks such as acting as security guards (Li *et al.*, 2010). This confirms Goetz *et al.* (2003) early findings, arguing that human-like robots seem preferred for tasks requiring a higher degree of sociability.

Further design factors also come into play when considering likeability and acceptance. In their review, Lambert *et al.* (2020) highlighted that a more feminine robot may be perceived as less threatening than a robot with a masculine appearance.

Thus, anthropomorphism and gender association seem to impact the perceived social qualities of robots.

Interestingly, anthropomorphism seems to also shape human empathy towards social robots significantly (Riek *et al.*, 2009). In a Web-based survey by Riek *et al.* (2009), participants watched film clips featuring protagonists with varying degrees of human likeness and rated their empathy towards them. The study found that people were more empathetic towards human-like robots than mechanical-looking ones. The emotionally evocative clips depicted humans acting cruelly towards the protagonist, while the neutral clips showed mundane activities. Following the film clips, participants were asked to imagine saving one of the robot protagonists in a hypothetical earthquake scenario. The results supported the hypothesis that people exhibit greater empathy towards robots resembling humans.

Building on this knowledge, Darling *et al.* (2015) conducted a study to examine the impact of anthropomorphic framing on people's reactions to animated robots. Participants were asked to observe a small robotic toy called Hexbug Nano and strike it with a mallet. The study found that participants hesitated significantly more to strike the robot when it was introduced with anthropomorphic framing, through factors such as a name and backstory. Darling (2015) also discovered a strong relationship between participants' tendency for empathic concern and their hesitation to harm robots introduced with anthropomorphic framing. Participants exhibited empathetic responses, with many expressing concerns about hurting the personified Hexbug. These findings highlight the influence of anthropomorphic framing on people's immediate reactions to robots and suggest that personifying robots or portraying them as having lifelike experiences can elicit empathetic responses. Considering the effects of framing, it becomes evident that designing robotic technology to accrue experiences or personifying them can influence users' perception of robots and enhance emotional responses (Darling *et al.*, 2015). Introducing robots with stories and narratives can facilitate the adoption of robotic technology by enabling users to relate to robots on an emotional level.

Awareness of the influence of anthropomorphism and framing in shaping empathy towards social robots is crucial for both individuals and institutions involved in the design and deployment of robotic technology. Understanding when and how to effectively use these strategies can enhance users' emotional engagement with robots, promote positive HRI and foster acceptance of robotic technology. It might raise concerns about the consequences of such an emotional response on users' well-being. Thus, the personalisation of the robot to the use case and the end users, as well as an inclusive design practice in the development phase of the tool, contribute to the ethical design of social robots in prioritising positive interactions and meaningful engagement with humans.

In conclusion, the design choices for social robots, regardless of whether they are anthropomorphic, zoomorphic or machine-like, play a crucial role in shaping people's reactions and potential empathy towards the technology. These choices affect how much people like the robot, how familiar it feels to them and how sociable they perceive it to be. Considering these factors is vital to ensure that humans have positive emotional experiences when interacting with robots.

2.4 Engagement factors

If appearance is considered a critical aspect of robot acceptance and positive feelings for the user, ethical design choices must also encompass specific behaviours and capabilities.

A longitudinal study conducted by de Graaf *et al.* (2015) sought to understand users' perspectives on the characteristics of social robots. The study identified eight main social

characteristics that users deemed crucial for a social robot to be perceived and accepted as a social entity in their homes. The most significant factor participants identified was the robot's ability to engage in two-way interaction. Users had high expectations for social robots to respond socially, and when these expectations were not met, people experienced disappointment and dissonance. Users also emphasised the importance of robots sharing the same environment, displaying thoughts and feelings, being socially aware, providing social support and demonstrating autonomy. These characteristics collectively contribute to the robot's social presence and perceived social capabilities. While participants consistently highlighted these five characteristics, they also mentioned three additional concepts: cosiness, self-similarity and mutual respect. However, these concepts were deemed relatively less relevant in users' perspectives.

In a later study by Dereshev *et al.* (2019), long-term users of the humanoid Pepper robot were interviewed to gain insights into their experiences and expectations. These participants interacted with the Pepper robot for extended periods, from eight months to over three years. One expectation that stood out was the robot's ability to engage in reciprocal conversation, aligning with de Graaf *et al.* (2015) findings. Participants expressed disappointment when the robot was limited to a one-sided conversation structure similar to a smart speaker. This finding is consistent with an earlier usability study by Rivoire and Lim (2016), which observed a quick loss of interest among users when interacting with Pepper over several weeks.

The phenomenon of reduced engagement over time, known as the "novelty effect", has been observed in various social robotics platforms (Leite *et al.*, 2013; Tanaka *et al.*, 2015). However, it is essential to note that these results are subject to controversy and appear to be context dependent (Hung *et al.*, 2019). Further long-term studies exploring sustained positive engagement with social robots in different contexts are necessary to gain a comprehensive understanding of this phenomenon.

In addition, Li *et al.* (2010) found that participants exhibited higher levels of active response and engagement in tasks with higher sociability, such as teaching, compared to tasks with lower sociability, such as acting as a security guard. Moreover, engagement in these tasks correlated more strongly with perceived likeability, trust and satisfaction than the mere level of active response. Therefore, sustained engagement also depends on the nature of the task and the robot's perceived sociability.

In conclusion, specific behaviours and characteristics influence the acceptance and sustained engagement with social robots. Ethical design choices should consider these factors to ensure meaningful, positive and lasting interaction between humans and social robots.

2.5 Behavioural factors

Fiske *et al.* (2007) identified warmth and competence as two universal dimensions in the assessment of individuals. Warmth refers to positive social traits and emotions, while competence is associated with perceived ability. Those evaluation criteria also seem to apply to the human evaluation of robots. Studies conducted by Nass and Moon (2000) and Nass and Brave (2005) revealed that people tend to attribute social qualities to autonomous agents, including social robots, drawing from their experiences in human-to-human interactions when interacting with non-living agents (Nass and Moon, 2000). Furthermore, the robot's use of warmth- or competence-based social cognitive strategies after making an error seems to influence people's perceptions of the robot along these dimensions (Honig and Oron-Gilad, 2018). However, it's important to note that the effectiveness of these strategies may be influenced by the frequency and severity of the robot's mistakes. Studies indicate that when

robots perform poorly, there is a noticeable drop in self-reported trust (Robinette *et al.*, 2017). In addition, both small and large errors during task execution adversely affect trust, with more significant errors having a more pronounced negative impact (Aliasghari *et al.*, 2021). While mistakes generally reduce trust, the robot's recovery strategy can mitigate this impact. Robots that acknowledge their errors and communicate their intention to rectify the situation are perceived as more trustworthy than those that do not effectively address their errors (Cameron *et al.*, 2021).

This is exemplified in a recent experiment by Cameron *et al.* (2021), where perceptions of a mobile guide robot were examined. The robot used synthetic social behaviours to elicit trust after making an error. The study involved 326 participants, and the results showed that when a robot identified its mistake and communicated its intention to rectify the situation, observers considered it more capable than a robot that only apologised for its mistake. However, the robot that apologised was perceived as likeable and uniquely increased people's intention to use the robot. In this context of service, warmth seemed to be more important than competence in the intention to use the robot.

On the other hand, using warmth-based strategies by a robot can sometimes hinder perceptions of the robot's competence, which is consistent with similar outcomes in human–human interactions (Kim *et al.*, 2006) and HRI research (Kaniarasu and Steinfeld, 2014). This suggests that in the context of engaging with social robots in assistive contexts, factors such as liking and warmth may have a more significant influence on people's intentions to use the robot compared to capability and competence, as predicted by affiliation models (Casciaro and Sousa-Lobo, 2005; Shazi *et al.*, 2015).

In conclusion, perceptions of warmth and competence impact individuals' evaluations of social robots, with warmth-based strategies influencing intentions to use the robot more significantly than competence in social contexts. Interestingly, the dimensions of warmth and competence underpin the stereotype content model (Nicolas *et al.*, 2022). This model can account for specific gender-based stereotypes concerning women, who may be perceived as "respected or liked, but not both" (Connor *et al.*, 2017, p.6). Crucially, although this relates to 'human' women, there may be interesting implications and intersections with findings concerning the warmth and competence of social robots designed to reflect 'female' attributes.

2.6 Personalisation

The personalisation of social robots to the use case of the target users might be the best option for positive interactions and acceptance of the technology. Personalisation involves tailoring social responses and adapting to individual users' preferences and needs. Studies have shown that personalising HRI can reinforce rapport, cooperation and engagement between humans and robots (Lee *et al.*, 2012; Cifuentes *et al.*, 2020). For example, a study by Lee *et al.* (2012) conducted a mixed factorial study to examine the social effects of personalisation using a robot with memory retention capabilities. The study involved personalisation and no personalisation conditions, evaluating the social interaction and engagement between humans and the robot. Participants, 21 individuals, were provided snacks by a Snackbot machine over a series of weeks. The study measured various aspects of social interaction, including self-disclosure, greeting the robot by name and self-connection. The evaluation included questions on service satisfaction and the perceived value of the provided service. The authors found that personalising interactions with a robot in a service context enhanced reported scores of social interaction, cooperation and developing a relationship between humans and robots.

Later, Cifuentes *et al.* (2020) investigated inclusive design and acceptance of robots in health care. Inclusive design is a collaborative approach that enables users to contribute to the decision-making process of developers and deployers, ensuring that the resulting robots are tailored to meet their unique, personalised needs. Their results highlighted the significance of inclusive design in increasing the acceptance and effectiveness of social robots in the context of care. Inclusive design considers the robot's functional aspects and the social, cultural and ethical dimensions influencing user acceptance.

3. Relationship development

We will now delve into the fascinating field of human–robot relationships, first exploring the theoretical models used to understand these relationships and the intriguing concept of developing empathy for robots. Furthermore, we explore the role of anthropomorphism in the emotional connection to robots and the creation of para-social relationships. We then further discuss this relationship's concerns, namely, the ethical implications of deception and the potential for excessive attachment to robots. By exploring these topics, we gain valuable insights into the complex dynamics of human–robot relationships and the implications they bring forth.

3.1 Models and theories of relationship

Understanding the intricacies of HRI and how the nature of these relationships may evolve is a central aim of the HRI field of research. In their analysis, Fox and Gambino (2021) advocate a cautious approach when exploring relationship theories in this domain.

One of the prominent theoretical frameworks in the field of human–computer interaction (HCI) and HRI is the application of the social response theory, the "computers are social actors" (CASA) perspective (Nass and Moon, 2000; Sung *et al.*, 2007). This theory suggests that humans react mindlessly and naturally to media representations, treating them like their natural counterparts. Specifically, people tend to engage in overlearned social behaviours, such as reciprocity and politeness, towards interactive technology (Nass and Moon, 2000; Brave *et al.*, 2005). CASA argues that computers can exhibit social interaction potential through anthropomorphic appearance cues or behaviours, leading human users to respond to them as social beings (Fox and Gambino, 2021). Empirical research has supported for CASA's claims, demonstrating that human–robot social interactions can be influenced by cues such as gendered facial features and that robots, in the context of social interactions, can create some degree of social presence and, thus, be perceived as a social entity by the user (Eyssel and Hegel, 2012; Van Doorn *et al.*, 2017).

However, it is essential to acknowledge the inherent limitations of social robots compared to humans (Fox and Gambino, 2021). Many HRI studies have focused on brief, one-time interactions that do not capture the dynamics of relationships, which crucially require repeated exposures over time to develop familiarity between parties (Fox and Gambino, 2021). Thus, findings from one-off studies may create a misleading impression that human–robot bonds can mirror interpersonal human–human relationships. Still, perceptions often change as familiarity increases through longitudinal interaction, revealing the robots' inability to meet human social and conversational standards. Maintaining familiar interaction quality over the long term is crucial for relationship formation, and more extensive, longitudinal investigations are needed to understand the potential for human-like relationships with social robots (Fox and Gambino, 2021).

On the other hand, the social exchange theory posits that relationships involve reciprocal sharing of resources between parties, a relevant framework to consider in human–robot relationships (Roloff, 1981). Robots epistemologically do not have the personal resources,

desires or autonomy to engage in genuine exchange, and evaluating costs and benefits is challenging as robots lack human motivation and experience of rewards and punishments (Roloff, 1981; Thibaut and Kelley, 1959). Moreover, robots cannot provide the depth of self-disclosure necessary for developing intimate relationships, as they have a limited breadth of information and lack subjective experiences (Altman and Taylor, 1973). While superficial interactions may mimic human-like effects, robots cannot meet the fundamental requirements for meaningful long-term interpersonal bonds as currently designed (Fox and Gambino, 2021).

In conclusion, transferring standard relationship theories from human–human interactions to HRIs requires caution. Social robots' unique attributes and limitations call for alternative frameworks that recognise the robot's distinct nature. Approaches that draw from the human-pet or companion perspective, as well as the exploration of superhuman relational abilities, may offer valuable avenues for understanding and designing human–robot relationships (Fox and Gambino, 2021; Dautenhahn, 2004; de Graaf, 2017; Krämer *et al.*, 2011). One paramount example of such reflection is Kate Darling's book "The New Breed" (2021), which proposes strong parallels between how human relationships with animals evolved from farm tools to friends and pets. Considering our past, she assesses that a similar pattern could happen with robots. As the field progresses, HRI designers and researchers must explore novel relationship understanding models encompassing HRI-specific dynamics and potentials (Riva *et al.*, 2012).

3.2 Para-social relationship and attachment theory

The field of social robots addresses the creation of dependencies and establishing relationships with users (de Graaf *et al.*, 2016; Fong *et al.*, 2003). These robots can express social and emotional cues through physical behaviours or spoken communication, which can foster attachment and emotional connections (Darling, 2015; García-Corretjer *et al.*, 2023). The degree of perceived autonomy and emotional capability in robots influences the strength of these attachments (Turkle, 2010; Scheutz, 2012; Darling, 2016) and their perceived animacy. The perception of animacy is influenced by robots' displayed intelligence and amiability, such as perceived competence and warmth (Bartneck *et al.*, 2007; Carpenter, 2013; Knight, 2014).

Due to this emotional attachment, individuals may develop structured, genuine and evolving one-sided relationships with social robots, also described as para-social relationships (Schiappa *et al.*, 2007; Perse and Rubin, 1989). While these relationships can be experienced as authentic by users, they present both opportunities and risks, particularly concerning emotional trust and responsibility (Glickson and Woolley, 2020; Fosch-Villaronga *et al.*, 2019).

The development of anthropomorphic and emotionally expressive robots has implications for emotional trust, particularly when issues concerning responsibility arise (Fosch-Villaronga *et al.*, 2019). Emotional trust within these relationships is driven by irrational factors and nurtured through affection (Glickson and Woolley, 2020). Interestingly, users may develop loyalty and trust towards robots based on false appearances, potentially leading to the disclosure of personal information and data that they would not usually share (Reig *et al.*, 2021). In vulnerable or sensitive populations, such as individuals with dementia or autism, as well as children interacting with care or educational robots, the establishment of para-social relationships with robots introduces unique challenges (Shamsuddin *et al.*, 2012; Calo *et al.*, 2011; Angel-Fernandez and Vincze, 2018).

The consequences of para-social relationships between users and robots remain largely unknown (Boch *et al.*, 2021). While some argue that genuine friendship can develop between

humans and robots (Danaher, 2019), others counter-argue the impossibility of such an occurrence due to its conditionality (Evans, 2010). Furthermore, Nyholm and Frank (2017) highlight the inherent deception issue in such relationships. It is essential to acknowledge that robots do not yet possess the capacity to genuinely experience emotions. This emotional deception occurs when users believe robots genuinely experience emotions, leading to unrealistic expectations (Sharkey and Sharkey, 2012). This can result in users prioritising the well-being of robots over that of other individuals or their well-being. In addition, users may rely excessively on robots as social assistants without exercising their critical judgement (Fulmer *et al.*, 2009).

Linked to this type of relationship is the question of attachment. From a psychological perspective, attachment refers to the bonds and cumulative experiences that individuals form with other individuals or objects (Huber *et al.*, 2016). These bonds are influenced by factors such as shared values, attractiveness, openness and reciprocity (Huber *et al.*, 2016). Researchers have examined how this theory can be applied to HRI (Richardson, 2015).

The initial statement for Bowlby's theory on attachment styles is that early attachment experiences with caretakers shape how people respond and relate to others and result in distinct attachment styles (Shaver *et al.*, 2005). The theory was initially developed through his seminal "Attachment and Loss" trilogy (Bowlby, 1982, 1984, 1998). In this work, Bowlby proposed that children have an innate motivation to form attachments as this is biologically driven for survival. He identified three primary attachment styles associated with distinct emotional, cognitive and behavioural tendencies.

The first is a secure attachment style. Children with secure attachments experience consistent and sensitive caregiving when distressed, allowing them to view the caregiver as a safe base for exploration. This results in secure internal working models of the self as worthy of care and relationships characterised by trust. Adults with secure styles have healthy, low-anxiety relationships. Secondly, the anxious–ambivalent style emerges from inconsistent caregiving responses. When needs are sometimes met but other times neglected, children learn that caregivers cannot always be relied upon. As adults, anxious–ambivalently attached individuals greatly desire approval and proximity, simultaneously pushing others away due to underlying mistrust. Finally, avoidant attachment arises from caregivers not responding to a distressed child regularly. The child then learns to deactivate their attachment system, understanding that relying on others is ineffective. As adults, avoidant individuals emphasise independence, avoid close connections and focus more on practical matters than emotional intimacy due to expectations that others will not meet their needs. If this theory has been reworked and reframed over the years, it remains a solid basis for scholarly work (Rabb *et al.*, 2022).

Research has found that individuals could form attachments to robots, even without explicit attachment-inducing behaviours, if the robot possesses human or animal characteristics (Keefer *et al.*, 2012; Scheutz, 2012; Norris *et al.*, 2012). Empirical work has also been done on the topic, furthering the initial theoretical frame of understanding the importance of attachment in HRI.

A study by Dziergwa *et al.* (2018) investigated interactions between three participants with different attachment styles – secure, anxious–ambivalent and avoidant – and an autonomous social robot they lived with for 10 days. The securely attached participants were highly engaged with the robot, EMYS, attributing human qualities to it despite its limitations. They found joy in teaching it colours and perceived that it could understand their emotions. Data showed that this group interacted with the robot the most. The anxiously attached participant focused on the robot's technical flaws, experiencing anxiety and anger. They wanted the robot to initiate interactions more, like greeting upon return. The avoidantly

attached participants were satisfied but kept distant from the robot. They only interacted to fulfil practical needs and wanted more personalised functions. Data confirmed that this group spent the least time with the robot. Overall, results demonstrated varied satisfaction, perceptions and opinions of the robot based on attachment style. While all became attached to the robot or its functions, securely attached participants had the most positive experience. Moreover, participants perceived emotions expressed similarly by the robot differently based on their attachment. This confirms that robots need personalised characteristics for different users' attachment patterns.

A separate study by Pozharliev *et al.* (2021) found that customers with low anxious attachment style scores responded more negatively to a frontline service robot than a human agent and perceived less empathy. In contrast, those with high anxious attachment style scores did not differ in their responses between their experience with their human and robot counterparts. These findings again illustrate how attachment styles could influence HRI. As social robots continue advancing to elicit more human-like behaviours, the potential for users to form attachments also increases, raising ethical concerns about emotional distress during separation from the robot (Coeckelbergh *et al.*, 2016; Sharkey and Sharkey, 2010, 2011; Sullins, 2012).

On the other hand, it seems crucial to acknowledge that in HRI research, attachment is sometimes understood as Norman's definition (Norman, 2004), stating the concept as the sum of cumulative emotional episodes a user experiences towards a robot. Furthermore, Rabb *et al.* (2022) propose an interesting attachment framework for HRI, introducing the notion of strong and weak attachment. They define strong attachment as the presence of attachment functions defined by psychological attachment theory, relevant proximity seeking or separation distress behaviours and presence in a significant sense. The strong attachment would, thus, translate into the systematic seeking of proximity when distressed, the robot's frequent fulfilment of security or comfort needs and potentially a high degree of distress present upon an event of separation. This last point would confirm ethical concerns regarding attachment to social robots. In addition, they define weak attachments as less significant relationships, including those described by Norman, which are solely formed by cumulative positive experience or those deemed "secondary attachments" (i.e. ones which fill in gaps otherwise left by primary attachment figures).

In essence, psychology provides insights into the process of relationship formation, and the acceptance criteria discussed earlier should be considered when ethically designing social robots. This design approach aims to mitigate potential risks associated with the technology by leveraging the dynamics of the human–robot relationship.

4. Potential consequences and ethical controversies

Further than descriptive work, empirical research and psychology can help us investigate ethical questions and controversies the ethics community raises regarding the consequences of human–robot relationships. Turkle (2006, 2012) and Scheutz (2012) voiced their concerns regarding the impact of robotic technology anthropomorphisation. They expressed apprehension that the emotional connections formed with anthropomorphised robots may supplant human relationships, engender undesirable behaviours or render individuals susceptible to emotional manipulation.

4.1 Deception

Deception in the context of social robots can have both harmful and beneficial aspects. Unintentional deception may arise due to discrepancies between a robot's behaviour and actual capabilities, while intentional deception involves deliberately creating false

expectations. A clear example of robots' deception is the renowned "Turing Test", also known as the Imitation Game. The imitation game explores the possibility of being deceived by machines rather than evaluating their accurate intelligence (Turing, 1950; Bertolini and Carli, 2022) and highlights the potential for manipulation in HRIs.

Deception can manifest in various ways, including emotional deception and attachment. Social robots are designed to elicit positive emotions and leverage anthropomorphism, potentially leading to emotional deception (Van Maris *et al.*, 2020). The appearance and behaviour of social robots play a crucial role in this deception, as they can be intentionally designed to evoke a friendly and lovable appearance (Lacey and Caudwell, 2019). However, emotional deception can raise concerns significantly when vulnerable populations, such as lonely older adults, develop an emotional attachment and potentially become emotionally dependent on robots (Gillath *et al.*, 2021).

Empirical evidence supports the ethical concerns of emotional deception and attachment in social robotics. For example, VA Maris *et al.* (2020) investigated emotional deception in interactions between social robots and older adults. The results indicated that participants perceived the emotional robot as a social entity, suggesting some level of successful deception. Interestingly, participants who perceived the robot as deceptive also found the interaction more pleasant. Participants with a higher level of attachment were more susceptible to emotional deception and potential over-trust, highlighting the risks associated with emotional attachment to robots (Van Maris *et al.*, 2020).

4.2 Manipulation

Another ethical concern is the possible influence of social robots on human decision-making, regardless of attachment.

A study by Hanoch *et al.* (2021) measured participants' risk-taking behaviour in the presence of a robot. The results showed that participants encouraged by the robot took more risks in a lab research setting, suggesting that robots can influence human decision-making to some extent. However, it is essential to note that the initial attitude towards robots may moderate this influence, as their influence seems lesser on individuals with a negative attitude towards robots (Hinz *et al.*, 2019).

Interestingly, recent research conducted by Hou *et al.* (2023) sheds further light on the influence of social robots on human decision-making. Participants were paired with a human and a robot to perform decision-making tasks in their experiment. The researchers manipulated the power dynamics by assigning one of the entities as the leader. They created three conditions: human as leader, robot as leader and a control condition with no power difference. The results revealed that participants were significantly more influenced by the leader, irrespective of whether it was a human or a robot. However, participants generally held a more positive attitude towards the human leader than the robot leader, although they perceived the entity in power as more competent. This suggests that social status and perceived power play a significant role in understanding their potential impact on humans.

Another level of manipulation studied is the one on human-to-human relationships. Sakamoto and Ono (2006) studied the impact of robots on human-to-human relationships. In this context, they evaluated the relevance of the "balance theory" in HRI. This theory refers to cognitive consistency, emphasising the preference for internal consistency and balance within a cognitive system. The American Psychological Association (APA, 2023) and Heider (1958) describe balanced systems as more stable and psychologically pleasant than imbalanced systems, where elements within the system lack consistency. In their study, Sakamoto and Ono (2006) used the balance theory to investigate how robot behaviour influences human relationships within the framework of P–O–X triads. Here, P represents

the person (self), O represents another person and X represents a stimulus or event. By applying the balance theory, the researchers aimed to gain insights into how robots can shape individual impressions of others, potentially impacting the stability and dynamics of human relationships. The findings of the study demonstrate that robots could have the capacity to both foster and disrupt human-to-human relations. Through their behaviour, robots can influence how individuals perceive others, leading to changes in the nature of their relationships. Consequently, this study emphasises the significant role of robots in social dynamics and highlights the need for further exploration in this domain.

4.3 The dehumanisation of companionship and (romantic) relationships

The controversy surrounding the dehumanisation of (romantic) relationships due to social robots has garnered significant attention from scholars and researchers. The first aspect to consider regarding the dehumanisation of relationships is the potential long-term impact relationships with social robots might have on human-to-human interactions regarding empathic abilities. While empathy is an inherent trait, the manifestation of empathic responses does not always occur automatically (Decety, 2015). Instead, these reactions appear to be skills partially acquired through interpersonal and contextual experiences (Tousignant *et al.*, 2017). Darling (2016) suggests that interactions with social robots could impede the development of general empathy due to the lack of realistic emotional responses from robots or the widespread dehumanisation of relationships. Others support this theoretical worry (Sharkey and Sharkey, 2012; Turkle, 2011; Fosch-Villaronga *et al.*, 2019).

Regarding romantic relationships, critics have directed their attention towards anthropomorphism in the context of social robots. Turkle (2010) bemoans the loss of authenticity, distinguishing biological beings and robotic entities (Turkle, 2007). The author also expresses worry that engaging in seductive robot relationships, which may be perceived as less challenging than human relationships, could result in individuals withdrawing from social interactions with friends and family (Turkle, 2010). Furthermore, Turkle (2011) argues that using relationship robots such as sex robots may dissuade individuals from investing the necessary effort into establishing genuine relationships with other humans. However, regarding sex robots, the trade-offs remain largely unknown due to a dearth of evidence-based research. While this technology may offer benefits in treating sexual disorders and supporting disabled patients, there is a potential risk of desensitising individuals and fostering adverse spillover effects on human interaction or objectification (Royakkers and van Est, 2015). Without proper validation through randomised control trials, the application of sex robots in therapeutic contexts is thought to possibly exacerbate issues such as sexual violence (Fiske *et al.*, 2019). Another discussion surrounding the potential positive use of sex robots is their possible impact on reducing human trafficking and involuntary sex work (Levy, 2007).

It is imperative to note that much of the current discourse on this topic leans towards the philosophical realm, necessitating further research through data collection to provide more concrete insights. Thus, there is a pressing need to bridge the gap between philosophical debates and empirical investigations.

4.4 Violence towards robots' impact on human-to-human interactions

The issue of violent or abusive behaviour towards social robots and its potential impact on human-to-human interaction has emerged as a contentious topic within the discourse surrounding social robots. Darling (2016) argues that mistreating humanoid and animal robots could lead to negative behaviour towards sentient animals and humans. Drawing upon Immanuel Kant's (1784) notion that cruelty or tenderness towards animals can extend to

humans, Darling suggests laws protecting anthropomorphic/zoomorphic robots, similar to animal cruelty laws. Furthermore, Calo (2015) proposes the concept of a new legal subject category, somewhere between personhood and objecthood, for social robots.

Recent research by Yamada *et al.* (2023) has shed light on our understanding of children's violence towards robots. Their study revealed a gradual process of robot abuse, akin to human bullying, which unfolds in four distinct stages: initial approach, mild abuse, physical abuse and serious abuse. In addition, the presence of certain environmental factors, specifically the presence of other children, was found to play a significant role in promoting and facilitating the progression of abuse. The study identified five key factors: the presence of other children encourages the target child to approach the robot; if other children have engaged in mild abuse, the target child is more likely to do the same; if other children have resorted to physical abuse, the target child is also inclined to follow suit; engaging in joint actions of abuse with other children escalates the severity of the target child's abuse; and if children around the target child encourage, the abuse escalates further. It is worth noting, however, that not all children escalated their abuse, with a majority remaining in the mild abuse stage. Therefore, the study suggests that precautions focused on addressing mild abuse, which is the most common stage, may be the most effective approach to mitigating children's violence towards robots.

According to Darling (2015), there is a concern that engaging in violent actions towards robots may hinder empathy development in individuals. For instance, preventing children from vandalising robots goes beyond respecting others' property, as lifelike robot behaviour could influence how children treat living beings (Walk, 2016). This concern extends beyond children, as violence towards lifelike robots may desensitise adults to violence in other contexts (Darling, 2016). Likewise, the repeated use of robots as sexual partners may encourage undesirable sexual acts or behaviours (Gutiu, 2016). These concerns are echoed by Coghlan *et al.* (2019), who argue that social robots, behaving similarly to certain lower animals, have the potential to elicit strong emotions such as pity, care, callousness and cruelty. Acts of kindness or cruelty towards these robots could, thus, influence similar responses towards nonhuman animals and humans, particularly in children whose moral responses are still developing.

Interestingly, these concerns align with assumptions put forth by moral psychology theories. For instance, the social learning theory proposed by Bandura and Walters (1977) posits that learning primarily occurs through modelling, imitation and social interactions. It suggests that behaviour development and regulation are influenced by external stimuli, such as the influence of others, as well as external reinforcement, including praise, blame and rewards. Bandura later expanded on this theory in 1986, introducing the social cognitive theory, which incorporates cognitive processes, such as conceptions, judgement and motivation, in shaping an individual's behaviour and the environment that influences them. According to this perspective, individuals actively interpret the outcomes of their actions, shaping their environments and personal factors, thereby informing and modifying subsequent behaviour. In essence, individuals learn through the experiences of positive or negative social responses, which help them determine acceptable behaviour. This perspective also aligns with Haidt (2001) theory of social intuitionism, which asserts that our environment shapes our moral values. Consequently, if our environment approves certain behaviours, we are more likely to perceive them as morally acceptable. This also resonates with the first stage of moral development proposed by Kohlberg (1971), particularly in children learning acceptable behaviour through punishment and reinforcement.

Finally, witnessing such behaviours could potentially induce trauma in bystanders, as studies suggest that the neural responses activated when witnessing violence towards robots

mirror those activated when witnessing violence towards humans (Rosenthal-von der Pütten *et al.*, 2013). However, it remains uncertain whether robots can alter long-term behavioural patterns in people positively or negatively (Darling, 2015). Moreover, whether HRI is more likely to encourage undesirable behaviour or serve as a healthy outlet for behaviour that would otherwise have negative consequences is unclear. Nevertheless, as discussions surrounding violent behaviour towards robots gain attention (Parke, 2015) and the emergence of companion (and more) robots becomes a reality (Freeman, 2016; Borenstein and Arkin, 2019), it is crucial to investigate this important question.

5. Discussion

The following discussion will outline some overarching recommendations for positive social robot design, important ethical considerations and dimensions to consider and propose endeavours for future research and wider applications.

5.1 General recommendations for positive social robot design

Considering the topics discussed within this paper, there are several recommendations we propose for a positive social robot design, informed by previous research and theoretical underpinnings from psychology. These include the important role of culture, various engagement factors, alongside appearance, behavioural and personalisation factors.

Firstly, there are important cultural factors, concerning the communication styles and the use case, to consider for positive social robot design. That is, there are differences concerning technology acceptance, philosophies and communication across Western and Eastern cultures and cultures which are individualistic versus collectivist. The following recommendations are made:

- Adapt social robots' communication styles and nonverbal cues to reflect cues corresponding with the culture for better acceptance.
- Consider the use case: Collectivist cultures favour socially tasked robots (e.g. service robots, companions) where individualistic cultures prefer robots as tools (e.g. industrial).

In the case of Eastern cultures, their holistic perspective means they may be more receptive to animism, facilitating their acceptance of social robots. In comparison, designing social robots for Western cultures should consider their systematic perspective, which may be less receptive to animism and, thus, informs their behaviour and engagement with certain social robots. Communication is an important recommendation for design, as collectivist cultures, which prioritise interdependence and use more implicit communication, may have a stronger adoption of social robots, with a role focus such as service and companionship and visual and non-verbal cues reflect that of culture. In comparison, western cultures may be more receptive to robots as robotic tools instead of companions. These recommendations are informed by previous research and theoretical considerations from cultural psychology and may inform a framework for the positive design of social robots considering the impact of culture on HRI and acceptance.

Alongside the recommendations proposed concerning cultural influences, there are several recommendations considering the role of engagement related to the interactions, likeability, autonomy, animacy, social support, expectations management and tasks. More engagement means more likeability, thus, the following recommendations are made:

- The ability to engage in a two-way interaction ongoingly with, for instance, the integration of large language models such as ChatGPT.

- The demonstration of autonomy and animacy.
- The ability to display feelings and social support.
- A clear explanation by the robot of its abilities and limitations to manage expectations.
- The engagement of robot is social tasks, such as teaching and service.

Engagement is the key overarching recommendation here, specifically related to the likeability of social robots. Considering and managing the expectations of users is significant. Users demonstrate a high expectations regarding engaging in ongoing two-way interactions, and when their expectations of sociability exceed the robot's capabilities, this causes disappointment and withdrawal. In addition, the ability to demonstrate autonomy through factors such as social awareness, providing social support, displaying thoughts and feelings were emphasised by users. Crucially, it is important to balance the perceived autonomy, to avoid causing unease if a social robot is too human-like. Users are more likely to engage with and respond to tasks that exhibit higher sociability. This engagement is correlated with perceived likeability and trust, which are linked to the appearance and behavioural recommendations below.

Finally, adhering to findings from previous research, there are crucial recommendations when designing social robots appearance, considering the appearance, behavioural and personalisation factors, some of which overlap with previous recommendations concerning engagement. The following recommendations are made:

- Zoomorphic, anthropomorphic and feminised robots are more likeable than machine-like robots for tasks requiring social qualities, thus, consider the overarching design choice based on the tasks performed.
- Implement behavioural warmth-based strategies (e.g. apologies) over competent ones (e.g. path to resolution of the problem) to increase likeability and cooperation with the robot.
- Ongoing personalised interactions enhance the positive perception, thus, ensure the robot has an ability to adapt to the user it is talking with.

This ability to adapt is increasingly driven by advances in AI, for instance, through methods like continual learning (CL), lifelong learning and meta-learning. CL enables robots to adapt their perception and behaviour models in real time to cater to individual user preferences, significantly improving the robot's likability and emotional understanding (Churamani *et al.*, 2022). Lifelong learning ensures that robots adapt to evolving user preferences and contexts over time, maintaining engagement and inclusivity in various settings (Irfan *et al.*, 2023). Meta-learning allows robots to rapidly adjust to new users with minimal data, enhancing their ability to accurately predict and respond to individual movements and actions (Moon and Seo, 2021). Hybrid hierarchical learning architectures can further refine personalisation by tailoring robot behaviours based on static and dynamic user characteristics, such as cognitive biases and emotional states, thereby improving the effectiveness of social interactions (Saunderson and Nejat, 2022). Personalised interactions through AI-driven natural dialogue strategies can enhance user trust and acceptance, particularly in household and service settings, by collecting and adapting to individual preferences (Kraus *et al.*, 2022). These AI-driven personalisation methods are integral to ensuring that social robots are effective in performing their tasks and capable of creating meaningful and positive HRIs.

Considering previous research, zoomorphic anthropomorphic and feminised robots are more likeable and, thus, recommended for tasks requiring social qualities, where

machine-like designs are recommended for low-sociability tasks. In particular, zoomorphic designs may be an optimal choice when targeting an older user base. There is the crucial balance when designing social robots, to avoid too close of an imitation to human appearance and behaviours to avoid unease, instead promoting likeability and positive user engagement and experiences. Alongside the role of appearance, the behaviours of the robots could be tailored considering the user perceptions, engagement and experience. Founded on research exploring warmth and competence, a robot with behaviours reflective of warmth can significantly impact on user intentions compared to perceived capability and competence. In addition, allowing for personalisation, where robots are tailored to user needs and preferences can enhance social interaction, cooperation, acceptance and perceived effectiveness of robots. These recommendations may be considered simultaneously for various contexts. For example, in the context of older adults in care settings, a social robot reflective of a zoomorphic appearance, with a focus of warmth-based behavioural strategies, with user personalisation, tailoring to their needs, may promote reflecting acceptance, likeability, positive HRI and meaningful engagement.

A wide range of recommendations for the design of social robots are outlined above. Crucially, further research on the validity and ease of application of these recommendations will encourage the validity of these recommendations. Furthermore, to ensure widespread implementation of these recommendations, precise tools for assessments and established frameworks need to be designed. Notably, the ethical implications of trusting social robots are complex and multi-faceted. Trust in these systems can lead to beneficial outcomes, such as increased user engagement and the practical completion of tasks. However, this trust also introduces ethical challenges, particularly around the diffusion of responsibility. As robots become more autonomous and are perceived as partners rather than tools, there is a risk that users may begin to delegate too much responsibility to these machines, potentially leading to reduced accountability and the erosion of moral agency in human decision-making (Carli and Najjar, 2021). Furthermore, recent research emphasises that the concept of responsibility in robotics is not monolithic. Instead, it is shaped by different and sometimes conflicting ideas of responsibility. This diffusion of responsibility can create ethical dilemmas, particularly when robots are expected to act autonomously and make decisions that traditionally require human judgement. The complexity of assigning responsibility in these contexts underlines the need for robust ethical frameworks that address the potential for irresponsibility in robot design and deployment (Liu and Zawieska, 2020). Indeed, building trust in AI systems, including social robots, necessitates a foundation of ethical governance. This involves ensuring transparency and fairness and recognising and mitigating the risks associated with overtrust. Overtrust in robots can lead to users abdicating responsibility, which poses significant ethical challenges, particularly in scenarios where human oversight is crucial. Thus, fostering appropriate levels of trust without encouraging overreliance is essential (Winfield and Jirotka, 2018). Therefore, these recommendations for positive social robot design should be informed by and adhere to important ethical considerations highlighted throughout this paper. The following section will propose dimensions to evaluate and ensure the ethical design of social robots.

5.2 Ethical dimensions and risk mitigation measures

Discussing the ambivalence of building trust and responsibility in social robots seems paramount. The growing integration of social robots into various facets of daily life has brought about significant advancements in HRI. However, these advancements also raise complex ethical concerns, particularly surrounding the ambivalence of trust and the diffusion of responsibility. As social robots become more anthropomorphic or zoomorphic, they can

more easily elicit trust from users. While this trust can enhance engagement and improve task completion, it simultaneously introduces the risk of users over-relying on these machines, treating them as partners rather than tools. This shift in perception can lead to delegating critical responsibilities to robots, which may have detrimental social effects, such as reduced human agency and accountability.

Recent research highlights the dual-edged nature of trust in social robots. On one hand, trust is essential for the acceptance and effectiveness of robots in social settings. On the other hand, this trust must be carefully managed to avoid scenarios where users abdicate responsibility, leading to ethical dilemmas and potential harm (Carli and Najjar, 2021). The ethical ambivalence of trust necessitates a balanced approach in robot design and deployment, ensuring that trust does not erode human moral agency.

The diffusion of responsibility is another critical ethical issue arising from social robots' increasing autonomy. As robots present more sophisticated decision-making capabilities, there is a growing risk that users may defer their moral and ethical responsibilities to these machines. This phenomenon is particularly concerning in contexts where the robot's decisions carry significant consequences, such as in health-care or elder-care settings. The ambiguity in assigning responsibility in such scenarios can lead to a lack of accountability, undermining the ethical foundation of HRI. Studies have shown that the concept of responsibility in robotics is inherently fragmented and shaped by varying interpretations of what it means to be responsible in a socio-technical context. The challenge lies in ensuring that while robots may assist in decision-making, the ultimate responsibility remains with human users or operators (Liu and Zawieska, 2020). To address this, ethical frameworks must be developed that delineate the roles and responsibilities of humans and robots, ensuring that robots are used as tools to enhance human decision-making rather than replace it.

Finally, to mitigate the ethical risks associated with trust and responsibility diffusion, it is crucial to implement strategies that promote transparency, accountability and informed decision-making in HRI. For instance, the design of social robots should incorporate mechanisms that regularly remind users of the robot's limitations and the boundaries of its decision-making capabilities. In addition, robots should be designed to encourage shared decision-making processes, where the robot's role is clearly defined as supportive rather than directive (Henriksen *et al.*, 2021).

Moreover, there should be continuous monitoring and assessment of HRI to identify and address instances where responsibility may be inappropriately shifted to robots. This includes the development of accountability frameworks that hold developers, operators and users responsible for the actions and decisions made by robots under their control. Such frameworks are essential for maintaining ethical standards in deploying social robots and preventing the negative social impacts that can arise from the misuse of trust and the diffusion of responsibility.

Addressing specific ethical recommendations, the European Union approach outlines four fundamental ethical principles that should guide the development of trustworthy AI (HLEG, 2019). Firstly, systems must respect human autonomy by empowering users and ensuring oversight rather than manipulation. Secondly, they must aim to prevent harm and protect human well-being, dignity and safety. Thirdly, fairness is crucial – benefits and costs should be distributed justly without unacceptable bias or impacts on opportunities. Fourthly, explicability requires transparency about a system's capabilities and purpose while ensuring decisions can be explained and contested.

Here, we propose precise dimensions to evaluate when designing a social robot to ensure ethical design technology, building on those principles.

Firstly, the personalisation dimension aligns with the principles of autonomy and fairness. Allowing users meaningful choices in customising robots respects their preferences and autonomy. Inclusive design fosters equitable experiences. The corresponding dimensions are:

- Consider user preferences and offer customisation options with consent:
  - Allow customisation of the robot's appearance, name and responses based on user input (e.g. initiating interactions, greetings upon return, etc.).
- Understand diverse user needs through inclusive design:
  - Include a range of users from the design phase for ethical, social and cultural perspectives.
  - Develop visual/behaviour guidelines based on age and ability to mitigate deception risks.
  - Gather feedback and iteratively update personalisation based on user testing and ongoing experience.

Secondly, the dimension of transparency, conceptualised here as the information deployed by the robot, allows for the understanding of the robot's abilities and limitations. This directly links to respecting users' autonomy by empowering them with appropriate expectations. It also enables the explicability and contestability of robot decisions and behaviour. Moreover, addressing potential overtrust issues through informed data practices and privacy protections respects autonomy through valid consent. This fosters explicability, fairness and equitable treatment of personal information. The outlined dimensions are:

- Outline functional limitations upfront:
  - State what the robot can/cannot do, using simple, unambiguous terms catering to layperson users.
  - Repeat limitation disclosures regularly in interactions.
  - Describe response capabilities and biases transparently, such as known limitations in the training data (e.g. "I was trained to assist adults, I am not adequate to play with children").
- Tailor the robot to a supportive tool role:
  - Label it clearly as technology, not social relationship replacement, and have the robot express this.
  - Design interactions to enhance, rather than replace, human connections (e.g. facilitate calls with family members or social relations).
- Manage expectations of social–emotional responses:
  - Specifically note the lack of emotional or social abilities beyond programming.
  - Communicate context triggers for emotional expressions.
- Ensure well-informed consent for data practices:
  - Clearly explain what information is collected and how it is used.
  - Obtain explicit consent and provide layperson privacy controls (e.g. every six months, re-ask the user to select their privacy settings choices).
  - Answer users' questions about data to establish accountability and promote trust.

Finally, implementing a dimension of safeguards such as restrictions on emotional support or targeted training responses prevents potential harm from abuse or overreliance. This upholds human dignity, safety and care for vulnerable individuals. The dimensions are:

- Implement strict user restrictions to prevent harm:
  - Limit display of emotion/support to avoid dependency or separation distress.
  - Enforce maximum interaction periods considering various well-being indicators (e.g. based on age, cognitive abilities, feedback from the user, etc.).
  - Shut down the robot passively in response to physical/emotional abuse.
- Provide alternative outlets for needs:
  - Signpost human/community supports for high-risk users.
  - Direct users to counselling for recurring distress or harmful behaviours.
  - Equip robot responses to de-escalate stress and redirect to calm activities.
- Obtain oversight and feedback:
  - Consult experts in relevant fields, such as psychology or caregiving, throughout design and development processes and following implementation.
  - Continuously improve through transparent evaluation programs.
  - Rapidly disable any functionality proven to endanger well-being.
- Propose and develop non-gendered robots to reduce the risks associated with gender association and the reinforcement of gender stereotypes.

In summary, we propose a set of precise dimensions for the ethical design of social robots based on an analysis of the literature and grounded in the European principles of trustworthy AI. By carefully considering personalisation, transparency and safeguards from the early stages of ideation through the entire development process, designers can better respect human values like autonomy, well-being, fairness and accountability. The dimensions offer tangible yet flexible guidance for balancing benefits and risks to maximise social robot potential, while minimising harm stemming from the relationship. As evaluations continue and use cases expand in real-world contexts, ongoing refinement will ensure these proposals evolve supported by emerging evidence. Establishing a foundation attentive to psychological, social and ethical issues from the start can help deliver compassionate technologies that empower all members of society, consistent with the overarching vision of the approach to ethical by design AI.

5.3 Outlook and future research

As outlined throughout this review, an interdisciplinary approach is imperative when designing positive social robots to understand the endeavours and ethical design of HRIs. Crucially, although the review and recommendations are informed by findings from an abundance of research and crucial psychological theoretical frameworks, there are ongoing challenges within the field to be addressed. The final section of this paper will outline and offer recommendations for future research, such as the need for rigorous findings generated from longitudinal studies, the important development of tools and assessment of the psychological impact on users, and the important real-world impact and legal protections.

Advancing the field of HRI in an ethically grounded manner will require addressing several well-defined methodological and conceptual challenges. Precisely quantifying psychological impacts will necessitate validated measurement tools. Variables such as changes to cognitive performance, social skills acquisition and variations in subjective well-being must be reliably assessed through standardised instruments. Careful consideration of potential moderators like age, gender or baseline characteristics will also be important to comprehensively capture individual differences in outcomes. However, to truly discern

interaction dynamics and long-term effects, investigations should operationally define research designs capable of prospectively tracking partnerships over extended durations. For example, randomised controlled trials with longitudinal follow-ups in scheduled intervals could provide rigorous data on the stability and direction of relationship quality indicators across maturational phases. Moreover, to conceptually represent observed experiences, theoretical frameworks must be pragmatically reviewed and revised.

Multidisciplinary research teams should systematically develop taxonomies and models incorporating technical, learning-related and socio-relational constructs through iterative evaluation. Rigorous qualitative methods, such as focused ethnography or structured observational coding schemes, can provide in-depth user data, aiding conceptual refinement. Relatedly, socio-legal perspectives need to be integrated through policy pilot studies. Experimental variations in marketing messaging, risk disclosure formats or permissible use cases could inform regulatory proposals aimed at optimising benefits while preventing misuse. Overall, advancing the field in an evidence-based yet inclusive manner will require forging collaborations between technical, behavioural and social scientists. Their combined efforts, operating within a clearly defined, rigorous, mixed-methods investigative plan, offer the most promising approach for continuing to build knowledge responsibly.

Across these various avenues, future research will significantly benefit from a cross-disciplinary approach, intersecting the perspectives of HRI, psychology, education and law, to continuously evaluate the real-world impacts and, thus, guide the development of responsible social robot design founded on empirical evidence as the field rapidly progresses. Fostering a multidisciplinary approach, using expert oversight and feedback throughout the development of positive social robot design, ensures this is founded on rigorous research with an extensive scope.

6. Conclusion

In conclusion, this comprehensive review delved into the intersection of psychology and social robotics to inform the ethical design of robots. By examining various factors that influence HRIs and relationships, this study provided valuable insights for the development of social robots in an ethical manner. The evidence demonstrates that appearance, behaviour, emotional expression, personalisation and perceived autonomy are pivotal in fostering positive engagement and acceptance of social robots. Therefore, it is crucial to consider elements like gender cues, personalised memory, warmth expressions and inclusive design to cultivate beneficial relationships.

Although further research is still needed, these findings establish best practices for user-centred development approaches. Individuals' perceptions of robots in terms of attributes, such as social competence and warmth, directly impacts the formation of bonds over time. While anthropomorphism cues can foster empathy and attachments, it is important to acknowledge that robots cannot fully replicate human relationship dynamics. Consequently, it is necessary to consider non-reciprocal, guidance-based bonds instead of attempting to translate human frameworks directly to robots.

Studies indicate that para-social and attachment relations can form with social robots, but developers must carefully balance the opportunities and risks involved. Relying solely on cue-based bonds for emotional connection can lead to unrealistic expectations, manipulation or over-reliance, which may harm well-being. Therefore, it is essential to ensure that users understand the limitations of robots while designing them to provide personalised support, thus better serving the goals of the relationship.

Concerns regarding dehumanisation arise when robot relationships replace rather than complement valuable human connections. Negative impacts can arise if robot bonds

substitute for social engagement and fail to alleviate pressures such as isolation. In addition, worries about manipulation stem from inappropriate influence over decisions or the creation of unrealistic perceptions without transparency. Further research is necessary to validate these concerns and identify effective mitigation strategies. However, principles such as emphasising the role of robots as tools to enhance relationships rather than replace them, educating users, ensuring transparency of robot capabilities and personalising robots to the users' needs could help proactively address these critiques.

Overall, a balanced and evidence-guided approach that avoids premature bans but safeguards users appears to be the most viable path forward. While progress in social robotics undoubtedly brings benefits, it also entails responsibility. The conclusions drawn from this review underscore the importance of considering diverse perspectives, including technical, psychological and ethical lenses. By using frameworks that prioritise relationships grounded in mutual guidance and well-being, we can safely unlock social robotics' potential rewards as the field continues to mature through thoughtful and collaborative efforts.

References

Alemi, M., Taheri, A., Shariati, A. and Meghdari, A. (2020), "Social robotics, education, and religion in the Islamic world: an Iranian perspective", *Science and Engineering Ethics*, Vol. 26 No. 5, pp. 2709-2734, doi: 10.1007/s11948-020-00225-1.

Aliasghari, P., Ghafurian, M., Nehaniv, C. and Dautenhahn, K. (2021), "Effect of domestic trainee robots' errors on human teachers' trust", *2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 81-88.

Altman, I. and Taylor, D.A. (1973), *Social Penetration: The Development of Interpersonal Relationships*, Holt, Rinehart and Winston.

American Psychological Association (APA) (2023), "Balance theory", available at: https://dictionary.apa.org/balance-theory (accessed 27 September 2023).

Angel-Fernandez, J.M. and Vincze, M. (2018), "Towards a definition of educational robotics", Austrian Robotics Workshop 2018,Vol. 37.

Bandura, A. (1977), *Social Learning Theory*, Englewood Cliffs.

Bartneck, C., Nomura, T., Kanda, T., Suzuki, T. and Kato, K. (2005), "Cultural differences in attitudes towards robots", *AISB '05 - Robot Companions: hard Problems and Open Challenges in Robot-Human Interaction*, University of Hertfordshire, Hatfield.

Bartneck, C., Van der Hoek, M., Mubin, O. and Al Mahmud, A. (2007), "Daisy give me your answer do! Switching off a robot", ACM/IEEE Human Robot Interaction, pp. 217-222.

Bertolini, A. and Carli, R. (2022), "Human–robot interaction and user manipulation", *International Conference on Persuasive Technology, Springer International Publishing*, pp. 43-57.

Biswas, M. and Murray, J. (2014), "Effect of cognitive biases on human–robot interaction: a case study of a robot's misattribution", *The 23rd IEEE International Symposium on Robot and Human Interactive Communication, IEEE*, pp. 1024-1029.

Boch, A. (2011), *Culture is "Tight" with Technology Adoption: Cultural and Governance Factors Involved in the Acceptance of AI-Powered Surveillance Technology Deployed to Manage Covid-19*, Technical University of Munich.

Boch, A., Lucaj, L. and Corrigan, C. (2021), *A Robotic New Hope: Opportunities, Challenges, and Ethical Considerations of Social Robots, Technical University of Munich*, pp. 1-12.

Boch, A., Ryan, S., Kriebitz, A., Amugongo, L.M. and Lütge, C. (2023), "Beyond the metal flesh: understanding the intersection between bio-and AI ethics for robotics in healthcare", *Robotics*, Vol. 12 No. 4, p. 110.

Borenstein, J. and Arkin, R. (2019), "Robots, ethics, and intimacy: the need for scientific research", On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence: Themes from IACAP 2016, pp. 299-309.

Bowlby, J. (1982), "Attachment and loss: retrospect and prospect", *American Journal of Orthopsychiatry*, Vol. 52 No. 4, p. 664.

Bowlby, J. (1984), "Attachment and loss, 2: separation: anxiety and ager", *Apego e Perda, 2: Separação: Angústia e Raiva*, pp. 451-451.

Bowlby, J. (1998), "Attachment and loss, 3: sadness and depression", *Apego e Perda, 3: Perda, Tristeza e Depressão*, pp. 486-486.

Brave, S., Nass, C. and Hutchinson, K. (2005), "Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent", *International Journal of Human-Computer Studies*, Vol. 62 No. 2, pp. 161-178.

Breazeal, C. (2003), "Toward sociable robots", *Robotics and Autonomous Systems*, Vol. 42 Nos 3/4, pp. 167-175.

Burgoon, J.K. (1994), "Nonverbal signals", in Knapp, M. L. and Miller, G. R. (Eds), *Handbook of Interpersonal Communication*, 2nd ed., Sage, pp. 229-285.

Calo, R. (2015), "Robotics and the lessons of Cyberlaw", *California Law Review*, Vol. 103, pp. 513-563.

Calo, C.J., Hunt-Bull, N., Lewis, L. and Metzler, T. (2011), "Ethical implications of using the PARO robot, with a focus on dementia patient care", *Workshops at the twenty-fifth AAAI Conference on Artificial Intelligence*.

Cameron, D., de Saille, S., Collins, E.C., Aitken, J.M., Cheung, H., Chua, A., . . . Law, J. (2021), "The effect of social-cognitive recovery strategies on likability, capability and trust in social robots", *Computers in Human Behavior*, Vol. 114, p. 106561.

Carli, R. and Najjar, A. (2021), "Rethinking trust in social robotics", *ArXiv*.

Carpenter, J. (2013), "The quiet professional: an investigation of US military explosive ordnance disposal personnel interactions with everyday field robots", Doctoral Dissertation, University of Washington.

Casciaro, T. and Sousa-Lobo, M. (2005), "Competent jerks, lovable fools, and the formation of social networks", *Harvard Business Review*, Vol. 83 No. 6, pp. 92-99.

Churamani, N., Axelsson, M., Caldir, A. and Gunes, H. (2022), *Continual Learning for Affective Robotics: A Proof of Concept for Wellbeing, ArXiv*.

Cifuentes, C.A., Pinto, M.J., Céspedes, N. and Múnera, M. (2020), "Social robots in therapy and care", *Current Robotics Reports*, Vol. 1 No. 3, pp. 59-74.

Coeckelbergh, M., Pop, C., Simut, R., Peca, A., Pintea, S., David, D., *et al.* (2016), "A survey of expectations about the role of robots in robot-assisted therapy for children with ASD: Ethical acceptability, trust, sociability, appearance, and attachment", *Science and Engineering Ethics*, Vol. 22 No. 1, pp. 47-65, doi: 10.1007/s11948-015-9649-x.

Coghlan, S., Vetere, F., Waycott, J. and Barbosa Neves, B. (2019), "Could social robots make us kinder or crueller to humans and animals?", *International Journal of Social Robotics*, Vol. 11 No. 5, pp. 741-751, doi: 10.1007/s12369-019-00531-1.

Connor, R.A., Glick, P. and Fiske, S.T. (2017), "Ambivalent sexism in the twenty-first century", In Sibley, C.G. and Barlow, F.K. (Eds.), *The Cambridge Handbook of the Psychology of Prejudice*, Cambridge University Press, pp. 295-320, doi: 10.1017/9781316161579.013.

Danaher, J. (2019), "The philosophical case for robot friendship", *Journal of Posthuman Studies*, Vol. 3 No. 1, pp. 5-24, doi: 10.5325/jpoststud.3.1.0005.

Darling, K. (2015), "Who's Johnny?' anthropomorphic framing in human–robot interaction, integration, and policy", *Robot Ethics*, Vol. 2, pp. 173-191.

Darling, K. (2016), "Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects", *Robot Law*, Edward Elgar Publishing.

Darling, K., Nandy, P. and Breazeal, C. (2015), "Empathic concern and the effect of stories in human–robot interaction", *24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), IEEE*, pp. 770-775.

Dautenhahn, K. (2004), "Robots we like to live with?! a developmental perspective on a personalized, life-long robot companion", *2004 13th IEEE International Workshop on Robot and Human Interactive Communication, IEEE*, pp. 17-22.

Dautenhahn, K. (2007), "Socially intelligent robots: Dimensions of human–robot interaction", *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 362 No. 1480, pp. 679-704, doi: 10.1098/rstb.2006.2004.

de Graaf, M.M., Allouch, S.B. and van Dijk, J.A. (2016), "Long-term evaluation of a social robot in real homes", *Interaction Studies*, Vol. 17 No. 3, pp. 462-491.

de Graaf, M.M.A. and Allouch, S.B. (2017), "The influence of prior expectations of a robot's lifelikeness on users' intentions to treat a zoomorphic robot as a companion", *International Journal of Social Robotics*, Vol. 9, pp. 17-32.

de Graaf, M.M.A., Ben Allouch, S. and van Dijk, J.A.G.M. (2015), "What makes robots social?: A user's perspective on characteristics for social human–robot interaction", In Tapus, A., André, E., Martin, J-C., Ferland, F. and Ammi, M. (Eds.), *Social Robot*, Springer International Publishing, pp. 184-193, doi: 10.1007/978-3-319-25554-5_19.

de Mooij, M. and Hofstede, G. (2010), "The Hofstede model: applications to global branding and advertising strategy and research", *International Journal of Advertising*, Vol. 29 No. 1, pp. 85-110, doi: 10.2501/S026504870920104X.

Decety, J. (2015), "The neural pathways, development, and functions of empathy", *Current Opinion in Behavioral Sciences*, Vol. 3, pp. 1-6.

Dereshev, D., Kirk, D., Matsumura, K. and Maeda, T. (2019), "Long-term value of social robots through the eyes of expert users", *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery*, pp. 1-12, doi: 10.1145/3290605.3300896.

Duffy, B.R. (2003), "Anthropomorphism and the social robot", *Robotics and Autonomous Systems*, Vol. 42 Nos 3/4, pp. 177-190.

Duffy, B.R., Collier, R.W., O'Hare, G.M., Rooney, C.F.B. and O'Donoghue, R.P.S. (1999), "Social robotics: reality and virtuality in agent-based robotics", *Bar-Ilan Symposium on the Foundations of Artificial Intelligence: Bridging Theory and Practice (BISFAI)*.

Dziergwa, M., Kaczmarek, M., Kaczmarek, P., Kędzierski, J. and Wadas-Szydłowska, K. (2018), "Long-term cohabitation with a social robot: a case study of the influence of human attachment patterns", *International Journal of Social Robotics*, Vol. 10 No. 1, pp. 163-176, doi: 10.1007/s12369-017-0428-5.

Evans, D. (2010), "Wanting the impossible. The dilemma at the heart of intimate human-robot relationships", *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical, and Design Issues*, John Benjamins Publishing Company, pp. 75-88.

Eyssel, F. and Hegel, F. (2012), "(S)he's got the look: gender stereotyping of robots", *Journal of Applied Social Psychology*, Vol. 42 No. 9, pp. 2213-2230, doi: 10.1111/j.1559-1816.2012.00939.x.

Fiske, A., Henningsen, P. and 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*, Vol. 21 No. 5, p. e13216, doi: 10.2196/13216.

Fiske, S.T., Cuddy, A.J. and Glick, P. (2007), "Universal dimensions of social cognition: warmth and competence", *Trends in Cognitive Sciences*, Vol. 11 No. 2, pp. 77-83, doi: 10.1016/j.tics.2006.11.005.

Fleischmann, C., Cardon, P.W. and Aritz, J. (2020), "Smart collaboration in global virtual teams: the influence of culture on technology acceptance and communication effectiveness", HICSS, pp. 1-11.

Fong, T., Nourbakhsh, I. and Dautenhahn, K. (2003), "A survey of socially interactive robots", *Robotics and Autonomous Systems*, Vol. 42 Nos 3/4, pp. 143-166.

Fosch-Villaronga, E. and Poulsen, A. (2020), "Sex care robots: exploring the potential use of sexual robot technologies for disabled and elder care", *Paladyn, Journal of Behavioral Robotics*, Vol. 11 No. 1, pp. 1-18.

Fosch-Villaronga, E., Lutz, C. and Tamò-Larrieux, A. (2019), "Gathering expert opinions for social robots' ethical, legal, and societal concerns: findings from four international workshops", *International Journal of Social Robotics*, pp. 1-18, doi: 10.1007/s12369-019-00546-0.

Fox, J. and Gambino, A. (2021), "Relationship development with humanoid social robots: Applying interpersonal theories to human–robot interaction", *Cyberpsychology, Behavior, and Social Networking*, Vol. 24 No. 5, pp. 294-299, doi: 10.1089/cyber.2020.0587.

Freeman, S. (2016), "Sex robots to become a reality", The Toronto Star.

Fulmer, I.S., Barry, B. and Long, D.A. (2009), "Lying and smiling: informational and emotional deception in negotiation", *Journal of Business Ethics*, Vol. 88 No. 4, pp. 691-709, doi: 10.1007/s10551-008-9975-x.

García-Corretjer, M., Ros, R., Mallol, R. and Miralles, D. (2023), "Empathy as an engaging strategy in social robotics: a pilot study", *User Modeling and User-Adapted Interaction*, Vol. 33 No. 2, pp. 221-259, doi: 10.1007/s11257-023-09372-5.

Gelfand, M.J. and Kashima, Y. (2016), "Editorial overview: culture: advances in the science of culture and psychology", *Current Opinion in Psychology*, Vol. 8, pp. iv-ix, doi: 10.1016/j.copsyc.2015.12.011.

Gillath, O., Ai, T., Branicky, M.S., Keshmiri, S., Davison, R.B. and Spaulding, R. (2021), "Attachment and trust in artificial intelligence", *Computers in Human Behavior*, Vol. 115, p. 106607, doi: 10.1016/j.chb.2020.106607.

Glickson, E. and Woolley, A.W. (2020), "Human trust in artificial intelligence: Review of empirical research", *Academy of Management Annals*, Vol. 14 No. 2, pp. 627-660, doi: 10.5465/annals.2018.0121.

Goetz, J., Kiesler, S. and Powers, A. (2003), "Matching robot appearance and behavior to tasks to improve human-robot cooperation", *Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication, IEEE*, pp. 55-60.

Gutiu, S. (2016), "The roboticization of consent", in Calo, R., Froomkin, M. and Kerr, I. (Eds), *Robot Law*, Edward Elgar Publishing, pp. 186-212.

Haidt, J. (2001). "The emotional dog and its rational tail: a social intuitionist approach to moral judgment", *Psychological Review*, Vol. 108 No. 4, p. 814.

Hanoch, Y., Arvizzigno, F., Hernandez Garcia, D., Denham, S., Belpaeme, T. and Gummerum, M. (2021), "The robot made me do it: human–robot interaction and risk-taking behavior", *Cyberpsychology, Behavior, and Social Networking*, Vol. 24 No. 5, pp. 237-243, doi: 10.1089/cyber.2020.0148.

Henriksen, A., Enni, S. and Bechmann, A. (2021), "Situated accountability: Ethical principles, certification standards, and explanation methods in applied AI", *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, ACM*, pp. 106-112 .

Henschel, A., Laban, G. and Cross, E.S. (2021), "What makes a robot social? A review of social robots from science fiction to a home or hospital near you", *Current Robotics Reports*, Vol. 2 No. 1, pp. 9-19, doi: 10.1007/s43154-021-00036-9.

Hinz, N., Ciardo, F. and Wykowska, A. (2019), "Individual differences in attitude toward robots predict behavior in human–robot interaction", In *Proceedings of the 28th IEEE International Symposium*

on Robot and Human Interactive Communication, *IEEE*, pp. 64-73, doi: 10.1007/978-3-030-35888-4_7.

Ho, C.C., MacDorman, K.F. and Pramono, Z.D. (2008), "Human emotion and the uncanny valley: a GLM, MDS, and isomap analysis of robot video ratings", *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pp. 169-176.

Honig, S. and Oron-Gilad, T. (2018), "Understanding and resolving failures in human–robot interaction: literature review and model development", *Frontiers in Psychology*, Vol. 9, p. 861, doi: 10.3389/fpsyg.2018.00861.

Hou, Y.T.Y., Lee, W.Y. and Jung, M. (2023), "'Should I follow the human, or follow the robot?'—Robots in power can have more influence than humans on decision-making", *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1-13.

Huber, A., Weiss, A. and Rauhala, M. (2016), "The ethical risk of attachment: How to identify, investigate and predict potential ethical risks in the development of social companion robots", *2016 11th ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, *IEEE*, pp. 367-374, doi: 10.1109/HRI.2016.7451776

Hung, L., Liu, C., Woldum, E., Au-Yeung, A., Berndt, A., Wallsworth, C., *et al.* (2019), "The benefits of and barriers to using a social robot PARO in care settings: a scoping review", *BMC Geriatrics*, Vol. 19 No. 1, p. 232, doi: 10.1186/s12877-019-1244-6.

Irfan, B., Ramachandran, A., Staffa, M. and Gunes, H. (2023), "Lifelong learning and personalization in long-term human–robot interaction (LEAP-HRI): adaptivity for all", *Companion of the 2023 ACM/IEEE International Conference on Human–Robot Interaction*, pp. 1-4.

Kacancioğlu, E., Klug, H. and Alonzo, S.H. (2012), "The evolution of social interactions changes predictions about interacting phenotypes", *Evolution*, Vol. 66 No. 7, pp. 2056-2064, doi: 10.1111/j.1558-5646.2012.01585.x.

Kamide, H. and Mori, M. (2016), "One being for two origins – a necessary awakening for the future of robotics", *2016 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, *IEEE*, pp. 1-6.

Kaniarasu, P. and Steinfeld, A.M. (2014), "Effects of blame on trust in human–robot interaction", *2014 23rd IEEE International Symposium on Robot and Human Interactive Communication*, *IEEE*, RO-MAN, pp. 850-855.

Ke, C., Lou, V., Tan, K., Wai, M.Y. and Chan, L.L. (2020), "Changes in technology acceptance among older people with dementia: the role of social robot engagement", *International Journal of Medical Informatics*, Vol. 141, p. 104241, doi: 10.1016/j.ijmedinf.2020.104241.

Keefer, L.A., Landau, M.J., Rothschild, Z.K. and Sullivan, D. (2012), "Attachment to objects as compensation for close others' perceived unreliability", *Journal of Experimental Social Psychology*, Vol. 48 No. 4, pp. 912-917, doi: 10.1016/j.jesp.2012.02.007.

Kim, P.H., Dirks, K.T., Cooper, C.D. and Ferrin, D.L. (2006), "When more blame is better than less: the implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation", *Organizational Behavior and Human Decision Processes*, Vol. 99 No. 1, pp. 49-65, doi: 10.1016/j.obhdp.2005.07.002.

Knight, H. (2014), "How humans respond to robots: building public policy through good design", *Brookings Report*.

Kohlberg, L. (1971), "Stages of moral development as a basis for moral education", *Ethics guidelines for trustworthy AI. Shaping Europe's Digital Future*, Center for Moral Education, Harvard University: Cambridge, pp. 24-84, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F. and Kircher, T. (2008), "Can machines think? Interaction and perspective taking with robots investigated via fMRI", *PLoS ONE*, Vol. 3 No. 7, p. e2597, doi: 10.1371/journal.pone.0002597.

Krämer, N.C., Eimler, S.C., Von der Pütten, A.M. and Payr, S. (2011), "Theory of companions: What can theoretical models contribute to applications and understanding of human–robot interaction?", *Applied Artificial Intelligence*, Vol. 25 No. 6, pp. 474-502, doi: 10.1080/08839514.2011.587153.

Kraus, M., Dettenhofer, V. and Minker, W. (2022), "Responsible interactive personalization for human-robot cooperation", *Proceedings of the 30th ACM Conference on User Modeling, Adaptation, and Personalization*, pp. 1-4.

Lacey, C. and Caudwell, C. (2019), "Cuteness as a 'dark pattern' in home robots", *Proceedings of the 14th ACM/IEEE International Conference on Human–Robot Interaction (HRI 2019), IEEE*, pp. 374-381, doi: 10.1109/HRI.2019.8673195.

Lambert, A., Norouzi, N., Bruder, G. and Welch, G. (2020), "A systematic review of ten years of research on human interaction with social robots", *International Journal of Human–Computer Interaction*, Vol. 36 No. 19, pp. 1804-1817, doi: 10.1080/10447318.2020.1828539.

Lee, M.K., Forlizzi, J., Kiesler, S., Rybski, P., Antanitis, J. and Savetsila, S. (2012), "Personalization in HRI: a longitudinal field experiment", *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human–Robot Interaction (HRI 2012)*, pp. 319-326, doi: 10.1145/2157689.2157791.

Leite, I., Martinho, C. and Paiva, A. (2013), "Social robots for long-term interaction: a survey", *International Journal of Social Robotics*, Vol. 5 No. 2, pp. 291-308, doi: 10.1007/s12369-013-0178-y.

Levy, D. (2007), "Robot prostitutes as alternatives to human sex workers", *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2007). Roboethics.org*. www.roboethics.org/icra2007/contributions/LEVY%20Robot%20Prostitutes%20as%20Alternatives%20to%20Human%20Sex%20Workers.pdf

Li, D., Rau, P.P. and Li, Y. (2010), "A cross-cultural study: effect of robot appearance and task", *International Journal of Social Robotics*, Vol. 2 No. 2, pp. 175-186, doi: 10.1007/s12369-010-0056-9.

Lim, V., Rooksby, M. and Cross, E.S. (2021), "Social robots on a global stage: establishing a role for culture during human–robot interaction", *International Journal of Social Robotics*, Vol. 13 No. 6, pp. 1307-1333, doi: 10.1007/s12369-020-00722-y.

Liu, H. and Zawieska, K. (2020), "From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence", *Ethics and Information Technology*, Vol. 22 No. 4, pp. 1-14, doi: 10.1007/s10676-019-09500-0.

MacDorman, K.F., Vasudevan, S.K. and Ho, C.-C. (2009), "Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures", *AI and Society*, Vol. 23 No. 4, pp. 485-510, doi: 10.1007/s00146-008-0181-2.

Malle, B.F. (2016), "Integrating robot ethics and machine morality: the study and design of moral competence in robots", *Ethics and Information Technology*, Vol. 18 No. 4, pp. 243-256, doi: 10.1007/s10676-015-9367-8.

Marchesi, S., Bossi, F., Ghiglino, D., De Tommaso, D. and Wykowska, A. (2021), "I am looking for your mind: pupil dilation predicts individual differences in sensitivity to hints of human-likeness in robot behaviour", *Frontiers in Robotics and AI*, Vol. 8, p. 653537, doi: 10.1007/978-3-319-25554-5_19.

Markus, H.R. and Kitayama, S. (1991), "Culture and the self: implications for cognition, emotion, and motivation", *Psychological Review*, Vol. 98 No. 2, pp. 224-253, doi: 10.1037/0033-295X.98.2.224.

Matsumoto, D. (2006), "Culture and nonverbal behaviour", in Manusov, V. and Patterson, M.L. (Eds), *The Sage Handbook of Nonverbal Communication*, Sage Publications, pp. 219-235.

Metallo, C., Agrifoglio, R., Lepore, L. and Landriani, L. (2022), "Explaining users' technology acceptance through national cultural values in the hospital context", *BMC Health Services Research*, Vol. 22 No. 1, pp. 1-10, doi: 10.1186/s12913-022-07811-7.

Moon, H.S. and Seo, J. (2021), "Fast user adaptation for human motion prediction in physical human–robot interaction", *IEEE Robotics and Automation Letters*, Vol. 7 No. 1, pp. 120-127, doi: 10.1109/LRA.2021.3063956.

Naneva, S., Sarda Gou, M., Webb, T.L. and Prescott, T.J. (2020), "A systematic review of attitudes, anxiety, acceptance, and trust towards social robots", *International Journal of Social Robotics*, Vol. 12 No. 6, pp. 1179-1201, doi: 10.1007/s12369-020-00659-w.

Nass, C. and Brave, S. (2005), *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*, MIT Press.

Nass, C. and Moon, Y. (2000), "Machines and mindlessness: social responses to computers", *Journal of Social Issues*, Vol. 56 No. 1, pp. 81-103, doi: 10.1111/0022-4537.00153.

Nicolas, G., Bai, X. and Fiske, S.T. (2022), "A spontaneous stereotype content model: taxonomy, properties, and prediction", *Journal of Personality and Social Psychology*, Vol. 123 No. 6, p. 1243.

Norman, D.A. (2004), *Emotional Design: Why we Love (or Hate) Everyday Things*, Basic Civitas Books.

Norris, J.I., Lambert, N.M., DeWall, C.N. and Fincham, F.D. (2012), "Can't buy me love?: Anxious attachment and materialistic values", *Personality and Individual Differences*, Vol. 53 No. 5, pp. 666-669, doi: 10.1016/j.paid.2012.05.009.

Nyholm, S. and Frank, L. (2017), "From sex robots to love robots: is mutual love with a robot possible?", *Robot Sex: Social and Ethical Implications*, MIT Press, pp. 219-243.

Oyibo, K. and Vassileva, J. (2020), "HOMEX: Persuasive technology acceptance model and the moderating effect of culture", *Frontiers in Computer Science*, Vol. 2, p. 10, doi: 10.3389/fcomp.2020.00010.

Paepcke, S. and Takayama, L. (2010), "Judging a bot by its cover: an experiment on expectation setting for personal robots", *Proceedings of the 5th ACM/IEEE International Conference on Human–Robot Interaction (HRI 2010)*, pp. 45-52, doi: 10.1145/1734454.1734462

Pandey, A.K. and Gelin, R. (2018), "A mass-produced sociable humanoid robot: pepper: the first machine of its kind", *IEEE Robotics and Automation Magazine*, Vol. 25 No. 3, pp. 40-48, doi: 10.1109/MRA.2018.2833157.

Parke, P. (2015), "Is it cruel to kick a robot dog?", CNN, available at: http://edition.cnn.com/2015/02/13/tech/spot-robot-dog-google/

Perse, E.M. and Rubin, R.B. (1989), "Attribution in social and parasocial relationships", *Communication Research*, Vol. 16 No. 1, pp. 59-77, doi: 10.1177/009365089016001003.

Pinillos, R., Marcos, S., Feliz, R., Zalama, E. and Gómez-García-Bermejo, J. (2016), "Long-term assessment of a service robot in a hotel environment", *Robotics and Autonomous Systems*, Vol. 79, pp. 40-57, doi: 10.1016/j.robot.2016.01.014.

Pozharliev, R., De Angelis, M., Rossi, D., Romani, S., Verbeke, W. and Cherubino, P. (2021), "Attachment styles moderate customer responses to frontline service robots: Evidence from affective, attitudinal, and behavioral measures", *Psychology and Marketing*, Vol. 38 No. 5, pp. 881-895, doi: 10.1002/mar.21470.

Rabb, N., Law, T., Chita-Tegmark, M. and Scheutz, M. (2022), "An attachment framework for human–robot interaction", *International Journal of Social Robotics*, Vol. 14 No. 2, pp. 641-661, doi: 10.1007/s12369-021-00828-7.

Reig, S., Carter, E.J., Tan, X.Z., Steinfeld, A. and Forlizzi, J. (2021), "Perceptions of agent loyalty with ancillary users", *International Journal of Social Robotics*, Vol. 13 No. 8, pp. 1521-1537, doi: 10.1007/s12369-020-00730-6.

Richardson, K. (2015), *An Anthropology of Robots and AI: Annihilation Anxiety and Machines*, Routledge.

Richardson, K. (2016), "The asymmetrical 'relationship': Parallels between prostitution and the development of sex robots", *ACM SIGCAS Computers and Society*, Vol. 45 No. 3, pp. 290-293, doi: 10.1145/2874239.2874284.

Riek, L.D., Rabinowitch, T.-C., Chakrabarti, B. and Robinson, P. (2009), "How anthropomorphism affects empathy toward robots", *Proceedings of the 4th ACM/IEEE International Conference on Human–Robot Interaction, ACM*, pp. 245-246, doi: 10.1145/1514095.1514158.

Riva, G., Banos, R.M., Botella, C., Wiederhold, B.K. and Gaggioli, A. (2012), "Positive technology: using interactive technologies to promote positive functioning", *Cyberpsychology, Behavior, and Social Networking*, Vol. 15 No. 2, pp. 69-77, doi: 10.1089/cyber.2011.0139.

Rivoire, C. and Lim, A. (2016), "Habit detection within a long-term interaction with a social robot: an exploratory study", *Proceedings of the International Workshop on Social Learning Multimodal Interaction Design for Artificial Agents, ACM*, pp. 1-5, doi: 10.1145/3005338.3005342.

Robinette, P., Howard, A. and Wagner, A.R. (2017), "Effect of robot performance on human–robot trust in time-critical situations", *IEEE Transactions on Human-Machine Systems*, Vol. 47 No. 4, pp. 425-436, doi: 10.1109/THMS.2017.2648849.

Roloff, M.E. (1981), *Interpersonal Communication: The Social Exchange Approach*, Sage Publications.

Royakkers, L. and van Est, R. (2015), "A literature review on new robotics: Automation from love to war", *International Journal of Social Robotics*, Vol. 7 No. 5, pp. 549-570, doi: 10.1007/s12369-015-0295-x.

Sakamoto, D. and Ono, T. (2006), "Sociality of robots: Do robots construct or collapse human relations?", *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human–Robot Interaction (HRI 2006), ACM*, pp. 355-356, doi: 10.1145/1121241.1121319.

Saunderson, S. and Nejat, G. (2022), "Hybrid hierarchical learning for adaptive persuasion in human–robot interaction", *IEEE Robotics and Automation Letters*, Vol. 7 No. 2, pp. 5520-5527, doi: 10.1109/LRA.2022.3182744.

Scheutz, M. (2012), "The inherent dangers of unidirectional emotional bonds between humans and social robots", in Lin, P., Abney, K. and Bekey, G. (Eds), *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, pp. 205-221.

Schiappa, E., Allen, M. and Gregg, P.B. (2007), "Parasocial relationships and television: a meta-analysis of the effects", *In Mass Media Effects Research: Advances through Meta-Analysis*, pp. 301-314, doi: 10.1111/j.1083-6101.1999.tb00354.x.

Schönmann, M., Bodenschatz, A., Uhl, M. and Walkowitz, G. (2024), "Contagious humans: a pandemic's positive effect on attitudes towards care robots", *Technology in Society*, Vol. 76, p. 102464, doi: 10.1016/j.techsoc.2024.102464.

Shamsuddin, S., Yussof, H., Ismail, L., Hanapiah, F.A., Mohamed, S., Piah, H.A. and Zahari, N.I. (2012), "Initial response of autistic children in human–robot interaction therapy with humanoid robot NAO", *Proceedings of the 2012 IEEE 8th International Colloquiumon Signal Processing and its Applications (CSPA 2012), IEEE*, pp. 188-193, doi: 10.1109/CSPA.2012.6194703.

Sharkey, N. and Sharkey, A. (2010), "The crying shame of robot nannies: an ethical appraisal", *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, Vol. 11 No. 2, pp. 161-190, doi: 10.1075/is.11.2.01sha.

Sharkey, A. and Sharkey, N. (2011), "Children, the elderly, and interactive robots", *IEEE Robotics and Automation Magazine*, Vol. 18 No. 1, pp. 32-38, doi: 10.1109/MRA.2010.940151.

Sharkey, A. and Sharkey, N. (2012), "Granny and the robots: Ethical issues in robot care for the elderly", *Ethics and Information Technology*, Vol. 14 No. 1, pp. 27-40, doi: 10.1007/s10676-010-9234-6.

Shaver, P.R., Schachner, D.A. and Mikulincer, M. (2005), "Attachment style, excessive reassurance seeking, relationship processes, and depression", *Personality and Social Psychology Bulletin*, Vol. 31 No. 3, pp. 343-359, doi: 10.1177/0146167204271709.

Shazi, R., Gillespie, N. and Steen, J. (2015), "Trust as a predictor of innovation network ties in project teams", *International Journal of Project Management*, Vol. 33 No. 1, pp. 81-91, doi: 10.1016/j.ijproman.2014.06.001.

Smith, E.R., Šabanović, S. and Fraune, M.R. (2021), "Human–robot interaction through the lens of social psychological theories of intergroup behavior", *Technology, Mind, and Behavior*, Vol. 1 No. 2, pp. 1-11, doi: 10.1037/tmb0000002.

Sullins, J.P. (2012), "Robots, love, and sex: the ethics of building a love machine", *IEEE Transactions on Affective Computing*, Vol. 3 No. 4, pp. 398-409, doi: 10.1109/T-AFFC.2012.31.

Sung, J.-Y., Guo, L., Grinter, R.E. and Christensen, H.I. (2007), "'My Roomba is Rambo': intimate home appliances", *Proceedings of the International Conference on Ubiquitous Computing (UbiComp 2007)*, pp. 145-162, doi: 10.1007/978-3-540-74853-3_9.

Tanaka, F., Isshiki, K., Takahashi, F., Uekusa, M., Sei, R. and Hayashi, K. (2015), "Pepper learns together with children: development of an educational application", *Proceedings of the 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids 2015)*, *IEEE*, pp. 270-275, doi: 10.1109/HUMANOIDS.2015.7363546.

Tousignant, B., Eugène, F. and Jackson, P.L. (2017), "A developmental perspective on the neural bases of human empathy", *Infant Behavior and Development*, Vol. 48, pp. 5-12, doi: 10.1016/j.infbeh.2016.11.002.

Trovato, G., Kishi, T., Endo, N., *et al.* (2013), "Cross-cultural perspectives on emotion expressive humanoid robotic head: recognition of facial expressions and symbols", *International Journal of Social Robotics*, Vol. 5 No. 4, pp. 515-527, doi: 10.1007/s12369-013-0213-z.

Turing, A. (1950), "Computing machinery and intelligence", *Mind*, Vol. LIX No. 236, pp. 433-460, doi: 10.1093/mind/LIX.236.433.

Turkle, S. (2007), "Authenticity in the age of digital companions", *Interaction Studies*, Vol. 8 No. 3, pp. 501-517.

Turkle, S. (2010), "In good company?", in Wilks, Y. (Ed.), *Close Engagements with Artificial Companions*, John Benjamins Publishing Company, pp. 3-10.

Turkle, S. (2011), *Alone Together: Why we Expect More from Technology and Less from Each Other*, Basic Books.

Turkle, S. (2012), *Alone Together: Why we Expect More from Technology and Less from Each Other*, Basic Books.

Turner, J.C. (1978), "Social comparison similarity and intergroup favouritism", in Tajfel, H. (Ed.), *Differentiation between Social Groups*, Academic Press.

Vallor, S. (2011), "Knowing what to wish for: human enhancement technology, dignity and virtue. Techne", *Research in Philosophy and Technology*, Vol. 15 No. 2, pp. 115-137, doi: 10.5840/techne201115215.

Van Doorn, J., Mende, M., Noble, S.M., Hulland, J., Ostrom, A.L., Grewal, D. and Petersen, J.A. (2017), "Domo arigato Mr. Roboto: emergence of automated social presence in organizational frontlines and customers' service experiences", *Journal of Service Research*, Vol. 20 No. 1, pp. 43-58, doi: 10.1177/1094670516679272.

Van Maris, A., Zook, N., Caleb-Solly, P., Studley, M., Winfield, A. and Dogramadzi, S. (2020), "Designing ethical social robots – a longitudinal field study with older adults", *Frontiers in Robotics and AI*, Vol. 7, pp. 1-20, doi: 10.3389/frobt.2020.00001.

Walk, H. (2016), "Amazon echo is magical", it's also turning my kid into an asshole. LinkedIn, available at: https://www.linkedin.com/pulse/amazon-echo-magical-its-also-turning-my-kid-asshole-hunter-walk/

Winfield, A. and Jirotka, M. (2018), "Ethical governance is essential to building trust in robotics and artificial intelligence systems", *Philosophical Transactions of the Royal Society A:*

Mathematical, Physical and Engineering Sciences, Vol. 376 No. 2133, pp. 1-20, doi: 10.1098/rsta.2018.0085.

Yamada, S., Kanda, T. and Tomita, K. (2023), "Process of escalating robot abuse in children", *International Journal of Social Robotics*, Vol. 15 No. 5, pp. 835-853.

Further reading

Abdi, J., Al-Hindawi, A., Ng, T. and Vizcaychipi, M.P. (2018), "Scoping review on the use of socially assistive robot technology in elderly care", *BMJ Open*, Vol. 8 No. 2, p. e018815, doi: 10.1136/bmjopen-2017-018815.

Abdollahi, H., Mollahosseini, A., Lane, J.T. and Mahoor, M.H. (2017), "A pilot study on using an intelligent life-like robot as a companion for elderly individuals with dementia and depression", *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids), IEEE*, pp. 541-546.

Abe, K., Hieida, C., Attamimi, M., Nagai, T., Shimotomai, T., Omori, T. and Oka, N. (2014), "Toward playmate robots that can play with children considering personality", *Proceedings of the second international conference on Human-agent interaction*, pp. 165-168.

Alemi, M., Ghanbarzadeh, A., Meghdari, A. and Moghadam, L.J. (2016), "Clinical application of a humanoid robot in Pediatric cancer interventions", *International Journal of Social Robotics*, Vol. 8 No. 5, pp. 743-759.

Alemi, M., Meghdari, A. and Haeri, N.S. (2017), "Young EFL learners' attitude towards RALL: an observational study focusing on motivation, anxiety, and interaction", *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9, Springer International Publishing*, pp. 252-261.

Alnajjar, F., Khalid, S., Vogan, A.A., Shimoda, S., Nouchi, R. and Kawashima, R. (2019), "Emerging cognitive intervention technologies to meet the needs of an aging population: a systematic review", *Frontiers in Aging Neuroscience*, Vol. 11, p. 291, doi: 10.3389/fnagi.2019.00291.

Alves-Oliveira, P., Tullio, E.D., Ribeiro, T. and Paiva, A. (2014), "Meet me halfway: Eye behaviour as an expression of robot's language", *AAAI Fall Symposium Series*, pp. 13-15.

Assad-Uz-Zaman, M., Rasedul Islam, M., Miah, S. and Rahman, M.H. (2019), "NAO robot for cooperative rehabilitation training", *Journal of Rehabilitation and Assistive Technologies Engineering*, Vol. 6, p. 2055668319862151, doi: 10.1177/2055668319862151.

Atkinson, R.K., Mayer, R.E. and Merrill, M.M. (2005), "Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice", *Contemporary Educational Psychology*, Vol. 30 No. 1, pp. 117-139.

Baxter, P., Ashurst, E., Read, R., Kennedy, J. and Belpaeme, T. (2017), "Robot education peers in a situated primary school study: personalisation promotes child learning", *Plos One*, Vol. 12 No. 5, p. e0178126.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B. and Tanaka, F. (2018), "Social robots for education: a review", *Science Robotics*, Vol. 3 No. 21, p. eaat5954, doi: 10.1126/scirobotics.aat5954.

Bemelmans, R., Gelderblom, G.J., Jonker, P. and de Witte, L. (2015), "Effectiveness of robot Paro in intramural psychogeriatric care: a multicenter quasi-experimental study", *Journal of the American Medical Directors Association*, Vol. 16 No. 11, pp. 946-950.

Blue Frog Robotics (2022), "L'éthique de la robotique sociale", available at: https://www.bluefrogrobotics.com/Uploads/Docs/LIVRE_BLANC_2022.pdf

Boucher, J.D., Pattacini, U., Lelong, A., Bailly, G., Elisei, F., Fagel, S., . . . Ventre-Dominey, J. (2012), "Reach faster when I see you look: gaze effects in human–human and human–robot face-to-face cooperation", *Frontiers in Neurorobotics*, Vol. 6 No. 3, pp. 1-11.

Bowlby, J. (1969), *Attachment and Loss: attachment*, Basic Books.

Bowlby, J. (1972), *Attachment and Loss: separation: anxiety and Anger*, Basic Books.

Bowlby, J. (1980), *Attachment and Loss: loss, Sadness and Depression*, Basic Books.

Breazeal, C. (2011), "Social robots for health applications", *2011 Annual international conference of the IEEE engineering in medicine and biology society, IEEE*, pp. 5368-5371.

Broekens, J., Heerink, M. and Rosendal, H. (2009), "Assistive social robots in elderly care: a review", *Gerontechnology*, Vol. 8 No. 2, pp. 94-103.

Byrne, S., Gay, G., Pollack, J.P., Gonzales, A., Retelny, D., Lee, T. and Wansink, B. (2012), "Caring for mobile phone-based virtual pets can influence youth eating behaviors", *Journal of Children and Media*, Vol. 6 No. 1, pp. 83-99.

Capgeris (2018), "Rapport Paro - Capgeris", available at: www.capgeris.com/docs/pu/1/rapport-utilisation-paro-en-ehpad.pdf

Carrillo, F., Butchart, J., Knight, S., Scheinberg, A., Wise, L., Sterling, L. and McCarthy, C. (2017), "In-situ design and development of a socially assistive robot for Pediatric rehabilitation", *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human–Robot Interaction*, pp. 199-200.

Catlin, D., Kandlhofer, M., Holmquist, S., Csizmadia, A.P., Angel-Fernandez, J. and Cabibihan, J.J. (2018), "EduRobot taxonomy and Papert's paradigm", in Dagiene, V. and Jasute, E. (Eds), *Constructionism 2018: Constructionism, Computational Thinking and Educational Innovation*, Vilnius, Lithuania, pp. 151-159.

Cespedes, N., Munera, M., Gomez, C. and Cifuentes, C.A. (2020), "Social human–robot interaction for gait rehabilitation", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 28 No. 6, p. 1307, doi: 10.1109/tnsre.2020.2987428.

Chamorro-Premuzic, T. and Furhnam, A. (2006), "Intellectual competence and the intelligent personality: a third way in differential psychology", *Review of General Psychology*, Vol. 10 No. 3, pp. 251-267, doi: 10.1037/1089-2680.10.3.251.

Chen, Y., Garcia-Vergara, S. and Howard, A.M. (2018), "Effect of feedback from a socially interactive humanoid robot on reaching kinematics in children with and without cerebral palsy: a pilot study", *Developmental Neurorehabilitation*, Vol. 21 No. 8, pp. 490-496, doi: 10.1080/17518423.2017.1360962.

Coleman, W.L. (2008), "Social competence and friendship formation in adolescents with attention-deficit/hyperactivity disorder", *Adolescent Medicine: State of the Art Reviews*, Vol. 19 No. 2, pp. 278-299.

da Silva, J., Kavanagh, D.J., Belpaeme, T., Taylor, L., Beeson, K. and Andrade, J. (2018), "Experiences of a motivational interview delivered by a robot: qualitative study", *Journal of Medical Internet Research*, Vol. 20 No. 5, p. e116, doi: 10.2196/jmir.7737.

Darling, K. (2021), *The New Breed: What Our History with Animals Reveals about Our Future with Robots*, Henry Holt and Company.

David, C., Fernando, S., Collins, E., Millings, A., Moore, R., Sharkey, A., . . . Prescott, T. (2015), "Presence of life-like robot expressions influences children's enjoyment of human–robot interactions in the field", New Frontiers, European Union Seventh Framework Programme (FP7-ICT-2013-10) (Grant Agreement No. 611971).

Dawe, J., Sutherland, C., Barco, A. and Broadbent, E. (2019), "Can social robots help children in healthcare contexts? A scoping review", *BMJ Paediatrics Open*, Vol. 3 No. 1, p. e000371, doi: 10.1136/bmjpo-2018-000371.

Esposito, J. (2011), "Negotiating the gaze and learning the hidden curriculum: a critical race analysis of the embodiment of female students of color at a predominantly white institution", *Journal for Critical Education Policy Studies*, Vol. 9 No. 2, pp. 143-164.

Fasola, J. and Mataric, M.J. (2013), "A socially assistive robot exercise coach for the elderly", *Journal of Human–Robot Interaction*, Vol. 2 No. 2, pp. 1-32.

Feil-Seifer, D. and Mataric, M.J. (2005), "Defining socially assistive robotics", *Proceedings of the 9th International Conference on Rehabilitation Robotics*, pp. 465-468.

Fridin, M. (2014), "Storytelling by a kindergarten social assistive robot: a tool for constructive learning in preschool education", *Computers and Education*, Vol. 70, pp. 53-64.

Geva, N., Uzefovsky, F. and Levy-Tzedek, S. (2020), "Touching the social robot PARO reduces pain perception and salivary oxytocin levels", *Scientific Reports*, Vol. 10 No. 1, p. 9814, doi: 10.1038/s41598-020-66982-y.

Góngora Alonso, S., Hamrioui, S., de la Torre, D.I., Motta Cruz, E., López-Coronado, M. and Franco, M. (2018), "Social robots for people with aging and dementia: a systematic review of literature", *Telemedicine and e-Health*, Vol. 25 No. 7, pp. 533-540, doi: 10.1089/tmj.2018.0051.

Han, J.-H., Jo, M.-H., Jones, V. and Jo, J.-H. (2008), "Comparative study on the educational use of home robots for children", *Journal of Information Processing Systems*, Vol. 4 No. 4, pp. 159-168, doi: 10.3745/JIPS.2008.4.4.159.

Hattie, J. (2009), *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement*, Routledge.

Heider, F. (1958), *The Psychology of Interpersonal Relations*, Wiley.

Henkemans, O.A.B., Bierman, B.P., Janssen, J., Looije, R., Neerincx, M.A., van Dooren, M.M. and Huisman, S.D. (2017), "Design and evaluation of a personal robot playing a self-management education game with children with diabetes type 1", *International Journal of Human-Computer Studies*, Vol. 106, pp. 63-76, doi: 10.1016/j.ijhcs.2017.05.003.

Hinds, P.J., Roberts, T.L. and Jones, H. (2004), "Whose job is it anyway? A study of human–robot interaction in a collaborative task", *Human-Computer Interaction*, Vol. 19 No. 1, pp. 151-181, doi: 10.1080/07370024.2004.9667343.

Hofstede, G. (1980), *Culture's Consequences: International Differences in Work-Related Values*, Sage.

Huang, I.S. and Hoorn, J.F. (2019), Having an Einstein in class: Teaching maths with robots is different for boys and girls, in Wang, X., Wang, Z., Wu, J. and Wang, L. (Eds), *Proceedings of the 13th World Congress on Intelligent Control and Automation (WCICA 2018), July 4-8, Changsha, China, IEEE*, pp. 424-427, doi: 10.1109/WCICA.2018.8630584.

Jamet, F., Masson, O., Jacquet, B., Stilgenbauer, J.-L. and Baratgin, J. (2018), "Learning by teaching with humanoid robot: a new powerful experimental tool to improve children's learning ability", *Journal of Robotics*, Vol. 2018, pp. 1-11, doi: 10.1155/2018/4578762.

Janssen, J.B., Van der Wal, C.C., Neerincx, M.A. and Looije, R. (2011), "Motivating children to learn arithmetic with an adaptive robot game", in Mutlu, B., Bartneck, C., Ham, J., Evers, V. and Kanda, T. (Eds), *Social Robotics. Lecture Notes in Computer Science*, Springer, Vol. 7072, pp. 153-160, doi: 10.1007/978-3-642-25504-5_16.

Jeong, S., Breazeal, C., Logan, D.E. and Weinstock, P. (2017), "Huggable: Impact of embodiment on promoting verbal and physical engagement for young Pediatric inpatients", *Proceedings of the 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2017), IEEE*, pp. 121-126, doi: 10.1109/ROMAN.2017.8172316.

Jeong, S., Logan, D.E., Goodwin, M.S., O'Connell, B. and Weinstock, P. (2015), "A social robot to mitigate stress, anxiety, and pain in hospital pediatric care", *Proceedings of the Tenth ACM/IEEE International Conference on Human–Robot Interaction Extended Abstracts (HRI 2015), ACM*, pp. 103-104, doi: 10.1145/2701973.2701983.

Johal, W. (2020), "Research trends in social robots for learning", *Current Robotics Reports*, Vol. 1 No. 3, pp. 75-83, doi: 10.1007/s43154-020-00009-6.

Johnsen, K., Ahn, S.J., Moore, J., Brown, S., Robertson, T.P., Marable, A. and Basu, A. (2014), "Mixed reality virtual pets to reduce childhood obesity", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 20 No. 4, pp. 523-530, doi: 10.1109/TVCG.2014.45.

Johnson, W.L. and Lester, J.C. (2016), "Face-to-face interaction with pedagogical agents, twenty years later", *International Journal of Artificial Intelligence in Education*, Vol. 26 No. 1, pp. 25-36, doi: 10.1007/s40593-015-0065-9.

Jones, A., Castellano, G. and Bull, S. (2014), "Investigating the effect of a robotic tutor on learner perception of skill-based feedback", Social Robotics: 6th International Conference, ICSR 2014, Sydney, NSW, Australia, October 27-29, 2014, Proceedings 6, *Springer International Publishing*, pp. 186-195.

Jøranson, N., Pedersen, I., Rokstad, A.M.M. and Ihlebæk, C. (2015), "Effects on symptoms of agitation and depression in persons with dementia participating in robot-assisted activity: a cluster-randomized controlled trial", *Journal of the American Medical Directors Association*, Vol. 16 No. 10, pp. 867-873, doi: 10.1016/j.jamda.2015.05.002.

Jøranson, N., Pedersen, I., Rokstad, A.M.M., Aamodt, G., Olsen, C. and Ihlebæk, C. (2016), "Group activity with Paro in nursing homes: Systematic investigation of behaviors in participants", *International Psychogeriatrics*, Vol. 28 No. 8, pp. 1345-1354, doi: 10.1017/S104161021600057X.

Kacancioğlu, E., Klug, H. and Alonzo, S.H. (2012), "The evolution of social interactions changes predictions about interacting phenotypes", *Evolution*, Vol. 66 No. 7, pp. 2056-2064, doi: 10.1111/j.1558-5646.2012.01585.x.

Kant, I. (1784–1785/1997), "Moral philosophy: Collins' lecture notes", in Heath, P. and Schneewind, J. B. (Eds), *Lectures on Ethics*, Cambridge University Press.

Kennedy, J., Baxter, P. and Belpaeme, T. (2015), "The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning", *Proceedings of the tenth annual ACM/IEEE International Conference on Human–Robot Interaction*, *ACM*, pp. 67-74, doi: 10.1145/2696454.2696457.

Kidd, C. and Breazeal, C. (2008), Robots at home: Understanding long-term human–robot interaction. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2008)*, *IEEE*, pp. 3644-3651, doi: 10.1109/IROS.2008.4650967.

Kim, P.H., Ferrin, D.L., Cooper, C.D. and Dirks, K.T. (2004), "Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations", *Journal of Applied Psychology*, Vol. 89 No. 1, pp. 104-118, doi: 10.1037/0021-9010.89.1.104.

Klein, B., Gaedt, L. and Cook, G. (2013), "Emotional robots", *GeroPsych*, Vol. 26 No. 2, pp. 89-99, doi: 10.1024/1662-9647/a000087.

Konijn, E.A. and Hoorn, J.F. (2020), "Robot tutor and pupils' educational ability: Teaching the times tables", *Computers and Education*, Vol. 157, p. 103970, doi: 10.1016/j.compedu.2020.103970.

Konijn, E.A., Smakman, M. and van den Berghe, R. (2020), "Use of robots in education", *The International Encyclopedia of Media Psychology*, pp. 1-8, doi: 10.1002/9781119011071.iemp0060.

Kory, J. and Breazeal, C. (2014), "Storytelling with robots: Learning companions for preschool children's language development", *2014 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2014)*, *IEEE*, pp. 643-648, doi: 10.1109/ROMAN.2014.6926348.

Kyong, I.K., Freedman, S., Mataric, M., Cunningham, M. and Lopez, B. (2005), "A hands-off physical therapy assistance robot for cardiac patients", *Proceedings of the 9th International Conference on Rehabilitation Robotics (ICORR 2005)*, *IEEE*, pp. 465-468, doi: 10.1109/ICORR.2005.1501142.

Laban, G., George, J.-N., Morrison, V. and Cross, E.S. (2021), "Tell me more! Assessing interactions with social robots from speech", *Paladyn, Journal of Behavioral Robotics*, Vol. 12 No. 1, pp. 136-159, doi: 10.1515/pjbr-2021-0011.

Laban, G., Kappas, A., Morrison, V. and Cross, E.S. (2023), "Human-robot relationships: Long-term effects on disclosure, perception, and well-being", *Frontiers in Psychology*, Vol. 14, p. 1100707, doi: 10.3389/fpsyg.2023.1100707.

Lane, G.W., Noronha, D., Rivera, A., Craig, K., Yee, C. and Mills, B. (2016), "Effectiveness of a social robot, 'paro', in a VA long-term care setting", *Psychological Services*, Vol. 13 No. 3, pp. 292-299, doi: 10.1037/ser0000100.

Lemaignan, S., Jacq, A., Hood, D., Garcia, F., Paiva, A. and Dillenbourg, P. (2016), "Learning by teaching a robot: the case of handwriting", *IEEE Robotics and Automation Magazine*, Vol. 23 No. 2, pp. 56-66, doi: 10.1109/MRA.2016.2546700.

Leyzberg, D., Spaulding, S. and Scassellati, B. (2014), "Personalizing robot tutors to individuals' learning differences", *Proceedings of the 2014 ACM/IEEE International Conference on Human–Robot Interaction (HRI 2014)*, pp. 423-430, doi: 10.1145/2559636.2559671.

Liang, A., Piroth, I., Robinson, H., MacDonald, B., Fisher, M., Nater, U.M., *et al.* (2017), "A pilot randomized trial of a companion robot for people with dementia living in the community", *Journal of the American Medical Directors Association*, Vol. 18 No. 10, pp. 871-878, doi: 10.1016/j.jamda.2017.05.019.

Logan, D.E., Breazeal, C., Goodwin, M.S., Jeong, S., O'Connell, B., Smith-Freedman, D., *et al.* (2019), "Social robots for hospitalized children", *Pediatrics*, Vol. 144 No. 1, pp. 1-10, doi: 10.1542/peds.2018-1511.

Looije, R., Neerincx, M.A., Peters, J.K. and Henkemans, O.A.B. (2016), "Integrating robot support functions into varied activities at returning hospital visits: Supporting child's self-management of diabetes", *International Journal of Social Robotics*, Vol. 8 No. 4, pp. 483-497, doi: 10.1007/s12369-016-0370-3.

Macedonia, M., Müller, K. and Friederici, A.D. (2011), "The impact of iconic gestures on foreign language word learning and its neural substrate", *Human Brain Mapping*, Vol. 32 No. 6, pp. 982-998, doi: 10.1002/hbm.21084.

Mann, J.A., MacDonald, B.A., Kuo, I., Li, X. and Broadbent, E. (2015), "People respond better to robots than computer tablets delivering healthcare instructions", *Computers in Human Behavior*, Vol. 43, pp. 112-117, doi: 10.1016/j.chb.2014.10.032.

Mervin, M.C., Moyle, W., Jones, C., Murfield, J., Draper, B., Beattie, E., *et al.* (2018), "The cost-effectiveness of using PARO, a therapeutic robotic seal, to reduce agitation and medication use in dementia: findings from a cluster-randomized controlled trial", *Journal of the American Medical Directors Association*, Vol. 19 No. 7, pp. 619-622, doi: 10.1016/j.jamda.2018.03.018.

Mohebbi, A. (2020), "Human–robot interaction in rehabilitation and assistance: a review", *Current Robotics Reports*, Vol. 1 No. 3, pp. 131-144, doi: 10.1007/s43154-020-00015-4.

Mori, M. (1970), "The uncanny valley: the original essay by Masahiro Mori", *IEEE Spectrum*, Vol. 6, doi: 10.1109/MSPEC.2012.6348159.

Moyle, W., Bramble, M., Jones, C. and Murfield, J. (2016), "Care staff perceptions of a social robot called Paro and a look-alike plush toy: a descriptive qualitative approach", *Aging and Mental Health*, Vol. 22 No. 3, pp. 330-335, doi: 10.1080/13607863.2016.1222820.

Moyle, W., Cooke, M., Beattie, E., Jones, C., Klein, B., Cook, G., *et al.* (2013), "Exploring the effect of companion robots on emotional expression in older adults with dementia: a pilot randomized controlled trial", *Journal of Gerontological Nursing*, Vol. 39 No. 5, pp. 46-53, doi: 10.3928/00989134-20130313-03.

Moyle, W., Jones, C.J., Murfield, J.E., Thalib, L., Beattie, E.R.R.A., Shum, D.K.K.H., *et al.* (2017), "Use of a robotic seal as a therapeutic tool to improve dementia symptoms: a cluster-randomized controlled trial", *Journal of the American Medical Directors Association*, Vol. 18 No. 9, pp. 766-773, doi: 10.1016/j.jamda.2017.03.018.

Nomura, T., Kanda, T. and Suzuki, T. (2006), "Experimental investigation into the influence of negative attitudes toward robots on human–robot interaction", *AI and Society*, Vol. 20 No. 2, pp. 138-150, doi: 10.1007/s00146-005-0012-7.

Petersen, S., Houston, S., Qin, H., Tague, C. and Studley, J. (2017), "The utilization of robotic pets in dementia care", *Journal of Alzheimer's Disease*, Vol. 55 No. 2, pp. 569-574, doi: 10.3233/JAD-160703.

Pino, M., Boulay, M., Jouen, F. and Rigaud, A.S. (2015), "Are we ready for robots that care for us? Attitudes and opinions of older adults toward socially assistive robots", *Frontiers in Aging Neuroscience*, Vol. 7, p. 141, doi: 10.3389/fnagi.2015.00141.

Rabbitt, S.M., Kazdin, A.E. and Scassellati, B. (2015), "Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use", *Clinical Psychology Review*, Vol. 35, pp. 35-46, doi: 10.1016/j.cpr.2014.07.001.

Ramachandran, A. and Scassellati, B. (2015a), "Fostering learning gains through personalized robot-child tutoring interactions", *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human–Robot Interaction Extended Abstracts*, *ACM*, pp. 193-194, doi: 10.1145/2701973.2701985.

Ramachandran, A. and Scassellati, B. (2015b), "Developing adaptive social robot tutors for children", *Proceedings of the 2015 AAAI Fall Symposium Series*.

Ritschel, H. (2018), "Socially-aware reinforcement learning for personalized human–robot interaction", *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1775-1777.

Robinson, N.L., Connolly, J. and Hides, L. (2020), "Social robots as treatment agents: Pilot randomized controlled trial to deliver a behavior change intervention", *Internet Interventions*, Vol. 21, p. 100320, doi: 10.1016/j.invent.2020.100320.

Robinson, H., Macdonald, B. and Broadbent, E. (2015), "Physiological effects of a companion robot on blood pressure of older people in residential care facilities: a pilot study", *Australasian Journal on Ageing*, Vol. 34 No. 1, pp. 27-32, doi: 10.1111/ajag.12099.

Robinson, H., MacDonald, B., Kerse, N. and Broadbent, E. (2013), "The psychosocial effects of a companion robot: a randomized controlled trial", *Journal of the American Medical Directors Association*, Vol. 14 No. 9, pp. 661-667, doi: 10.1016/j.jamda.2013.02.007.

Rogé, B. (2017), "Robots et autisme", *Enfance*, Vol. 2 No. 2, pp. 283-287, doi: 10.3917/enf.172.0283.

Roger, K., Guse, L., Mordoch, E. and Osterreicher, A. (2012), "Social commitment robots and dementia", *Canadian Journal on Aging / La Revue Canadienne du Vieillissement*, Vol. 31 No. 1, pp. 87-94, doi: 10.1017/S0714980811000663.

Rosanda, V. and Istenic, S.A. (2020), "The robot in the classroom: a review of a robot's role", in Popescu, E., Hao, T., Hsu, T.-C., Xie, H., Temperini, M. and Chen, W. (Eds), *Emerging Technologies for Education*, Springer International Publishing, pp. 347-357, doi: 10.1007/978-3-030-57717-9_40.

Rosenthal-von der Pütten, A.M., Schulte, F.P., Eimler, S.C., Hoffmann, L., Sobieraj, S., Maderwald, S. and Brand, M. (2013), "Neural correlates of empathy towards robots", *Proceedings of the 8th ACM/IEEE International Conference on Human–Robot Interaction*, *IEEE Press*, pp. 215-216, doi: 10.1109/HRI.2013.6483571.

Roshdy, A., Karar, A.S., Al-Sabi, A., Al Barakeh, Z., El-Sayed, F., Beyrouthy, T. and Nait-Ali, A. (2019), "Towards human brain image mapping for emotion digitization in robotics", *Proceedings of the 2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART)*, *IEEE*, pp. 1-5, doi: 10.1109/BioSMART.2019.8734285.

Rowe, M.L., Silverman, R.D. and Mullan, B.E. (2013), "The role of pictures and gestures as nonverbal aids in preschoolers' word learning in a novel language", *Contemporary Educational Psychology*, Vol. 38 No. 2, pp. 109-117, doi: 10.1016/j.cedpsych.2012.12.001.

Šabanović, S., Bennett, C.C., Chang, W.-L. and Huber, L. (2013), "Paro robot affects diverse interaction modalities in group sensory therapy for older adults with dementia", *Proceedings of the 2013*

IEEE International Conference on Rehabilitation Robotics (ICORR), *IEEE*, pp. 1-7, doi: 10.1109/ICORR.2013.6650427.

Saerbeck, M., Schut, T., Bartneck, C. and Janse, M.D. (2010), "Expressive robots in education: Varying the degree of social supportive behavior of a robotic tutor", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2010)*, *ACM*, pp. 1613-1622, doi: 10.1145/1753326.1753567.

Santarossa, S., Kane, D., Senn, C.Y. and Woodruff, S.J. (2018), "Exploring the role of in-person components for online health behavior change interventions: can a digital person-to-person component suffice?", *Journal of Medical Internet Research*, Vol. 20 No. 4, pp. e144, doi: 10.2196/jmir.8480.

Sartorato, F., Przybylowski, L. and Sarko, D.K. (2017), "Improving therapeutic outcomes in autism spectrum disorders: Enhancing social communication and sensory processing through the use of interactive robots", *Journal of Psychiatric Research*, Vol. 90, pp. 1-11, doi: 10.1016/j.jpsychires.2017.02.004.

Scassellati, B., Admoni, H. and Mataric, M.J. (2012), "Robots for use in autism research", *Annual Review of Biomedical Engineering*, Vol. 14 No. 1, pp. 275-294, doi: 10.1146/annurev-bioeng-071811-150036.

Scassellati, B., Boccanfuso, L., Huang, C.-M., Mademtzi, M., Qin, M., Salomons, N., Ventola, P. and Shic, F. (2018), "Improving social skills in children with ASD using a long-term, in-home social robot", *Science Robotics*, Vol. 3 No. 21, p. eaat7544, doi: 10.1126/scirobotics.aat7544.

Scoglio, A.A., Reilly, E.D., Gorman, J.A. and Drebing, C.E. (2019), "Use of social robots in mental health and well-being research: Systematic review", *Journal of Medical Internet Research*, Vol. 21 No. 7, p. e13322, doi: 10.2196/13322.

Serholt, S., Barendregt, W., Vasalou, A., Alves-Oliveira, P., Jones, A., Petisca, S. and Paiva, A. (2017), "The case of classroom robots: Teachers' deliberations on the ethical tensions", *AI and Society*, Vol. 32 No. 4, pp. 613-631, doi: 10.1007/s00146-016-0667-2.

Sharkey, A. (2016), "Should we welcome robot teachers?", *Ethics and Information Technology*, Vol. 18 No. 4, pp. 283-297, doi: 10.1007/s10676-016-9387-z.

Shen, Z. and Wu, Y. (2016), "Investigation of practical use of humanoid robots in elderly care centres", *Proceedings of the 4th International Conference on Human-Agent Interaction (HAI 2016)*, pp. 63-66, doi: 10.1145/2974804.2974831.

Shibata, T. and Coughlin, J.F. (2014), "Trends of robot therapy with neurological therapeutic seal robot, PARO", *Journal of Robotics and Mechatronics*, Vol. 26 No. 4, pp. 418-425, doi: 10.20965/jrm.2014.p0418.

Smakman, M., Vogt, P. and Konijn, E.A. (2021), "Moral considerations on social robots in education: a multi-stakeholder perspective", *Computers and Education*, Vol. 174, p. 104317, doi: 10.1016/j.compedu.2021.104317.

Smarr, C., Prakash, A., Beer, J.M., Mitzner, T.L., Kemp, C.C. and Rogers, W.A. (2012), "Older adults' preferences for and acceptance of robot assistance for everyday living tasks", *Proceedings of the Human Factors and Ergonomics Society. Annual Meeting. Human Factors and Ergonomics Society. Annual Meeting*, Vol. 56 No. 1, pp. 153-157, doi: 10.1177/1071181312561001.

Spaulding, S. (2018), "Personalized robot tutors that learn from multimodal data", *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2018)*, *International Foundation for Autonomous Agents and Multiagent Systems*, pp. 1781-1783.

Sung, H.-C., Chang, S.-M., Chin, M.-Y. and Lee, W.-L. (2015), "Robot-assisted therapy for improving social interactions and activity participation among institutionalized older adults: a pilot study", *Asia-Pacific Psychiatry*, Vol. 7 No. 1, pp. 1-6, doi: 10.1111/appy.12127.

Takayanagi, K., Kirita, T. and Shibata, T. (2014), "Comparison of verbal and emotional responses of elderly people with mild/moderate dementia and those with severe dementia in responses to seal robot, PARO", *Frontiers in Aging Neuroscience*, Vol. 6, p. 257, doi: 10.3389/fnagi.2014.00257.

Tanaka, F. and Matsuzoe, S. (2012), "Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning", *Journal of Human–Robot Interaction*, Vol. 1 No. 1, pp. 78-95, doi: 10.5898/JHRI.1.1.Tanaka.

Thibaut, J.W. and Kelley, H.H. (1959), *The Social Psychology of Groups*, Wiley.

Tiberius, R. and Billson, J.M. (1991), "The social context of teaching and learning", *New Directions for Teaching and Learning*, Vol. 1991 No. 45, pp. 67-86, doi: 10.1002/tl.37219914509.

Turkle, S., Breazeal, C., Duffy, B., *et al*. (2006), "A nascent robotics culture: new complicities for companionship", *AAAI Technical Report Series*, doi: 10.1609/aimag.v30i3.2250.

Van Der Drift, E.J., Beun, R.J., Looije, R., Blanson Henkemans, O.A. and Neerincx, M.A. (2014), A remote social robot to motivate and support diabetic children in keeping a diary, *Proceedings of the 2014 ACM/IEEE International Conference on Human–Robot Interaction*, pp. 463-470, doi: 10.1145/2559636.2559650.

Wada, K., Kouzuki, Y. and Inoue, K. (2012), "Field test of caregiver's manual for robot therapy using therapeutic seal robot", *Alzheimer's and Dementia*, Vol. 8 No. 4, pp. S636-S637, doi: 10.1016/j.jalz.2012.05.1729.

Wada, K., Ikeda, Y., Inoue, K. and Uehara, R. (2010), "Development and preliminary evaluation of a caregiver's manual for robot therapy using the therapeutic seal robot Paro", *Proceedings of the 19th International Symposium on Robot and Human Interactive Communication (RO-MAN 2010)*, pp. 533-538, doi: 10.1109/ROMAN.2010.5598607.

Wayne, A. and Youngs, P. (2003), "Teacher characteristics and student achievement gains: a review", *Review of Educational Research*, Vol. 73 No. 1, pp. 89-122, doi: 10.3102/00346543073001089.

Westlund, J.K., Gordon, G., Spaulding, S., Lee, J.J., Plummer, L., Martinez, M. and Breazeal, C. (2016), "Lessons from teachers on performing HRI studies with young children in schools", *Proceedings of the 2016 11th ACM/IEEE International Conference on Human–Robot Interaction (HRI 2016)*, pp. 383-390, doi: 10.1109/HRI.2016.7451777.

Woo, H., LeTendre, G.K., Pham-Shouse, T. and Xiong, Y. (2021), "The use of social robots in classrooms: a review of field-based studies", *Educational Research Review*, Vol. 33, p. 100388, doi: 10.1016/j.edurev.2021.100388.

Xie, Y. and Peng, S. (2009), "How to repair customer trust after negative publicity: the roles of competence, integrity, benevolence, and forgiveness", *Psychology and Marketing*, Vol. 26 No. 7, pp. 572-589, doi: 10.1002/mar.20289.

Corresponding author
Auxane Boch can be contacted at: auxane.boch@tum.de

A3 Essay 3. Introduction to Video Games Ethics

Reference: Boch, A., Thomas, B., Wanick, V., & Clifford, C. (2024). Introduction to video games ethics. In *Edward Elgar Publishing eBooks* (pp. 313–343). https://doi.org/10.4337/9781803928241.00021

Information about the Book Chapter

Title	Introduction to video games ethics
Authors	Boch, A., Thomas, B., Wanick, V., & Clifford, C.
Publication Date	14 November 2024
Book	The Elgar Companion to Applied AI Ethics
Publisher	Edward Elgar Publishing
Pages	313-343

Copyright information

https://www.e-elgar.com/author-hub/reuse-of-your-work/#accordion-27

Use in PhD theses

You may include the final published versions of your Elgar chapter, article, figure or table in your PhD thesis, citing the original version in the Elgar book or journal, as long as this is no more than 1 or 2 chapters/articles. The thesis may then subsequently be made available in your university's open repository without an embargo.

The chapter/ article should not be reused or posted independently of the thesis without further permission. If you wish to post a standalone chapter or article in an open repository please refer to our author reuse policy or our Open Access publishing options.

This permission does not cover any part of the material that is attributed to another source for which you would have to obtain separate permission.

To obtain free permission please apply through PLSClear.

12. Introduction to video games ethics

Auxane Boch, Bethany Thomas, Vanissa Wanick, and Cacie Clifford[1]

12.1 INTRODUCTION

Video games are interactive electronic games played on a screen, typically using a game console, computer, or mobile device, designed to entertain players through engaging gameplay, graphics, and sound effects. Their history dates back to the 1950s, with the creation of early computer games such as Tennis for Two[2] and Spacewar! (*Spacewar! History*). However it was not until the 1970s and 80s that video games became widely popular, with the introduction of arcade games like Pac-Man and home consoles such as the Atari 2600. Since then, video games have continued to evolve and grow in popularity, with new technologies and game designs continually pushing the boundaries of what is possible in interactive entertainment.

In 2022 and 2023, Minecraft is arguably the most popular game worldwide with 238 million copies sold across all platforms and 1 trillion views of Minecraft videos on YouTube.[3] However, Candy Crush Saga holds the first place position for the most played game worldwide in 2022 with 2.73 billion registered accounts. These statistics highlight the vast reach of video games, with an estimated 3.09 billion active gamers globally,[4] representing about 35% of the world's population at this point in time. Considering their vast outreach, understanding the impacts of video games on individuals, groups, and societies is of utmost importance. And especially in regards to ethics.

'Video game ethics' refers to the study of dilemmas related to video games and their positive or negative impact on individuals and societies. Topics such as the potential effects of

[1] We would like to thank the fantastic Women in Games Ambassador community and the members who enthusiastically participated in discussions with us regarding the content of this chapter, namely; Urooj Iqbal, Growth & Partnerships manager at Playdew; Mariana Heggholmen; Kaela O'Neill, AI Programmer at The Molasses Flood. We also warmly thank Sophie Elizabeth Smith for her proof-reading efforts.

[2] Abby, A. (2022, December 8). *The Complete History of Tennis for Two*. History-Computer. https://history-computer.com/tennis-for-two-complete-history/.

[3] James, P. (2023, March 6). *Most Played & Most Popular Games In The World (2022-2023)*. Gamer Tweak. https://gamertweak.com/most-played-popular-games/.

[4] Howarth, J. (2023, January 18). *How Many Gamers Are There? (New 2023 Statistics)*. Exploding Topics. https://explodingtopics.com/blog/number-of-gamers.

violent[5] or sexist[6] content on gamers' real life behaviours, or video game addiction,[7] are prevalent in the media and scholarly discourse. The field of video game ethics aims to deliberate on those issues based on philosophical and empirical work. By essence, those discussions are meant to constantly evolve, adapting to new challenges and issues that arise in the gaming industry, as well as changes in societal values and norms. Different stakeholders in the gaming industry may have different perspectives on what constitutes ethical behaviour for, in, and around video games, which can add to the complexity of the field. However, there is a growing awareness within the industry regarding the importance of understanding the impact of games on society, with proposals for developers' codes of ethics (see IGDA, Open Gamification, Ethical Games), and conferences such as Games for Change Festival focusing on the potential positive impact of games.

In this chapter, we aim to introduce and discuss key aspects of video games ethics as they relate to gamers and their direct gaming culture. We understand video game ethics here as an ongoing discussion needing to bridge the gap between philosophical opinions and empirical research in pursuit of reaching the best possible understanding of what is good, and what needs to evolve in the sector. We thus decide to build this outline from an empirical basis, offering a psychology driven approach to the impact of games on gamers and their real life behaviours. We will thus introduce the discussions regarding the impact of games on mental health, cognitive skills, social behaviours, and the overall gaming culture. Through those steps, we will utilise theories and research grounded in clinical, cognitive, and social psychology perspectives, while integrating perspectives from video game research and social sciences in general. We will thus provide a solid understanding of the concept and of existing debates surrounding video game ethics as they relate to 'doing good', and the empirical background supporting them.

12.2 THE COGNITIVE SIDE OF THE STORY

To first start our journey, we will discuss the game mechanics from a design perspective and their reported effects on cognitive skills such as executive function and attention. We will then discuss how multiplayer gaming affects social cognition skills. Our aim here is to better understand from a cognitive perspective what are the benefits and challenges of playing video games.

---

5 Hern, A. (2020, July 22). Playing video games doesn't lead to violent behaviour, study shows. *The Guardian*. https://www.theguardian.com/games/2020/jul/22/playing-video-games-doesnt-lead-to-violent-behaviour-study-shows; Triggs, L. (2022, September 19). Violent video games can lead to violent behavior. https://www.Ksnblocal4.com.

6 Gleeson, J. (2014). Vitriolic abuse of Anita Sarkeesian: why the games industry needs her. *The Conversation*.; Gentile, D. (2017, March 28). Researchers find video games influence sexist attitudes News Service Iowa State University. *News Iowa State University*.

7 Barr, S. (2019, October 8). Children with gaming addiction to be offered treatment by NHS. *The Independent*.; Orlando, J. (2022). Can gaming 'addiction' lead to depression or aggression in young people? Here's what the evidence says. *The Conversation*.

12.2.1 Game Mechanics

Game mechanics are the large range of rules and design that shape a game in order to restrict aspects of gameplay to create challenges and interesting game worlds (Tyler, 2023). Each game genre and each game itself will be different in the rules and mechanics it will add to diversify the game from already existing games. Whether it's a merging of two genres or a remake or sequel of a previous game, different mechanics serve to force a player to navigate gameplay in a way that the creator of the game wishes.

As a non comprehensive list, there are a few game mechanics that will be quite obvious when looking at the game itself:[8]

- **Turn based mechanics** such as Civilisation VI in which players take turns to build and expand their society.
- **Card based gameplay and deck building** such as Hearthstone in which players collect cards to build specific card decks and defeat their adversary in combat.
- **Dice mechanics**, like the Dice Block featured in the Mario Party series, dictate the number of spaces a character will move.
- **Resource management**, as seen in Stardew Valley, in which the player becomes a self-sufficient farmer.
- **Movement and area control dynamics**, such as seen in League of Legends, in which players try to, in teams, control the battlefield by destroying the enemies base by moving their character on the map, controlling territories, and engaging in battles.
- **Abilities and levelling systems mechanics**, like in the open-world action role-playing game Skyrim in which the player can level up by completing quests, exploring the world, and defeating enemies. By levelling up, they also improve their existing abilities and gain new ones.
- **Fog of war mechanics** such as in Starcraft, in which players control one of three alien species, each with unique units and abilities. Areas of the map that have not yet been explored by the player are hidden from view until they are revealed by a unit or a building, lifting the fog on them.

Each game can be broken down into a systemised list of these types of game mechanics such as the ones presented above, and more. Interestingly, games within the same genre can be differentiated based on how they implement and use these mechanics. A number of different genres exist, allowing for each player or potential player to 'find the shoe that fits'. We here give an incomplete overview of genres to allow for a full understanding of the different layers of mechanics and design involved in video games.

- **Adventure games**, usually emphasising exploration and storytelling and involving puzzles and other challenges that advance the plot. Series of games such as Uncharted, Tomb Raider, and Zelda incarnate this type remarkably.

[8] Tyler, D. (2023, February 23). *The Beginner's Guide to Game Mechanics*. Video Game Design and Development.

- **Role Playing Games (RPGs)**, involving character customisation and advancement, engage the player to make choices that can affect the story and the game world. Notably renowned franchises of such games are The Elder Scrolls, World of Warcraft, and Assassin's Creed.
- **Strategy Games**, usually requiring the player to plan and execute actions carefully to achieve their goals, often involve resource management and building up an army or civilisation. Prime examples of such games are series such as StarCraft, Total War, and Warcraft.
- **Narrative games** are games that focus on storytelling and the narrative experience, often using interactive gameplay elements to enhance the narrative. These games typically prioritise plot, character development, and dialogue over other gameplay mechanics such as action or puzzle-solving. Some examples of narrative games include Life is Strange, The Walking Dead, and Detroit: Become Human.
- **Multiplayer games** allow multiple gamers to compete or collaborate with each other in real-time. Different ways to reach this exist, through wired networks between gaming devices, or over the internet. The latter is more common in recent years. When played online, some games regroup a large number of players, and are then called Massively Multiplayer Online games, or MMOs. Examples of popular MMOs include Fortnite, League of Legends, and EVE online.

Game genres will each make use of their mechanics differently and as such, each genre and game will have different impacts on the players abilities, including cognitive ones.

12.2.2 Cognitive Skills and Gaming

According to a review published in 2020,[9] video games have the ability to train players in various skills, which can lead to changes in their perceptions. The review identified two key areas of focus: problem-solving abilities and time management. Puzzle games, like Dr. Leighton, are effective in training cognitive abilities and can enhance problem-solving skills. Time management is another skill that is developed through gaming, as players learn to make decisions within specific time frames, which can vary from game to game. Games like Detroit: Become Human for example place a heavy emphasis on decision making, where each choice impacts the storyline and time pressure is often present. As a result of their game experience, players tend to make decisions more confidently and with greater belief in their choices.

In another review, authors highlight the positive impact of action video games, such as Uncharted, for attentional control and learning.[10] Interestingly, when it comes to language learning specifically, the same areas of the brain were shown to be used similarly as in 'real life' language learning.[11] Thus, playing games seems to heighten the attention skills of

---

[9] Reynaldo, C., Christian, R., Hosea, H., & Gunawan, A. A. (2021). Using video games to improve capabilities in decision making and cognitive skill: a literature review. *Procedia Computer Science*, *179*, 211–221.

[10] Cardoso-Leite, P., & Bavelier, D. (2014). Video game play, attention, and learning: how to shape the development of attention and influence learning?. *Current opinion in neurology*, *27*(2), 185–191.

[11] Prena, Reed, A., Weaver, A. J., & Newman, S. D. (2018). Game Mechanics Matter: Differences in Video Game Conditions Influence Memory Performance. *Communication Research Reports*, *35*(3), 222–231

a person, skills that can translate in other areas of one's life such as learning, justifying the gamification efforts as it relates to education.[12]

Finally, a meta-analysis of twenty experimental studies revealed that video game training seemed to have positive effects on several cognitive functions including reaction time, linking back to our previous argument in regards to time management, memory, learning, and cognition globally.[13] In other words, video games seem to show a positive effect on the evolution of players' cognitive skills in general, skills that translate in other areas of their lives.

All in all, game mechanics can affect players not just through varying difficulty levels in playing a game, but through additional skill building. These skills can come from individual types of game mechanics or a collection of multiple types. Varying the design of these game mechanics is what allows for the diverse nature of games. Game mechanics can also translate to other areas such as education to reach more adapted learning practices.

12.2.3 Social Cognition and Multiplayer Games

If we see effects for cognition in general, it is interesting to consider the social cognitive aspect of video games, especially as it relates to multiplayer games. In those games, by design, interaction with other players is necessary to reach goals and evolve in the game.

Social cognition refers to the way we think about and understand the social world around us. In Fiske and Taylor's[14] words, it can be defined as 'thinking about people'. Korman et al. (2015) expand on this definition and explain it as a branch of cognitive psychology that focuses on how we perceive, interpret, and process social information. This includes recognising (sometimes cultural) social cues, forming opinions about people and their behaviour as well as understanding what others might be thinking or feeling, which here relates to empathy. Our ability to navigate the social world is crucial for successful social interactions and relationships, and social cognition plays a key role in guiding our behaviour and responses to others. In other words, social cognition is also involved in teaching us which behaviours are appropriate or not, and how to understand our social environment and navigate it better.

When looking into video games, social cognition skills seem to be especially involved and impacted by playing multiplayer games. In such games, social interactions usually have the purpose of collaboration or competition. Interestingly, the primary motivation for young adults to engage in online multiplayer gaming such as Counter Strike and World of Warcraft seem to be social interaction in the specific shapes of cooperation and communication.[15] Competition is also an important factor, and it has been discussed that this may create aggressive behaviour

---

[12] Manzano-León, A., Camacho-Lazarraga, P., Guerrero, M. A., Guerrero-Puerta, L., Aguilar-Parra, J. M., Trigueros, R., & Alias, A. (2021). Between level up and game over: A systematic literature review of gamification in education. *Sustainability*, *13*(4), 2247.

[13] Toril, P., Reales, J. M., & Ballesteros, S. (2014). Video game training enhances cognition of older adults: a meta-analytic study. *Psychology and aging*, *29*(3), 706.

[14] Fiske, S. T., & Taylor, S. E. (1991). *Social cognition.* Mcgraw-Hill Book Company.

Fredrickson, B. L., Roberts, T. A., Noll, S. M., Quinn, D. M., & Twenge, J. M. (1998). That swimsuit becomes you: sex differences in self-objectification, restrained eating, and math performance. *Journal of personality and social psychology*, *75*(1), 269.

[15] Frostling-Henningsson, M. (2009). First-person shooter games as a way of connecting to people: 'Brothers in blood'. *Cyberpsychology & behavior, 12*(5), 557–562.

from opponents towards each other.[16] These results are disputed by more recent field research, in opposition to in-lab research that builds on self-reported measures instead of observation of real life behaviour.[17] Thus, the impact of competition on social skills is still debated.

On the other hand, researchers on the topic of cooperation and social skills enhancement agree on the benefit of such game design choices for developing social abilities. In an early study investigating the educational value of digital games, researchers evaluated the collaborative puzzle game Super Word Search based on different dimensions of learning.[18] In their assessment, the authors highlight multiple social skills improvements due to the collaboration aspect of the game; teamwork and social skills in general were improved by the experience, mutual support was shown through the initiative of better-performing students to help the slower ones, and the participants, who were children, developed coping strategies when frustration or other negative emotions arose to calm themselves and find alternative solutions as a group. Those findings regarding the beneficence of collaborative games for social skills can be seen in other studies. For example, in 2017, Koiva et al. investigated the use of the game Emotion Detectives in two Finnish daycare centres. The children taking part showed an improvement in their socio-emotional knowledge in peer-interactions. In other words, the understanding and ability to recognise and interpret emotions in oneself and others, as well as the ability to use this knowledge in social context was trained by playing this collaborative game. Those specific qualities relate to empathy abilities. A 2022 review of 33 papers also revealed that playing collaborative and teamwork games was shown to improve communication skills over time,[19] especially in the case of high level players such as seen in the competitive Esports scene.

Such social and communication skill development is beneficial to the players social abilities in and out of the game setting. Indeed, cooperation in-game seems to translate to the players' 'real world' behaviour in the form of prosocial behaviours such as proactively helping others.[20] On the other hand, communication skills such as learning a new language more easily through collaborative playing can also be seen in the context of massively multiplayer online role-playing games (MMORPGs), illustrated by Guild Wars 2 gamers learning English as a second language, driven by the collaborative aspect of the game.[21]

---

[16] Adachi, P. J., & Willoughby, T. (2011). The effect of video game competition and violence on aggressive behavior: Which characteristic has the greatest influence?. *Psychology of violence, 1*(4), 259.

[17] McLean, D., Waddell, F., & Ivory, J. (2020). Toxic teammates or obscene opponents? Influences of cooperation and competition on hostility between teammates and opponents in an online game. *Journal For Virtual Worlds Research, 13*(1).

[18] Hong, J. C., Cheng, C. L., Hwang, M. Y., Lee, C. K., & Chang, H. Y. (2009). Assessing the educational values of digital games. *Journal of Computer Assisted Learning, 25*(5), 423–437.

[19] Lie, A., Stephen, A., Supit, L. R., Achmad, S., & Sutoyo, R. (2022, October). Using Strategy Video Games to Improve Problem Solving and Communication Skills: A Systematic Literature Review. In 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS) (pp. 1–5). IEEE.

[20] Dolgov, I., Graves, W. J., Nearents, M. R., Schwark, J. D., & Volkman, C. B. (2014). Effects of cooperative gaming and avatar customization on subsequent spontaneous helping behavior. Computers in human behavior, 33, 49–55.

[21] Ng, L. L., Azizie, R. S., & Chew, S. Y. (2022). Factors influencing ESL players' use of vocabulary learning strategies in massively multiplayer online role-playing games (MMORPG). The Asia-Pacific education researcher, 31(4), 369–381.

Thus, video games have the potential to enhance cognitive and social skills in players. In terms of cognitive skills, puzzle games are effective in training problem-solving abilities, while action video games improve attentional control and learning. Multiplayer games, on the other hand, offer opportunities for social interaction and cooperation, which can enhance social cognition skills, including empathy and communication. Although competition in multiplayer games can be controversial, game mechanics can be designed to encourage collaboration and teamwork, leading to improved socio-emotional knowledge and communication skills. Overall, the diverse nature of game mechanics allows for the development of various skills, which can be translated into other areas of players' lives, including education.

12.3 FROM COGNITIVE TO MENTAL HEALTH DISCUSSIONS

If video games show great results to teach cognitive and social skills, they can be controversial in regards to mental health or wellbeing. Practices such as the dark patterns are sometimes included in the design of the game, possibly influencing the gamer towards unwanted behaviours such as spending money on loot boxes. In addition, controversies relating to gaming addiction disorder's (newly) defined diagnosis, and the general mental and physical well being of gamers are discussed. In this part, we will investigate those sometimes controversial conversations and provide empirical elements behind the given arguments.

12.3.1 Loot Boxes and Dark Patterns

In a study conducted in 2020 in the Steam platform,[22] found that loot boxes and cosmetic microtransactions (e.g., aesthetics in-game purchases, such as decorations and fashion) grew exponentially during the period of 2012 to 2014, with the peak in 2019. Compared to cosmetic microtransactions, pay-to-win strategies in desktop games had a great exposure in 2015, but did not achieve the same growth. Yet, in the same study, the authors found that at least 70 % of gamers played games with loot boxes, which raised concerns about the long exposure to these strategies. Loot boxes usually use a 'gacha' mechanic (similar to a toy vending machine), alluring players to spend money (or in-game currency, watching an ad, etc.) for the chance to get a random reward. In this case, these rewards can be rare and players might not be able to acquire such items any other way. These loot boxes can appear in many ways in games. An example is Animal Crossing: Pocket Camp. In the game, loot boxes are represented by new Fortune Cookies that can give the players random rewards. In this case, most of the rewards are cosmetics and can be used to personalise your character or your camp. League of Legends (LoL), on the other hand, has a Hextech Chest, which is a treasure chest that can be purchased in the game with specific rules (e.g., the player cannot open three chests in a row, etc.). This design decision might be because some of these loot boxes are similar to 'pay-to-win', in other words, giving players some advantage in the game such as bonus, power-ups, etc. in exchange for payment. Yet, loot boxes have their own loop, which makes them different from 'pay-to-win'. This type of loop can be considered as a small 'Skinner box', in which players

[22] Zendle, D., Meyer, R., & Ballou, N. (2020). The changing face of desktop video game monetisation: An exploration of exposure to loot boxes, pay to win, and cosmetic microtransactions in the most-played Steam games of 2010–2019.

repeat an activity in order to receive a reward. Different from the core loop of the game, loot boxes can be presented at different times for players and might have their own rules (e.g., Hextech Chest) to avoid loopholes in the game.

The use of loot boxes and pay-to-win strategies in video games has given rise to concerns over the employment of 'dark patterns'. The term refers to strategies that can cause negative experiences to players without their explicit consent.[23] The ego depletion theory sheds light on how dark patterns operate in the context of video games.[24] According to this theory, emotional and attentional cognitive resources are used up after complex and engaging tasks, leading to a decrease in self-control, and thus, willpower. With this lower willpower, impulsive behaviours such as making purchases are more likely to happen. Such behaviours are important to differentiate from a case of compulsive disorder, as it is context dependent. Madigan[25] highlights the role of goal-oriented behaviour in games in his book, *Getting Gamers*; as players engage in these games, their emotional and cognitive resources can be drained, leading to a state of ego depletion. This can result in players agreeing to the status quo or caving in to lesser self-control, thereby increasing their acceptance of microtransactions in games. Moreover, humans have a natural bias in favour of accepting default options; it is named the status quo bias.[26] A real life example of such bias is seen when renewing subscriptions almost automatically, without deliberation or thought, because it requires less effort and can lead to a continuation of the current status quo. The timing of microtransaction proposals is thus critical. For instance, offering microtransactions immediately after a long and challenging combat sequence in game may increase the likelihood of players making purchases due to the depletion of psychological resources. Thus, the use of dark patterns in video games lies in the understanding of players' cognitive functioning, and is a concerning issue that has implications for players' experiences and well-being. Yet, the term 'dark pattern' can be dubious and perhaps not the best way to discuss ethical approaches as there is still a need to establish an ethical framework that can be used for future discussions.[27]

The other issue with loot boxes, in particular, is that it blurs the line with gambling mechanisms. The reason for that might be because the feeling players have whilst opening a loot box is very similar to addictive qualities in gambling.[28] However, research in this area is still in its infancy and it is not very clear what gaming disorder really is. In this next part, we will present the discussion relating to this newly defined disorder.

---

[23] Mori, S. (2021, May 25). *Help Bring Dark Patterns To Light*. Electronic Frontier Foundation.

[24] Muraven, M., Buczny, J., & Law, K. F. (2019). Ego Depletion: Theory and Evidence. *The Oxford Handbook of Human Motivation*, 111–134.

[25] Madigan, J. (2016). *Getting Gamers: The Psychology of Video Games and Their Impact on the People who Play Them*. Rowman & Littlefield Publishers.

[26] Anderson (2003).

[27] Deterding, C. S., Stenros, J., & Montola, M. (2020, January). Against 'Dark Game Design Patterns'. In *DiGRA'20-Abstract Proceedings of the 2020 DiGRA International Conference*. York.

[28] Brady, A., & Prentice, G. (2021). Are loot boxes addictive? Analyzing participant's physiological arousal while opening a loot box. *Games and Culture*, *16*(4), 419–433.

12.3.2 Addiction or (Internet) Gaming Disorder

In 2013, the Internet Gaming Disorder (IGD) was first mentioned in the *Diagnostic and Statistical Manual of Mental Disorders*[29] as a Condition for Further Studies, and compared to the now well established Gambling Disorder; both conditions have in common their substance-less addiction characteristic. To understand better how IGD translates in the real world, it is important to understand Gambling Disorder. The gambling addiction seems to activate the reward system of a player in a similar way that a drug or substance such as an opioid would do, and induce comparable behavioural symptoms such as strong cravings, repeated unfruitful efforts to reduce the prevalence of the behaviour, and the persistence of the behaviour regardless of its impact on important personal and affective relationships, studies, or career.[30] In other words, IGD could be simplified in excessive online gaming behaviours despite their negative consequences.[31] In 2018, the World Health Organisation published an updated *International Classification of Disease*[32] in which Gaming Disorder (GD) can be found as, this time, a fully recognised diagnosis. Interestingly, contrary to the nine possible symptoms present in the DSM-5 proposition, the ICD-11 reduces the disorder to three main recurring and persistent ones: a loss of control over gaming, an elevated priority placed on gaming activities, and a persistent inclination towards gaming despite negative repercussions. Importantly, both diagnostic manuals highlight the need for significant impairment or distress by the patient to be able to pose a clinical diagnosis on their situation.[33] In all, the diagnosis of Gaming Disorder can now be assessed by trained clinicians. This addition in the diagnosis manuals is meant to support individuals suffering by identifying their issue and proposing best care opportunities. However, given the lack of clarity and ongoing discussion about the details of this new official diagnosis, its creation was somewhat controversial.

In a notable example, academics across the globe joined forces and published a paper discussing the implications of this proposal, highlighting that it might cause 'more harm than good' as the Game Disorder label could be used to stigmatise players/gamers.[34] In the paper, Aarseth et al[35] also mentioned that there is a need to properly define gaming addiction or 'dis-

---

[29] *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*; APA, 2013).

[30] APA (2013).

[31] Feng, W., Ramo, D. E., Chan, S. R., & Bourgeois, J. A. (2017). Internet gaming disorder: Trends in prevalence 1998–2016. *Addictive behaviors, 75*, 17–24.

[32] *ICD-11*; WHO, 2018.

[33] Borges, G., Orozco, R., Benjet, C., Mart´Inez, K. I. M., Contreras, E. V., P´Erez, A. L. J., Cedr´Es, A. J. P., Uribe, P. C. H., Couder, M. a. C. D., Gutierrez-Garcia, R. U. A., Ch´Avez, G. E. Q., Albor, Y., Mendez, E., Medina-Mora, M. E., Mortier, P., & Ayuso-Mateos, J. L. (2020). (Internet) Gaming Disorder in DSM-5 and ICD-11: A Case of the Glass Half Empty or Half Full: (Internet) Le trouble du jeu dans le DSM-5 et la CIM-11: Un cas de verre à moitié vide et à moitié plein. *The Canadian Journal of Psychiatry*, 66(5), 477–484. https:// doi .org/ 10 .1177/ 0706743720948431

[34] Aarseth, E., Bean, A. M., Boonen, H., Colder Carras, M., Coulson, M., Das, D., ... & Van Rooij, A. J. (2017). Scholars' open debate paper on the World Health Organization ICD-11 Gaming Disorder proposal. *Journal of behavioral addictions*, 6(3), 267–270.

[35] Aarseth, E., Bean, A. M., Boonen, H., Colder Carras, M., Coulson, M., Das, D., ... & Van Rooij, A. J. (2017). Scholars' open debate paper on the World Health Organization ICD-11 Gaming Disorder proposal. Journal of behavioral addictions, 6(3), 267–270.

order' via a proper diagnosis as it is not very clear if this behaviour occurs as a coping mechanism, and therefore a symptom of another problem, or as a disorder independent of others. Indeed, depression, loneliness and social anxiety are known factors to predict problematic use of video games as they allow for escapism and, when playing MMO, non face-to-face social interactions.[36] A lack of clarity also occurs in the chronic or episodic aspect of the disorder. In other words, it is unclear if the disordered behaviours happen: sometimes, during a given timeframe; episodically, that is, in reaction to something and thus possibly as a coping mechanism; or chronically, regardless of the individual's context.

Going back to the practical diagnostic aspect of this addiction, recent research was added to the ICD-11 website and discusses a variety of diagnostic methods and reports that need to be carefully scrutinised.[37] Finally, it seems important to highlight that regardless of the increase of the accessibility of gaming technologies from 1998 to 2016, a thorough systematic review of the literature on the topic concluded that the average prevalence of GD did not increase.[38]

Thus, conversations are still ongoing regarding the practicality and veracity of the current diagnosis itself, and the toolkit available to properly assess the issue when it arises, while pointing out the harm a misunderstanding of this condition by the general public could have on the perception of gamers. More research needs to clarify what is and what isn't a disorder when it comes to gaming behaviours, and provide proper assessment methodologies for practitioners.

12.3.3 Overall Well Being

As seen previously, video games have been the subject of much debate in terms of their potential impact on players' well-being. Interestingly, despite the concerns that have been raised, games also have the potential to increase gamers contentment. In a recent study conducted by Johannes et al.,[39] the authors investigated the relationship between video game playtime and affective well-being, using objective measures of playtime instead of self-reported data. Affective well being is here the frequency and intensity with which individuals experience positive or negative emotions.[40] The games used for this study were Plant vs. Zombie: Battle for Neighborville (Electronic Arts) and Animal Crossing: New Horizon (Nintendo). Contrary to concerns regarding excessive playtime possibly leading to addiction and poor mental health, the study found only a (small) positive relationship between game play and well-being. The study also revealed that the need-satisfaction and motivations during play were independently related to well-being, but did not interact with playtime. In other words, the fulfilment and

---

[36] Maroney, N., Williams, B. J., Thomas, A., Skues, J., & Moulding, R. (2019). A stress-coping model of problem online video game use. *International Journal of Mental Health and Addiction, 17*, 845–858.

[37] Darvesh, N., Radhakrishnan, A., Lachance, C. C., Nincic, V., Sharpe, J. P., Ghassemi, M., ... & Tricco, A. C. (2020). Exploring the prevalence of gaming disorder and Internet gaming disorder: a rapid scoping review. *Systematic reviews*, *9*, 1–10.

[38] Feng et al. (2017).

[39] Johannes, N., Vuorre, M., & Przybylski, A. K. (2021). Video game play is positively correlated with well-being. *Royal Society open science, 8*(2), 202049.

[40] Luhmann, M., Krasko, J., & Terwiel, S. (2021). Subjective well-being as a dynamic construct. *Elsevier EBooks*, 1231–1249.

gratification gained in the game by reaching the motivational objectives or desires impacted the gamers' moods positively regardless of how long they played. Similarly, Raith et al.[41] found a significant positive relationship between playing massively multiplayer online games (MMOs) and social well-being, regardless of age and gaming habits. Thus, games can also have a positive impact on gamers' social happiness.

On the other hand, Cruea[42] highlights that certain video games, such as Playne, or Journey, have been developed with the specific goal of improving meditation skills and mindfulness. These video games build on the induction of a flow state; a mental state in which an individual is completely focused on a single task or activity, sometimes referred to as 'being in the zone'. Interestingly, a flow state can also happen when not focused on a task, and is then usually referred to as mindfulness. Enjoying this kind of focus comes with positive impacts for enjoyment, and is known to remove the barriers of self-referential thinking such as worrying, or self-reflection.[43] This state is also usually accompanied with a feeling of accomplishment, meaningfulness, and positive mood.[44]

Interestingly, video games have also been tested to support the development of physical well-being over mental well-being. In a study evaluating the effects of active video games on physical fitness, reaction times, self-perception, and enjoyment levels in 106 inactive and technologically preoccupied children, results showed that the active video game group had significant improvements in weight, body mass index, corresponding z scores, reaction times, and self-perception.[45] Children in the active video game group also reported high levels of enjoyment, indicating a motivational aspect for continued play. The authors concluded that active video games could thus be used as beneficial tools to prevent obesity in inactive non-obese children and improve personal, social, and intellectual development. Active video games also showed to raise self-esteem, induce enjoyment, improve the personal and intellectual development of children in addition to socialising and seem to be a safe alternative to indoor sedentary video games.

Thus, while the effects of video games on well-being may be small, cumulative effects over time could still have significant positive impact on gamers, and video games have the potential to be used as readily-available tools for promoting mindfulness, teaching positive behaviours for one's physical health, and improving gamers' overall well-being. As we touch upon the behavioural impact of video games, our next section will regard the well debated topic of the

---

[41] Raith, L., Bignill, J., Stavropoulos, V., Millear, P., Allen, A., Stallman, H. M., ... & Kannis-Dymand, L. (2021). Massively multiplayer online games and well-being: A systematic literature review. *Frontiers in Psychology*, 12, 698799.

[42] Cruea, M. D. (2020). Gaming the mind and minding the game: mindfulness and flow in video games. *Video Games and Well-being: Press Start,* 97–107.

[43] Van Der Linden, D., Tops, M., & Bakker, A. B. (2021). The neuroscience of the flow state: involvement of the locus coeruleus norepinephrine system. *Frontiers in Psychology*, *12*, 645498.

[44] Csikszentmihalyi, M., & Nakamura, J. (2010). Effortless attention in everyday life: A systematic phenomenology. *Effortless attention: A new perspective in the cognitive science of attention and action*, 179–189.

[45] Coknaz, D., Mirzeoglu, A. D., Atasoy, H. I., Alkoy, S., Coknaz, H., & Goral, K. (2019). A digital movement in the world of inactive children: favourable outcomes of playing active video games in a pilot randomized trial. *European Journal of Pediatrics, 178*, 1567–1576.

impact of gaming on pro- or anti- social real life behaviour, also sometimes referred to as moral behaviour.

12.4 VIDEO GAMES AND SOCIAL BEHAVIOURS

12.4.1 The Gamer's Dilemma

The moral dimension of video games is quite complex, and games such as Detroit: Become Human,[46] or Telltale's The Walking Dead,[47] require players to confront difficult moral choices. The Gamer's Dilemma refers to the moral dilemma raised by actions in video games that have a moral significance similar to their real-world counterparts, while others do not have moral significance.[48] Two main theories are at odds according to Davnall:[49] 'inflationary' theories consider virtual worlds a metaphysical novelty generated by video games, while 'deflationary' theories consider video games purely as systems for generating images. However, neither theory fully explains the moral complexity of the gamer's actions and reactions to their gaming experience. Davnall[50] proposes a new theory of the gamer as a performer, analogous to stage and cinema actors, that can capture this complexity. Ultimately, the gamer's dilemma is a moral issue that relates to what sorts of performance are morally permissible to engage in.

To understand better the complexity and consequences on behaviours of moral decision making in game, we will first define what we understand as morality.

12.4.2 The Concept of Morality

The concept of morality can be defined in two distinct senses, as outlined by the Stanford Encyclopedia of Philosophy (2020). In the normative sense, morality refers to a code of conduct that would be put forward by all rational people under specified conditions. This code must involve impartiality and have the function of enabling people to live together in groups. It governs interpersonal interactions and commits individuals to regarding certain behaviours as immoral. The Western tradition of moral philosophy has historically focused on deontology, consequentialism, and virtue ethics as the major forms of normative ethical theory for determining ethical actions:[51] deontology considers the action in and of itself, while consequen-

---

[46] Holl, E., & Melzer, A. (2022). Moral minds in gaming: A quantitative case study of moral decisions in Detroit: Become human. *Journal of Media Psychology: Theories, Methods, and Applications, 34*(5), 287–298.

[47] Stang, S. (2019). This action will have consequences: Interactivity and player agency. *Game Studies*, *19*(1).

[48] Luck, M. (2009). The gamer's dilemma: An analysis of the arguments for the moral distinction between virtual murder and virtual paedophilia. *Ethics and Information Technology, 11*(1), 31–36. https://doi.org/10.1007/s10676-008-9168-4

[49] Davnall, R. (2021). What does the gamer do?. *Ethics and Information Technology*, *23*(3), 225–237.

[50] Davnall (2021).

[51] Thomas, A. J. (2011). Normative Ethics. *Philosophy.* https:// doi .org/ 10 .1093/ obo/ 9780195396577-0082.

tialism focuses on the end result of behaviour. Virtue ethics, on the other hand, emphasises the importance of being charitable or benevolent, and a virtue is an excellent trait of character that goes all the way down and involves the wholehearted acceptance of a distinctive range of considerations as reasons for action.[52]

In contrast, the descriptive sense of morality refers to certain codes of conduct put forward by a society or a group, or accepted by an individual for their own behaviour. This stems from a relational nature and thus is the lens through which research fields such as psychology, anthropology, and sociology consider morality. In this context, morality has been characterised as a set of psychological adaptations that permit self-interested individuals to benefit from cooperation. This definition is supported by the notion of moral systems, which consist of a complex interplay of various elements such as values, virtues, or norms as well as evolved psychological mechanisms. The ultimate goal of these systems is, according to this definition, to suppress or regulate self-interest in order to facilitate the creation of cooperative societies.[53] Now that we better understand the concept of morality and the scope through which we will consider it in this chapter, a question arises: how do humans make moral decisions?

12.4.3 Moral Decision Making and Video Games

Haidt's moral intuitionism model[54] proposes that intuition is the precursor of moral judgement, followed by slow moral reasoning to justify the initial gut feeling. Social moral judgments are based on feelings and autobiographically established beliefs, which are taught by the culture and environment, also called norms. Moreover, Greene and Haidt[55] suggest that reflection on one's belief through perspective taking can lead to changing beliefs when reasoned consideration conflicts with intuitions. This last point is important as it suggests that in game experience might impact real life attitudes of the player, depending on empathy. Empathy is an innate capacity that allows individuals to comprehend and share the thoughts and emotions of others, and it can be distinguished into three categories: emotional, motivational, and cognitive.[56] Cognitive empathy, also called perspective taking, may play a role in shaping moral behaviours and reasoning, as explained earlier.[57]

In the context of video games, perspective-taking refers to the player's identification with the avatar.[58] This identification or self-representation is discussed by Yee & Bailenson[59] as the Proteus Effect, describing a behavioural conformity with a virtual self-representation, or

---

[52] Thomas (2011).

[53] Greene, J. (2014). *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. Penguin; Haidt, J. (2012). *The Righteous Mind: Why Good People are Divided by Politics and Religion*. Penguin.

[54] Haidt (2012).

[55] Greene (2014) and Haidt (2002).

[56] Camassa (2019); Decety, J., & Cowell, J. M. (2014). The complex relation between morality and empathy. *Trends in cognitive sciences, 18*(7), 337–339; Decety (2015).

[57] Haidt (2012).

[58] Cohen, J. (2001). Defining identification: A theoretical look at the identification of audiences with media characters. *Mass Communication and Society, 4*(3), 245–264. https://doi.org/10.1207/S15327825MCS0403_01; Klimmt. (2009).

[59] Yee & Bailenson (2007)

avatar. It has been observed that this effect could last post-gaming experience.[60] Going back to the term perspective taking, its consequences seem to have a significant impact on one's motivation towards justice for a group or in general.[61] In other words, it can motivate prosocial behaviours, or positive behaviours, towards others. To sum up, taking someone's perspective in a video game, and thus experiencing some type of situation through identification/self-representation with an avatar, might lead to adding to the library of knowledge of a gamer, and participating in changing their social beliefs, or social norms, finally having an impact on their behaviour.

Interestingly, a study based on Red Dead Redemption 2 found that video game based empathy for characters seemed to relate to morality when it comes to harmful situations and moral decisions, which are often based on several moral intuitions.[62] This entails that negative emotional situations would lead to higher empathy for the character, and thus stronger effect on the moral deliberation and decisions of the player in game, with possible repercussions outside of the game for the player. This suggestion is supported by Smiley[63] who studied this identification effect on morality in books, concluding that when a character a reader identifies with suffers social immorality, participants had a tendency to change their moral judgements regarding this injustice in society. The effect is explained by Tousignant et al;[64] individuals might feel the negative emotions of peers through this perspective taking, and thus be motivated to change their own behaviours to decrease the uncomfortable emotional experience. This could lead to the reduction of partiality in the attitude towards a prejudiced social group. Thus, when exposed to morally and emotionally uncomfortable choices such as for instance, allowing a little girl to see an android's corpse in Detroit: Become Human, the perspective taking a player has might impact their beliefs towards android as objects or not (in the context of the game), and may impact their overall perception of artificial intelligence systems.

On the other hand, it is argued that in certain contexts, immoral content or actions in video games can result in reflection and moral deliberation for gamers.[65] Indeed, technological advances in graphics and social features, such as emotions, can also contribute to a greater sense of interactivity and immersion in video games.[66] This feeling of being socially and self-present in the game world can make players feel more connected to the moral dilemmas

---

[60] Pena, J., Hernández Pérez, J. F., Khan, S., & Cano Gómez, Á. P. (2018). Game perspective-taking effects on players' behavioral intention, attitudes, subjective norms, and self-efficacy to help immigrants: the case of "papers, please". *Cyberpsychology, Behavior, and Social Networking, 21*(11), 687-693. https://doi.org/10.1089/cyber.2018.0030

[61] Decety & Yoder (2016).

[62] Grohmann, L., Holl, E., & Melzer, A. (2021). Moral Judgment in Video Games: Effects of Medium, Moral Intuitions and Media-Based Empathy. In *12th Media Psychology Conference (MediaPsych 2021)*, p. 100.

[63] Smiley, J. (2008). *13 Ways of Looking at the Novel*. Anchor.

[64] Tousignant, B., Eugène, F., & Jackson, P. L. (2017). A developmental perspective on the neural bases of human empathy. *Infant Behavior and Development, 48*, 5–12.

[65] Bowman, N. D., Ahn, S. J., & Mercer Kollar, L. M. (2020). The paradox of interactive media: The potential for video games and virtual reality as tools for violence prevention. *Frontiers in Communication*, 104; Katsarov, J., Christen, M., Mauerhofer, R., Schmocker, D., & Tanner, C. (2019). Training moral sensitivity through video games: A review of suitable game mechanisms. *Games and Culture, 14*(4), 344–366.

[66] Hartmann & Vorderer (2010).

presented, leading them to make decisions that are consistent with their personal values and beliefs.[67] Thus, personal morality plays an important role in video games decisions and can affect the choices made by players when presented with moral dilemmas. But when considering why gamers play morally challenging video games, Holl et al. (2020) find that three main motivations prevail: escapism, exploration, and situational motivation. Escapism can lead players to abandon their moral beliefs in order to enjoy moral transgressions, while exploration is characterised by experimentation and lower emotional engagement, and multiple playthroughs may reduce moral engagement. Thus, moral decisions in game might not alway impact the gamers the same way depending on why and how they are playing. This is important to consider to understand better how video games can reflect or reinforce moral codes, but also contribute (or not) to their formation for an individual.

The question we will now address is how do video game moral decision making experiences impact gamers in their everyday moral deliberation, and thus, social behaviour, building on three moral controversies raised by scholars and the media alike; violence, sexism, and cultural appropriation.

12.4.4 Violence and Aggressive Behaviour

We will here first discuss the ongoing debate about the link between violent video games and real-life aggression, and summarise recent research findings on this topic. The issue of whether violent video games cause real-life aggression has been a long-standing and heavily debated topic in the research community.[68]

On one hand, a segment of the academic community argues that violent video games may have negative effects on social behaviour. This is in line with social learning theory, which posits that children learn to exhibit aggressive behaviour by observing others acting aggressively and witnessing how such behaviours are reinforced over time.[69] It is argued that playing violent video games, which often feature approved and rewarded in-game violence, may lead to an increase in aggression that is then translated into the player's everyday behaviour.[70] Furthermore, meta-analyses on this topic provide evidence for the impact of violent video games on behaviour.[71] However, concerns have been raised about the methodological approaches used in these studies. For instance, Hilgard et al.[72] contend that a publication bias

---

[67] Weaver & Lewis (2012).

[68] Greitemeyer, T. (2022). The dark and bright side of video game consumption: Effects of violent and prosocial video games. *Current Opinion in Psychology*, 101326.

[69] Bandura, A., & Walters, R. H. (1977). *Social learning theory* (Vol. 1). Prentice Hall: Englewood Cliffs.

[70] Gentile, D. A., & Gentile, J. R. (2008). Violent video games as exemplary teachers: A conceptual analysis. *Journal of Youth and Adolescence*, *37*, 127–141.

[71] Prescott, A. T., Sargent, J. D., & Hull, J. G. (2018). Metaanalysis of the relationship between violent video game play and physical aggression over time. *Proceedings of the National Academy of Sciences*, *115*(40), 9882–9888; Greitemeyer, T., & Mügge, D. O. (2014). Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play. *Personality and social psychology bulletin*, *40*(5), 578–589.

[72] Hilgard, J., Engelhardt, C. R., & Bartholow, B. D. (2013). Individual differences in motives, preferences, and pathology in video games: the gaming attitudes, motives, and experiences scales (GAMES). *Frontiers in Psychology*, *4*, 608.

exists in support of the short-term effects of violent video games on aggression, whereas pre-registered studies - which are not subject to this bias - have failed to find such effects.[73] Despite ongoing methodological discussions about the validity of results, the statistical link between violent video games and aggressive behaviour appears to be weak at best.[74] A recent meta-analysis also revealed that the relationship between violent games and aggressive or prosocial behaviour is negligible, with only small relationships found with aggressive affect and cognitions.[75]

In contrast, the relationship between violent games and desensitisation appears to be much stronger.[76] Interestingly, further research suggests that competition within games rather than violence may be more impactful in terms of provoking aggressive behaviour in gamers.[77] In addition, it appears that specific game features, such as levels of frustration or the pace of action, may contribute to the observed association between violent video game play and heightened aggression.[78]

Ultimately, the debate regarding the link between violent video games and real-life aggression has been ongoing and heavily discussed in the research community. While some scholars argue that violent video games may have negative effects on social behaviour, others maintain that the statistical relationship between violent video games and aggressive behaviour appears to be weak at best. This debate is thus not settled yet, and requires further research with reproducible design and control of different variables such as, e.g., frustration, pace, and competition.

---

[73] Hilgard, J., Engelhardt, C. R., Rouder, J. N., Segert, I. L., & Bartholow, B. D. (2019). Null effects of game violence, game difficulty, and 2D: 4D digit ratio on aggressive behavior. *Psychological Science*, *30*(4), 606–616; McCarthy, R. J., Coley, S. L., Wagner, M. F., Zengel, B., & Basham, A. (2016). Does playing video games with violent content temporarily increase aggressive inclinations? A pre-registered experimental study. *Journal of Experimental Social Psychology*, *67*, 13–19; Kühn, S., Kugler, D. T., Schmalen, K., Weichenberger, M., Witt, C., & Gallinat, J. (2019). Does playing violent video games cause aggression? A longitudinal intervention study. *Molecular Psychiatry*, *24*(8), 1220–1234.

[74] Mathur, M. B., & VanderWeele, T. J. (2019). Finding common ground in meta-analysis "wars" on violent video games. *Perspectives on Psychological Science, 14*(4), 705–708; López-Fernández, F. J., Mezquita, L., Etkin, P., Griffiths, M. D., Ortet, G., & Ibáñez, M. I. (2021). The role of violent video game exposure, personality, and deviant peers in aggressive behaviors among adolescents: A two-wave longitudinal study. *Cyberpsychology, Behavior, and Social Networking*, *24*(1), 32–40.

[75] Ferguson & Wang (2022).

[76] Lindner, D., Trible, M., Pilato, I., & Ferguson, C. J. (2020). Examining the effects of exposure to a sexualized female video game protagonist on women's body image. *Psychology of Popular Media*, *9*(4), 553.

[77] Dowsett, A., & Jackson, M. (2019). The effect of violence and competition within video games on aggression. *Computers in Human Behavior, 99*, 22–27.

[78] Devilly, G. J., O'Donohue, R. P., & Brown, K. (2023). Personality and frustration predict aggression and anger following violent media. *Psychology, Crime & Law*, *29*(1), 83–119; Elson, M., Breuer, J., Van Looy, J., Kneer, J., & Quandt, T. (2015). Comparing apples and oranges? Evidence for pace of action as a confound in research on digital games and aggression. *Psychology of Popular Media Culture*, *4*(2), 112.

12.4.5 Sexualisation and Violence Towards Female Characters

While video games can serve as a tool to promote diversity and inclusivity, the depiction of female characters has been a topic of debate. In this section, we will explore the potential influence of sexual objectification and virtual violence against female characters in video games on real-life attitudes towards women.

One relevant question is how the presence of sexualised female characters in video games can affect female players. Objectification theory[79] posits that frequent exposure to sexually objectifying media messages can socialise women to turn this sexualisation inward, engaging in self-objectification. This can have consequences on attitudes such as valuing one's body in terms of its appearance rather than its competence, thinking about one's appearance primarily from the perspective of others, and treating one's body as if it is capable of representing the self.[80] However, correlational research suggests that women who engage in self-objectification are more likely to objectify other women,[81] contributing to the cultural occurrence of sexualisation.

The first question to ask is thus the impact of exposure to sexualised female bodies on female players. Lindner et al.[82] investigated the consequences on female players' body satisfaction and aggression towards other women, finding mixed results. In their study, female players were randomly assigned to play either a more or less sexualised avatar in a Tomb Raider game. Exposure to a sexualised avatar in a video game did not significantly impact on participants' reported self-objectification and body dissatisfaction, as well as hostility and aggression towards other female confederates. These suggest that video games exposure to sexualised females may not have a substantial impact on female players.

Other interesting correlational research suggests that individuals who sexually objectify women are likely to hold negative attitudes towards them[83] and perceive them as less competent.[84] Therefore, it is necessary to examine the impact of exposing players to sexual objectification and virtual violence against females on players' real life attitudes.

Beck and Rose[85] conducted a study to investigate the potential link between exposure to sexual objectification and virtual violence against female characters in video games and negative attitudes towards women. Participants were randomly assigned to either a control group

---

[79] Fredrickson & Roberts (1997).

[80] Lindner, D., & Tantleff-Dunn, S. (2017). The development and psychometric evaluation of the Self-Objectification Beliefs and Behaviors Scale. *Psychology of Women Quarterly*, *41*(2), 254–272.

[81] Lindner, D., Tantleff-Dunn, S., & Jentsch, F. (2012). Social comparison and the 'circle of objectification'. *Sex roles*, *67*, 222–235; Strelan, P., & Hargreaves, D. (2005). Women who objectify other women: The vicious circle of objectification? *Sex Roles*, *52*.

[82] Lindner et al. (2020).

[83] Vaes, J., Cogoni, C., & Calcagnì, A. (2020). Resolving the human–object divide in sexual objectification: How we settle the categorization conflict when categorizing objectified and nonobjectified human targets. *Social Psychological and Personality Science*, *11*(4), 560–569/

[84] Johnson, V., & Gurung, R. A. (2011). Defusing the objectification of women by other women: The role of competence. *Sex Roles*, *65*, 177–188.

[85] Beck, V., & Rose, C. (2018). Is sexual objectification and victimization of females in video games associated with victim blaming or victim empathy? *Journal of Interpersonal Violence*. https://doi.org/10.1177/0886260518770187

that played Madden NFL 12 or an experimental group that played Grand Theft Auto, a game in which strong sexual violence against sexualised women can be carried out. The study found that initially, the group playing Grand Theft Auto did not show any significant difference in rape myth acceptance compared to the control group. However, the same experimental group showed a decrease in rape myth acceptance over time, and the effect became statistically significant by the end of the study, six months later. These findings suggest that exposure to sexual objectification and violence against female characters in video games may have a more complex and nuanced impact on attitudes towards women than previously believed.

Furthermore, Ferguson et al.[86] conducted a recent meta-analysis investigating the relationship between sexualisation in video games and players' mental well-being, as well as sexist and misogynistic attitudes and behaviours. The meta-analysis found that current evidence does not support concerns that sexualised game content negatively affects players' mental well-being or is positively associated with sexist and misogynistic attitudes and behaviours.

Therefore, the relationship between video games and attitudes towards women is multifaceted. While exposure to sexual objectification and violence against women in video games may not necessarily lead to negative attitudes towards women, there is still a societal need to improve the representation of women and female characters in games.

12.4.6 Cultural (Mis)Representation or Appropriation

Video games can be a platform to explore and celebrate different cultures, encouraging exchange and appreciation, but different cultures might also be misrepresented or appropriated by developers.

Culture is a complex concept, and can be grasped as a sum of knowledge, belief, customs, habits, and laws, in addition to other capabilities of humans in a given society.[87] Its misrepresentation or appropriation can lead to strong negative repercussions such as nonrecognition, misrecognition and exploitation.[88] However, Lenard and Balint[89] argue that there are not so many occurrences of serious wrongful cultural appropriation, but misrepresentation of cultures is still of importance. Thus, the representation of culture in video games is to be looked into closely to ensure respect of different and diverse cultures.

In video games, cultural representation is discussed as a heritage displayed through, e.g., architecture reproduction, people portrayal, the inclusion of artefacts, and language use.[90] In their study investigating and evaluating the cultural heritage and representation of muslims in Assassin's Creed 1 and Unearthed: Trail of Ibn battuta, Balela and Mundy[91] conclude that despite investing substantial time and resources into design, video game developers make

---

[86] Ferguson et al. (2022).

[87] Birukou, A., Blanzieri, E., Giorgini, P., and Giunchiglia, F. (2013). A formal definition of culture. In Sycara, K., Gelfand, M. & Abbe, A. (eds), *Models for Intercultural Collaboration and Negotiation*, Springer, pp. 1–26.

[88] Lalonde, D. (2021). Does cultural appropriation cause harm?. *Politics, Groups, and Identities*, *9*(2), 329–346.

[89] Lenard and Balint (2020).

[90] Balela, M. S., & Mundy, D. (2015). Analysing Cultural Heritage and its Representation in Video Games. In *DiGRA Conference* (pp. 1–16).

[91] Balela and Mundy (2015).

decisions that impact the accuracy of cultural representation in their games. Designers should thus improve their awareness of cultural production and give greater consideration of cultural impact in design to propose more informed products. In other words, not that all games should mirror reality, but that greater awareness leads to greater consideration.

On the other hand, games such as Nishan Shaman, a music video game based on Chinese ethnic minorities, are an example of appropriate cultural representation in video games.[92] The game is based on the Manchu shaman legend and integrates Manchu culture into its narrative to make it immersive with the shamanic culture. It provides a new way for a traditional ethnic minority culture to spread to younger generations and the general public through modern media.

With the aim to understand and represent cultures accurately in games, co-design should be implemented. It has already been done in other contexts to represent child protection and dementia sufferance for example,[93] but also in the case of design for location based games (LBGs) for cultural heritage.[94] This last study argues the importance of including communities and heritage professionals for contextual relevance and content validity in such game design, as well as promoting awareness and learning about intangible cultural heritage of, e.g., craftsmanship and artisanal technologies. This spread of culturally accurate representations is of importance as it might impact gamers' perception of the world surrounding them.

A systematic review of sixty-two studies identified to analyse behavioural-change, content understanding, knowledge acquisition, and perceptual impacts was carried out by Shliakhovchuk and Muñoz García.[95] The authors' findings suggest that video games have the potential to help to acquire cultural knowledge and develop intercultural literacy, socio-cultural literacy, cultural awareness, self-awareness, and the cultural understanding of different geopolitical spaces, to reinforce or weaken stereotypes, and to some extent facilitate the development of intercultural skills. Shliakhovchuk[96] designed Chuzme, an educational digital game designed to raise cultural self-awareness and acknowledge cultural biases. The author concludes that through the development of cultural literacy of gamers, their attitude towards, and communication with specific stereotypical groups such as migrants, or refugees can be ameliorated.

Thus, video games can promote cultural exchange and appreciation, but developers must avoid misrepresentation and appropriation. Co-design with communities and heritage professionals is important for accurate cultural representation, which can impact gamers' perception of the world. Accurate representation in video games can also help to develop intercultural

---

[92] Balela and Mundy (2015).

[93] Davies, R., & Flynn, R. Explicit and Implicit Narratives in the Co-Design of Videogames. *in Maragiannis, A. Final Paper/Proceedings of the Digital Research in the Humanities and Arts Conference, DRHA2014, London.* (p. 73).

[94] Koutsabasis, P., Partheniadis, K., Gardeli, A., ... & Filippidou, D. E. (2022). Co-designing the user experience of location-based games for a network of museums: Involving cultural heritage professionals and local communities. *Multimodal Technologies and Interaction, 6*(5), 36.

[95] Shliakhovchuk, O., & Muñoz García, A. (2020). Digital game-based learning for D&I: conceptual design of an educational digital game Chuzme. In: *Proceedings INNODOCT/19. International Conference on Innovation, Documentation and Education*.

[96] Shliakhovchuk, O. (2019). *Cultural literacy acquisition through video game environments of a digitally born generation* (Doctoral dissertation, Universitat Politècnica de València).

literacy, awareness, and understanding. As video games gain popularity and cultural representation becomes more important, developers should strive to be more aware of cultural production and consider the cultural impact of their design decisions.

To conclude, in regards to violence and aggressivity, sexualisation and violence towards women, and cultural appropriation and misrepresentation, the empirical discourse is still at a discussion stage. Indeed, the transfer effects of morality from in-game behaviour and context to social real life behaviour should be viewed with great caution. Behaviours in video games and their possible impact on real life attitudes in specific cases have now been discussed, and lead us to the final layer of our overview; the gaming culture as a whole.

12.5 GAMING CULTURE AS A WHOLE

Although video games have been at the forefront of various recurrent, and controversial debates, such as aggression, violence,[97] and addiction.[98] There are various benefits to engaging with video games, including the opportunity for social connection to players and the community, as demonstrated throughout the covid pandemic lockdowns.[99] The video gaming domain captures a diverse population thus, promoting socialisation across various demographic groups.[100] In this first part, we will review what it entails to identify as a gamer.

12.5.1 Identifying as a Gamer

Individuals within the video game community have been categorised in two primary ways – as players or gamers. As discussed by Grooten and Kowert,[101] in the simplest terms, 'gamer' has been used to categorise anyone who plays video games, whilst others have adopted 'gamer' as an aspect of social identity. The former definition, which simply refers to anybody who plays video games, is questionable when considering that not all players identify as gamers. This is expected as 'gamer' is steeped in negative societal conceptions.[102] The simplistic construction of 'gamer' fades when considering the distinctions between players and gamers, as the gaming domain extends beyond simply playing – encapsulating multiple levels of user, and non-user, engagement.

The level of personal significance can distinguish the terms player and gamer. There is also a differentiation between the short-term and long-term effects of socialisation through gaming in that a player is a temporary, functional status obtained while playing a digital game

---

[97] Ferguson (2007); Anderson et al. (2010); Huesmann (2010).

[98] Griffiths et al., 2012; Bean et al. (2017).

[99] Barr, M., & Copeland-Stewart, A. (2022). Playing video games during the COVID-19 pandemic and effects on players' well-being. *Games and Culture, 17*(1), 122–139.

[100] Kowert, R. (2020). Dark Participation in Games. *Frontiers In Psychology, 11*. doi: 10.3389/fpsyg.2020.598947.

[101] Grooten, J., & Kowert, R. (2015). Going beyond the game: Development of gamer identities within societal discourse and virtual spaces. *Loading, 9*, 70–87. https://journals.sfu.ca/loading/index.php/loading/article/view/151.

[102] Hoffswell, J. (2017). Factoring in gamer identity: the application of social identity theory and flow to understanding video game violence effects. *2017 MU dissertations.*

(Grooten & Kowert, 2015). Subsequently, an individual is considered a player in the first instance of game interaction. In contrast, 'gamer' is a concept comprising the longitudinal aspects of self-construction, self-perception, individual societal and cultural positioning.[103]

Critically, gaming culture is frequently considered toxic culture due to the exclusion of individuals who do not adhere to a certain 'mould'[104] reflected in the outdated, stereotypical characteristics assigned to 'gamers'. This is further reflected in the abundance of video games containing age, ethnicity, and gender inequalities – they often feature narrow and dated ideas of femininity, are male-only, or are degrading towards females.[105]

12.5.2 Gamergate or a Tale of Toxicity Towards Women

Gaming culture has historically cultivated prejudicial attitudes and negative-gender based stereotypes towards female players and gender-based hostility is evident across the community.[106] The societal importance of these issues is evident when considering the harassment campaign coined 'Gamergate'[107] and the anti-discrimination lawsuit against Activision Blizzard.[108] The Gamergate discourse emphasised issues of ethics in journalism and the oppressive, sexist portrayals of female characters in video games, following several reports of harassment against women journalists and game designers.[109] The movement encouraged broader discussions of privilege and identity politics across the gaming domain.[110] But findings from Buyukozturk et al.[111] demonstrated how within gaming communities, specifically those within Reddit (a popular online forum with a large presence of gaming subgroups), the nature of discussions created barriers and facilitated the breakdown of collective action efforts stemming from the movement. Since then, 'anti-feminist' gamers continue to contest those who support the development of progressive and inclusive gaming content.

The campaign against Activision Blizzard exposed that not only do female players within the gaming community face hostility, but female employees within the industry are also targets of discrimination. The company was found to foster a sexist culture, in which women experienced instances of sexual misconduct, harassment, threats, and were paid far less than their male colleagues (Winslow & Gurwin, 2021), emphasising widespread discriminatory behaviours targeted at women across the domain. Research from Kivijärvi and Katila (2021) has crucial links to this issue, suggesting women within the gaming industry submit to the hegemonic gamer discourse, reproducing the masculine gamer notions in order to gain recognition as viable members of the community. This highlights just some of the many consequences of

---

[103] Grooten & Kowert (2015).

[104] Kowert (2020).

[105] Lavigne, C. (2015). "I'm batman" (and you can be too): Gender and constructive play in the Arkham game series. *Cinema Journal, 55*, 133–168

[106] Hussain & Griffiths (2008); Maclean (2016); McLean & Griffiths (2013); (2019).

[107] Dewey, C. (2014, October 14). The only guide to Gamergate you will ever need to read. *The Washington Post*; Romano, A. (2021, January 7). What we still haven't learned from Gamergate.

[108] Winslow & Gurwin (2021).

[109] Ferguson & Glasgow (2021).

[110] Braithwaite (2016).

[111] Buyukozturk et al. (2018).

a masculine-superordinate gaming culture, as players with diverse gender identities are arguably forced to align themselves with the masculine ideal, or face hostility.

A data leak from the streaming site Twitch depicted that the issues women face within the domain extend beyond verbal stigmatisation. The unprecedented leak posted vast amounts of data from Twitch, displaying the disproportion of revenue between gaming streamers. The leak revealed no female streamers were in the top third of highest earners and, overall, women only made up 16% of the top 1,000 streamers on the site.[112] These instances highlight significant issues within the community, which are a product of the culture the video game domain has cultivated.

12.5.3 A Stereotype to Break and a Culture to Change

Despite the diverse population of players, and the extensive presence of women engagement with video games, the industry and culture remain fixated on a masculine gamer identity.[113] This is an ongoing challenge that researchers face. The stereotypical characteristics of 'gamers' have been historically considered as 'mostly young, mostly nerdy and most definitely male'[114] and they are reinforced through ridicule and satire across media.[115] In contrast to this stereotype, the average player is 35–44 years old, and women represent a significant proportion of players. Lessening this juxtaposition is crucial, particularly as identity is a catalyst for gender-based discrimination across the domain.[116] Although the community has somewhat shifted to separate from the outdated, stereotypical representation of gamers, new labels adopted within gaming communities have since spurred separation and in-turn, hostility between players.

For example, discussions of 'hardcore' and 'casual' gamers are frequently used across research[117] and have gained a widespread presence in gaming communities. Critically, male players are more likely to align with the characteristics of hardcore players.[118] The impact of players' gender in these typologies is crucial, as these labels have since been used by some players to distinguish 'real' gamers (i.e., hardcore, male players) from 'fake' (i.e., casual, female players) gamers – further perpetuating negative gender-based stereotypes within the culture.[119] Moreover, categorisations such as 'gamer girls' have quickly become embedded in the domain as a term that sexualises, stigmatises, and stereotypes female players.[120] 'Gamer

---

[112] Asarch, S. (2021, October 8). 10 Biggest Revelations from The Unprecedented Twitch Leak. *Inverse.*

[113] Kafai, Y. B., Heeter, C., Denner, J., & Sun, J. Y. (2008). *Beyond Barbie and Mortal Kombat: New Perspectives on Gender and Gaming.* Cambridge, Massachusetts: The MIT Press.

[114] Harwell, D. (2014, Oct 18). More women play video games than boys, and other surprising facts lost in the mess of gamergate. *The Washington Post*, 17.

[115] Kowert, R., Griffiths, M. D., & Oldmeadow, J. A. (2012). Geek or chic? Emerging stereotypes of online gamers. *Bulletin of science, technology & society, 32*(6), 471–479.

[116] De Grove et al. (2015); Kaye & Pennington (2016); Powell & Kaye (2018).

[117] Morin et al. (2016); Kapalo et al. (2015); Shaw (2012); Kuittinen et al. (2007); Fritsch et al. (2006); Ip & Jacobs (2005).

[118] Kapalo et al. (2015); Phan et al. (2012).

[119] Paaßen et al. (2017); Vanderhoef (2013); Harrison et al.(2016).

[120] Harrison et al.(2016); Turner, (2021).

girl' has connotations that women players are less skilled, less engaged gamers who do not belong in the domain, implying they 'aren't real gamers'.

12.5.4 Changing the Game to Change the Gamer

Representation in video games is an on-going issue with various factors to consider, including the representation of industry figures, video game protagonists, accessibility of games, and the audiences games are targeted towards. Although communities, developers and organisations have moved towards an inclusive representation of players, further work is necessary to combat the historical connotations associated with 'gamers' and the video game community.

Various aspects of gaming allow players to explore their identity.[121] Games that create playful spaces, allowing for meaningful exploration, can provide positive experiences and safe spaces for players' to explore their gender identity.[122] The ability to 'gender swap' through avatar creation is one example of how identity can be explored diversely through digital games,[123] where players have used this opportunity to emphasise the appearance of their avatar. The significance of representation within video games is well-demonstrated when exploring research on avatars.[124] That is, avatars are an important aspect of identity embodiment and expression for players, and require a significant shift to improve inclusion of non-white, non-cis, non-male players.[125] If current avatars lack crucial representation of all individuals encompassed within the video game community, then the benefits of embodiment and expression are accessible only to those represented.

Significant work is being undertaken by communities within video games, to shift the culture. Crucial work by organisations such as Women in Games, GaymerX, and TakeThis.org are creating positive, inclusive discourses within the community to shift the culture and represent all players. Although the gaming community sees active groups in its core, it is crucial to consider the role of games themselves. Despite the positive shift, groups and individuals within the community still engage and promote negative stereotypes and hostility towards players who have been underrepresented by the culture. Considering the representation of all players in games is crucial in order to shift the culture, and thus to change individual behaviour.

---

[121] Bessière et al (2007); Wonica (2014).

[122] George, L. (2021, October). Investigating the Role of Technology in Supporting Exploration of Gender Identity Through Games and Play. In *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play* (pp. 399-400). doi: 10.1145/3450337.3483516.

[123] Paik, P., & Shi, C. (2013). Playful gender swapping: user attitudes toward gender in MMORPG avatar customisation. *Digital Creativity, 24*(4), 310–326. https:// doi .org/ 10 .1080/ 14626268.2013.767275.

[124] Yoon & Vargas (2014); McMenomy (2011); Eklund (2011).

[125] Euteneuer, J. S. (2016). Default characters and the embodied nature of play: Race, gender, and gamer identity. *Press Start, 3*(1), 115–125. https://press-start.gla.ac.uk/index.php/press-start/article/view/50.

12.6 CONCLUSION

In conclusion, video game ethics is a complex topic that requires a multidisciplinary approach to fully understand. From a psychology-driven perspective, we have discussed the impact of video games on mental health, cognitive skills, social behaviours, and the overall gaming culture. We have shown that video games have the potential to enhance cognitive and social skills in players, and the diverse nature of game mechanics allows for the development of various skills, which can be translated into other areas of players' lives, including education.

However, there are still ongoing debates and controversies surrounding the impact of video games on players. For example, while the effects of video games on well-being may be small, cumulative effects over time could still have a significant positive impact on gamers, and video games have the potential to be used as readily-available tools for promoting mindfulness, teaching positive behaviours for one's physical health, and improving gamers' overall well-being. On the other hand, there are concerns about the potential negative impacts of video games on players, such as addiction and dark patterns manipulations.

In terms of the impact of video games on pro- or anti-social real-life behaviour, the relationship between video games and attitudes towards violence, sexism, and cultural appropriation is complex and multifaceted. While some studies suggest a link between violent video games and real-life aggression, others argue that the relationship is weak at best. Similarly, exposure to sexual objectification and violence against female characters in video games may not necessarily lead to negative attitudes towards women, but there is still a societal need to improve the representation of women and female characters in games.

Finally, we have discussed the role of the gaming culture as a whole. While there have been positive shifts towards greater representation and inclusivity in the gaming community, there are still groups and individuals who engage in negative stereotypes and hostility towards players who have been underrepresented by the culture. Therefore, it is crucial to consider the representation of all players in games in order to shift the culture and change individual behaviour.

In summary, video game ethics is a complex and ongoing discussion that requires continued multidisciplinary research and collaboration to fully understand the impacts of video games on players and society as a whole. While video games have the potential to enhance cognitive and social skills and promote positive behaviours, there are also concerns about potential negative impacts on players. As such, it is important to continue to critically examine the impact of video games on players and the culture surrounding them to ensure that they are used in ways that benefit individuals and society as a whole.

REFERENCES

Aarseth, E., Bean, A. M., Boonen, H., Colder Carras, M., Coulson, M., Das, D., ... & Van Rooij, A. J. (2017). Scholars' open debate paper on the World Health Organization ICD-11 Gaming Disorder proposal. *Journal of behavioral addictions*, *6*(3), 267–270.

Abby, A. (2022, December 8). *The Complete History of Tennis for Two*. History-Computer. https://history-computer.com/tennis-for-two-complete-history/.

Adachi, P. J., & Willoughby, T. (2011). The effect of video game competition and violence on aggressive behavior: Which characteristic has the greatest influence?. *Psychology of violence, 1*(4), 259.

Aligeng. (2021). A Study on the Spread of Ethnic Minority Traditional Legends in the New Era: Taking the Video Game 'Nishan Shaman' as an Example. In *Proceedings of the 2nd International Conference on Language, Art and Cultural Exchange (ICLACE 2021), Advances in Social Science, Education and Humanities Research, Volume 559*.

American Psychiatric Association (APA) (2013). *DSM-5 Classification*. American Psychiatric Publishing.

Anderson, C. (2003). The psychology of doing nothing: Forms of decision avoidance result from reason and emotion. *Psychological Bulletin*, *129*(1), 139–167. https://doi.org/10.1037/0033-2909.129.1.139.

Anderson, C., Shibuya, A., Ihori, N., Swing, E., Bushman, B., & Sakamoto, A. et al. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries: A meta-analytic review. *Psychological Bulletin, 136*(2), 151–173. doi: 10.1037/a0018251.

Asarch, S. (2021, October 8). 10 Biggest Revelations from The Unprecedented Twitch Leak. *Inverse.* https://www.inverse.com/gaming/twitch-leak-hack-data-breach-streamer-payout-earnings.

Balela, M. S., & Mundy, D. (2015). Analysing Cultural Heritage and its Representation in Video Games. In *DiGRA Conference* (pp. 1–16).

Barr, S. (2019, October 8). Children with gaming addiction to be offered treatment by NHS. *The Independent*. https://www.independent.co.uk/life-style/health-and-families/gaming-addiction-nhs-treatment-children-video-games-health-a9146136.html.

Barr, M., & Copeland-Stewart, A. (2022). Playing video games during the COVID-19 pandemic and effects on players' well-being. *Games and Culture*, *17*(1), 122–139. DOI: 10.1177/15554120211017036.

Bandura, A., & Walters, R. H. (1977). *Social learning theory* (Vol. 1). Prentice Hall: Englewood Cliffs.

Bean, A. M., Nielsen, R. K. L., van Rooij, A. J., & Ferguson, C. J. (2017). Video game addiction: The push to pathologize video games. *Professional Psychology: Research and Practice, 48*(5), 378–389. https://doi.org/10.1037/pro0000150.

Beck, V., & Rose, C. (2018). Is sexual objectification and victimization of females in video games associated with victim blaming or victim empathy? *Journal of Interpersonal Violence*. https://doi.org/10.1177/0886260518770187

Bessière, K., Seay, A., & Kiesler, S. (2007). The Ideal Elf: Identity Exploration in World of Warcraft. *Cyberpsychology & Behavior, 10*(4), 530–535. https://doi.org/10.1089/cpb.2007.9994.

Birukou, A., Blanzieri, E., Giorgini, P., and Giunchiglia, F. (2013). A formal definition of culture. In Sycara, K., Gelfand, M. & Abbe, A. (eds), *Models for Intercultural Collaboration and Negotiation*, Springer, pp. 1–26.

Borges, G., Orozco, R., Benjet, C., Mart´Inez, K. I. M., Contreras, E. V., P´Erez, A. L. J., Cedr´Es, A. J. P., Uribe, P. C. H., Couder, M. a. C. D., Gutierrez-Garcia, R. U. A., Ch´Avez, G. E. Q., Albor, Y., Mendez, E., Medina-Mora, M. E., Mortier, P., & Ayuso-Mateos, J. L. (2020). (Internet) Gaming Disorder in DSM-5 and ICD-11: A Case of the Glass Half Empty or Half Full: (Internet) Le trouble du jeu dans le DSM-5 et la CIM-11: Un cas de verre à moitié vide et à moitié plein. The Canadian Journal of Psychiatry, 66(5), 477–484. https://doi.org/10.1177/0706743720948431.Bowman, N. D., Ahn, S. J., & Mercer Kollar, L. M. (2020). The paradox of interactive media: The potential for video games and virtual reality as tools for violence prevention. *Frontiers in communication*, 104.

Boyan, A., GrizzArd, M., & Bowman, N. (2015). A massively moral game? Mass Effect as a case study to understand the influence of players' moral intuitions on adherence to hero or antihero play styles. *Journal of Gaming & Virtual Worlds*, *7*(1), 41–57.

Brady, A., & Prentice, G. (2021). Are loot boxes addictive? Analyzing participant's physiological arousal while opening a loot box. *Games and Culture*, *16*(4), 419–433.

Braithwaite, A. (2016). It's about ethics in games journalism? Gamergaters and geek masculinity. *Social Media and Society, 2*(4). https://doi.org/10.1177/2056305116672484.

Buyukozturk, B., Gaulden, S., & Dowd-Arrow, B. (2018). Contestation on Reddit, Gamergate, and movement barriers. *Social Movement Studies*, *17*(5), 592–609. https://doi.org/10.1080/14742837.2018.1483227.

Cardoso-Leite, P., & Bavelier, D. (2014). Video game play, attention, and learning: how to shape the development of attention and influence learning?. *Current opinion in neurology*, *27*(2), 185–191.

Cohen, J. (2001). Defining identification: A theoretical look at the identification of audiences with media characters. *Mass Communication and Society, 4*(3), 245–264. https://doi.org/10.1207/S15327825MCS0403_01

Coknaz, D., Mirzeoglu, A. D., Atasoy, H. I., Alkoy, S., Coknaz, H., & Goral, K. (2019). A digital movement in the world of inactive children: favourable outcomes of playing active video games in a pilot randomized trial. *European Journal of Pediatrics, 178*, 1567–1576.

Cruea, M. D. (2020). Gaming the mind and minding the game: mindfulness and flow in video games. *Video Games and Well-being: Press Start,* 97–107.

Csikszentmihalyi, M., & Nakamura, J. (2010). Effortless attention in everyday life: A systematic phenomenology. *Effortless attention: A new perspective in the cognitive science of attention and action*, 179–189.

Darvesh, N., Radhakrishnan, A., Lachance, C. C., Nincic, V., Sharpe, J. P., Ghassemi, M., ... & Tricco, A. C. (2020). Exploring the prevalence of gaming disorder and Internet gaming disorder: a rapid scoping review. *Systematic reviews*, *9*, 1–10.

Davies, R., & Flynn, R. Explicit and Implicit Narratives in the Co-Design of Videogames. *in Maragiannis, A. Final Paper/Proceedings of the Digital Research in the Humanities and Arts Conference, DRHA2014, London.* (p. 73).

Davnall, R. (2021). What does the gamer do?. *Ethics and Information Technology*, *23*(3), 225–237.

De Grove, F., Courtois, C., & Looy, V. J. (2015). How to be a gamer! Exploring personal and social indicators of gamer identity. *Journal of Computer-Mediated Communication, 20*, 346–361. doi:10.1111/jcc4.12114

Decety, J., & Cowell, J. M. (2014). The complex relation between morality and empathy. *Trends in Cognitive Sciences, 18*(7), 337–339

Deterding, C. S., Stenros, J., & Montola, M. (2020, January). Against 'Dark Game Design Patterns'. In *DiGRA'20-Abstract Proceedings of the 2020 DiGRA International Conference*. York.

Devilly, G. J., O'Donohue, R. P., & Brown, K. (2023). Personality and frustration predict aggression and anger following violent media. *Psychology, Crime & Law*, *29*(1), 83–119.

Dewey, C. (2014, October 14). The only guide to Gamergate you will ever need to read. *The Washington Post*. https://www.washingtonpost.com/news/the-intersect/wp/2014/10/14/the-only-guide-to-gamergate-you-will-ever-need-to-read/.

Dolgov, I., Graves, W. J., Nearents, M. R., Schwark, J. D., & Volkman, C. B. (2014). Effects of cooperative gaming and avatar customization on subsequent spontaneous helping behavior. *Computers in Human Behavior, 33*, 49–55.

Dowsett, A., & Jackson, M. (2019). The effect of violence and competition within video games on aggression. *Computers in Human Behavior, 99*, 22–27.

Eklund, L. (2011). Doing gender in cyberspace: The performance of gender by female World of Warcraft players. Convergence: *The International Journal of Research into New Media Technologies, 17*(3), 323–342. https://doi.org/10.1177/1354856511406472

Elson, M., Breuer, J., Van Looy, J., Kneer, J., & Quandt, T. (2015). Comparing apples and oranges? Evidence for pace of action as a confound in research on digital games and aggression. *Psychology of Popular Media Culture*, *4*(2), 112.

Ethical Games. (2020, December 3). Ethical Games. http://ethicalgames.org/

Euteneuer, J. S. (2016). Default characters and the embodied nature of play: Race, gender, and gamer identity. *Press Start, 3*(1), 115–125. https://press-start.gla.ac.uk/index.php/press-start/article/view/50.

Feng, W., Ramo, D. E., Chan, S. R., & Bourgeois, J. A. (2017). Internet gaming disorder: Trends in prevalence 1998–2016. *Addictive Behaviors, 75*, 17–24.

Ferguson, C. (2007). Evidence for publication bias in video game violence effects literature: A meta-analytic review. *Aggression and Violent Behavior, 12*(4), 470–482. doi: 10.1016/j.avb.2007.01.001.

Ferguson, C. J. (2015). Do angry birds make for angry children? A meta-analysis of video game influences on children's and adolescents' aggression, mental health, prosocial behavior, and academic performance. *Perspectives on psychological science*, *10*(5), 646–666.

Ferguson, C., & Glasgow, B. (2021). Who is GamerGate? A descriptive study of individuals involved in the GamerGate controversy. *Psychology of Popular Media, 10*(2), 243–247. https://doi.org/10.1037/ppm0000280

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition.* McGraw-Hill Book Company.

Fredrickson, B. L., Roberts, T. A., Noll, S. M., Quinn, D. M., & Twenge, J. M. (1998). That swimsuit becomes you: sex differences in self-objectification, restrained eating, and math performance. *Journal of Personality and Social Psychology*, *75*(1), 269.

Fritsch, T., Voigt, B., & Schiller, J. (2006). Distribution of online hardcore player behavior. *Proceedings Of 5Th ACM SIGCOMM Workshop on Network and System Support for Games – Netgames '06*. https://doi.org/10.1145/1230040.1230082.

Frostling-Henningsson, M. (2009). First-person shooter games as a way of connecting to people: 'Brothers in blood'. *Cyberpsychology & Behavior, 12*(5), 557–562.

Game for Change Festival. (n.d.) *Join the 20th Games for Change Festival!* https://festival.gamesforchange.org/.

Gentile, D. (2017, March 28). Researchers find video games influence sexist attitudes News Service Iowa State University. *News Iowa State University*. https://www.news.iastate.edu/news/2017/03/28/gamessexism.

Gentile, D. A., & Gentile, J. R. (2008). Violent video games as exemplary teachers: A conceptual analysis. *Journal of Youth and Adolescence*, *37*, 127–141.

George, L. (2021, October). Investigating the Role of Technology in Supporting Exploration of Gender Identity Through Games and Play. In *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play* (pp. 399-400). doi: 10.1145/3450337.3483516.

Gleeson, J. (2014). Vitriolic abuse of Anita Sarkeesian: why the games industry needs her. *The Conversation*. https://theconversation.com/vitriolic-abuse-of-anita-sarkeesian-why-the-games-industry-needs-her-31826.

Greene, J. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Greitemeyer, T. (2022). The dark and bright side of video game consumption: Effects of violent and prosocial video games. *Current Opinion in Psychology*, 101326.

Greitemeyer, T., & Mügge, D. O. (2014). Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play. *Personality and Social Psychology Bulletin*, *40*(5), 578–589.

Griffiths, M. D., Kuss, D. J., & King, D. L. (2012). Video Game Addiction: Past, Present and Future. *Current Psychiatry Reviews, 8*(4), 308–318. doi: 10.2174/157340012803520414.

Grohmann, L., Holl, E., & Melzer, A. (2021). Moral Judgment in Video Games: Effects of Medium, Moral Intuitions and Media-Based Empathy. In *12th Media Psychology Conference (MediaPsych 2021)*, p. 100.

Grooten, J., & Kowert, R. (2015). Going beyond the game: Development of gamer identities within societal discourse and virtual spaces. *Loading..., 9*, 70–87. https://journals.sfu.ca/loading/index.php/loading/article/view/151.

Haidt, J. (2012). *The Righteous Mind: Why Good People are Divided by Politics and Religion*. Penguin UK.

Harrison, R. L., Drenten, J., & Pendarvis, N. (2016). Gamer girls: navigating a subculture of gender inequality. In *Consumer Culture Theory* (Vol. 18, pp. 47–64). Emerald Group Publishing Limited.

Harwell, D. (2014, Oct 18). More women play video games than boys, and other surprising facts lost in the mess of gamergate. *The Washington Post*, 17. https://www.washingtonpost.com/news/the-switch/wp/2014/10/17/more-women-play-video-games-than-boys-and-other-surprising-facts-lost-in-the-mess-of-gamergate/.

Hern, A. (2020, July 22). Playing video games doesn't lead to violent behaviour, study shows. *The Guardian*. https://www.theguardian.com/games/2020/jul/22/playing-video-games-doesnt-lead-to-violent-behaviour-study-shows.

Hilgard, J., Engelhardt, C. R., & Bartholow, B. D. (2013). Individual differences in motives, preferences, and pathology in video games: the gaming attitudes, motives, and experiences scales (GAMES). *Frontiers in Psychology, 4*, 608.

Hilgard, J., Engelhardt, C. R., & Rouder, J. N. (2017). Overstated evidence for short-term effects of violent games on affect and behavior: A reanalysis of Anderson et al.(2010).

Hilgard, J., Engelhardt, C. R., Rouder, J. N., Segert, I. L., & Bartholow, B. D. (2019). Null effects of game violence, game difficulty, and 2D: 4D digit ratio on aggressive behavior. *Psychological Science*, *30*(4), 606–616.

Hoffswell, J. (2017). Factoring in gamer identity: the application of social identity theory and flow to understanding video game violence effects. *2017 MU dissertations.* https://mospace.umsystem.edu/xmlui/handle/10355/65507.

Holl, E., & Melzer, A. (2022). Moral minds in gaming: A quantitative case study of moral decisions in Detroit: Become human. *Journal of Media Psychology: Theories, Methods, and Applications, 34*(5), 287–298. https://doi.org/10.1027/1864-1105/a000323.

Hong, J. C., Cheng, C. L., Hwang, M. Y., Lee, C. K., & Chang, H. Y. (2009). Assessing the educational values of digital games. *Journal of Computer Assisted Learning, 25*(5), 423–437.

Howarth, J. (2023, January 18). *How Many Gamers Are There? (New 2023 Statistics)*. Exploding Topics. https://explodingtopics.com/blog/number-of-gamers.

Huesmann, L. (2010). Nailing the coffin shut on doubts that violent video games stimulate aggression: Comment on Anderson et al. (2010). *Psychological Bulletin, 136*(2), 179–181. doi: 10.1037/a0018567.

Hussain, Z., & Griffiths, M. D. (2008). Gender swapping and socialising in cyberspace: An exploratory study. *Cyberpsychology and Behavior, 11*(1), 47–53. doi: 10.1089/cpb.2007.0020.

International Game Developers Association (IGDA). (n.d.). *Code of Ethics*. https://members.igda.org/page/codeofethics.

Ip, B., & Jacobs, G. (2005). Segmentation of the games market using multivariate analysis. *Journal of Targeting, Measurement and Analysis for Marketing, 13*(3), 275–287. https://doi.org/10.1057/palgrave.jt.5740154.

James, P. (2023, March 6). *Most Played & Most Popular Games In The World (2022–2023)*. Gamer Tweak. https://gamertweak.com/most-played-popular-games/.

Joeckel, S., Bowman, N. D., & Dogruel, L. (2012). Gut or game? The influence of moral intuitions on decisions in video games. *Media psychology*, *15*(4), 460–485.

Joeckel, S., Bowman, N. D., & Dogruel, L. (2013). The influence of adolescents' moral salience on actions and entertainment experience in interactive media. *Journal of Children and Media*, *7*(4), 480–506.

Johannes, N., Vuorre, M., & Przybylski, A. K. (2021). Video game play is positively correlated with well-being. *Royal Society Open Science, 8*(2), 202049.

Johnson, V., & Gurung, R. A. (2011). Defusing the objectification of women by other women: The role of competence. *Sex Roles*, *65*, 177–188.

Kafai, Y. B., Heeter, C., Denner, J., & Sun, J. Y. (2008). *Beyond Barbie and Mortal Kombat: New Perspectives on Gender and Gaming.* Cambridge, Massachusetts: The MIT Press.

Kapalo, K., Dewar, A., Rupp, M., & Szalma, J. (2015). Individual Differences in Video Gaming. *Proceedings Of the Human Factors and Ergonomics Society Annual Meeting, 59*(1), 878–881. https://doi.org/10.1177/1541931215591261.

Kaye, L. K., & Pennington, C.R. (2016). 'Girls can't play': The Effects of Stereotype Threat on Females' Gaming Performance. *Computers in Human Behavior, 59*, 202–209. https://doi.org/10.1016/j.chb.2016.02.020.

Katsarov, J., Christen, M., Mauerhofer, R., Schmocker, D., & Tanner, C. (2019). Training moral sensitivity through video games: A review of suitable game mechanisms. *Games and Culture*, *14*(4), 344–366.

Kivijärvi, M., & Katila, S. (2021). Becoming a Gamer: Performative Construction of Gendered Gamer Identities. *Games And Culture*, 1-21. https://doi.org/10.1177/15554120211042260.

Koivula, M., Huttunen, K., Mustola, M., Lipponen, S., & Laakso, M. L. (2017). The emotion detectives game: Supporting the social-emotional competence of young children. *Serious games and edutainment applications: Volume II,* 29–53.

Korman, J., Voiklis, J., & Malle, B. F. (2015). The social life of cognition. *Cognition, 135*, 30–35.

Koutsabasis, P., Partheniadis, K., Gardeli, A., Vogiatzidakis, P., Nikolakopoulou, V., Chatzigrigoriou, P., ... & Filippidou, D. E. (2022). Co-Designing the User Experience of Location-Based Games for a Network of Museums: Involving Cultural Heritage Professionals and Local Communities. *Multimodal Technologies and Interaction*, *6*(5), 36.

Kowert, R. (2015). Social outcomes: Online game play, social currency, and social ability. In *The Video Game Debate* (pp. 94–115). Routledge. https://doi.org/10.4324/9781315736495-6

Kowert, R. (2020). Dark Participation in Games. *Frontiers In Psychology, 11*. doi: 10.3389/fpsyg.2020.598947.

Kowert, R., Griffiths, M. D. & Oldmeadow, J. A. (2012). Geek or chic? Emerging stereotypes of online gamers. *Bulletin of Science Technology & Society, 32*(6), 471–479. https://doi.org/10.1177/0270467612469078.

Krcmar, M., & Eden, A. (2019). Rational versus intuitive processing: The impact of cognitive load and moral salience on in-game aggression and feelings of guilt. *Journal of Media Psychology: Theories, Methods, and Applications*, *31*(1), 2.

Kühn, S., Kugler, D. T., Schmalen, K., Weichenberger, M., Witt, C., & Gallinat, J. (2019). Does playing violent video games cause aggression? A longitudinal intervention study. *Molecular Psychiatry*, *24*(8), 1220–1234.

Kuittinen, J., Kultima, A., Niemelä, J., & Paavilainen, J. (2007). Casual games discussion. *Proceedings of the 2007 Conference on Future Play – Future Play '07*. https://doi.org/10.1145/1328202.1328221.

Lalonde, D. (2021). Does cultural appropriation cause harm?. *Politics, Groups, and Identities*, *9*(2), 329–346.

Lavigne, C. (2015). 'I'm batman' (and you can be too): Gender and constructive play in the Arkham game series. *Cinema Journal, 55*, 133–168. doi:10.1353/CJ.2015.0069.

Lenard, P. T., & Balint, P. (2020). What is (the wrong of) cultural appropriation?. *Ethnicities*, *20*(2), 331–352.

Lie, A., Stephen, A., Supit, L. R., Achmad, S., & Sutoyo, R. (2022, October). Using Strategy Video Games to Improve Problem Solving and Communication Skills: A Systematic Literature Review. In *2022 4th International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1–5). IEEE.

Lindner, D., & Tantleff-Dunn, S. (2017). The development and psychometric evaluation of the Self-Objectification Beliefs and Behaviors Scale. *Psychology of Women Quarterly*, *41*(2), 254–272.

Lindner, D., Tantleff-Dunn, S., & Jentsch, F. (2012). Social comparison and the 'circle of objectification'. *Sex roles*, *67*, 222–235.

Lindner, D., Trible, M., Pilato, I., & Ferguson, C. J. (2020). Examining the effects of exposure to a sexualized female video game protagonist on women's body image. *Psychology of Popular Media*, *9*(4), 553.

López-Fernández, F. J., Mezquita, L., Etkin, P., Griffiths, M. D., Ortet, G., & Ibáñez, M. I. (2021). The role of violent video game exposure, personality, and deviant peers in aggressive behaviors among adolescents: A two-wave longitudinal study. *Cyberpsychology, Behavior, and Social Networking*, *24*(1), 32–40.

Luck, M. (2009). The gamer's dilemma: An analysis of the arguments for the moral distinction between virtual murder and virtual paedophilia. *Ethics and Information Technology, 11*(1), 31–36. https://doi.org/10.1007/s10676-008-9168-4

Luhmann, M., Krasko, J., & Terwiel, S. (2021). Subjective well-being as a dynamic construct. *Elsevier EBooks*, 1231–1249. https://doi.org/10.1016/b978-0-12-813995-0.00048-0.

Maclean, E. (2016). Girls, guys, and games: How news media perpetuate stereotypes of male and female gamers. *Press Start: Negotiating Gamer Identities, 3*(1), 17–45. http://orcid.org/0000-0002-6468-6174.

Madigan, J. (2016). *Getting Gamers: The Psychology of Video Games and Their Impact on the People who Play Them*. Rowman & Littlefield Publishers.

Manzano-León, A., Camacho-Lazarraga, P., Guerrero, M. A., Guerrero-Puerta, L., Aguilar-Parra, J. M., Trigueros, R., & Alias, A. (2021). Between level up and game over: A systematic literature review of gamification in education. *Sustainability*, *13*(4), 2247.

Maroney, N., Williams, B. J., Thomas, A., Skues, J., & Moulding, R. (2019). A stress-coping model of problem online video game use. *International Journal of Mental Health and Addiction, 17*, 845–858.

Mathur, M. B., & VanderWeele, T. J. (2019). Finding common ground in meta-analysis 'wars' on violent video games. *Perspectives on Psychological Science, 14*(4), 705–708.

McCarthy, R. J., Coley, S. L., Wagner, M. F., Zengel, B., & Basham, A. (2016). Does playing video games with violent content temporarily increase aggressive inclinations? A pre-registered experimental study. *Journal of Experimental Social Psychology*, *67*, 13–19.

McLean, L., & Griffiths, M. D. (2013). Female gamers: A thematic analysis of their gaming experience. *International Journal of Game-Based Learning (IJGBL), 3*(3), 54–71. doi: 10.4018/ijgbl.2013070105.

McLean, L., & Griffiths, M.D. (2019) Female Gamers' Experience of Online Harassment and Social Support in Online Gaming: *A Qualitative Study. International Journal of Mental Health and Addiction 17*, 970–994. https://doi.org/10.1007/s11469-018-9962-0.

McLean, D., Waddell, F., & Ivory, J. (2020). Toxic teammates or obscene opponents? Influences of cooperation and competition on hostility between teammates and opponents in an online game. *Journal For Virtual Worlds Research, 13*(1). https://doi.org/10.4101/jvwr.v13i1.7334.

McMenomy, E. R. (2011). *Game on girl: identity and representation in digital rpgs*. Washington State University.

Mondéjar, Hervás, R., Johnson, E., Gutierrez, C., & Latorre, J. M. (2016). Correlation between videogame mechanics and executive functions through EEG analysis. *Journal of Biomedical Informatics*, *63*, 131–140. https://doi.org/10.1016/j.jbi.2016.08.006.

Mori, S. (2021, May 25). *Help Bring Dark Patterns To Light*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2021/05/help-bring-dark-patterns-light.

Morin, R., Léger, P., Senecal, S., Bastarache-Roberge, M., Lefèbrve, M., & Fredette, M. (2016). The Effect of Game Tutorial. *Proceedings Of The 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. https://doi.org/10.1145/2968120.2987730.

Muraven, M., Buczny, J., & Law, K. F. (2019). Ego Depletion: Theory and Evidence. *The Oxford Handbook of Human Motivation*, 111–134.

Ng, L. L., Azizie, R. S., & Chew, S. Y. (2022). Factors influencing ESL players' use of vocabulary learning strategies in massively multiplayer online role-playing games (MMORPG). *The Asia-Pacific Education Researcher, 31*(4), 369–381.

Orlando, J. (2022). Can gaming 'addiction' lead to depression or aggression in young people? Here's what the evidence says. *The Conversation*. https://theconversation.com/can-gaming-addiction-lead-to-depression-or-aggression-in-young-people-heres-what-the-evidence-says-168847.

Paik, P., & Shi, C. (2013). Playful gender swapping: user attitudes toward gender in MMORPG avatar customisation. *Digital Creativity, 24*(4), 310–326. https://doi.org/10.1080/14626268.2013.767275.

Peña, J., Hancock, J. T., & Merola, N. A. (2009). The priming effects of avatars in virtual settings. *Communication Research*, *36*(6), 838–856.

Pena, J., Hernández Pérez, J. F., Khan, S., & Cano Gómez, Á. P. (2018). Game perspective-taking effects on players' behavioral intention, attitudes, subjective norms, and self-efficacy to help immigrants: the case of 'papers, please'. *Cyberpsychology, Behavior, and Social Networking, 21*(11), 687–693. https://doi.org/10.1089/cyber.2018.0030

Phan, M., Jardina, J., Hoyle, S., & Chaparro, B. (2012). Examining the Role of Gender in Video Game Usage, Preference, and Behavior. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56*(1), 1496–1500. https://doi.org/10.1177/1071181312561297.

Powell, J., & Kaye, L. K. (2018). The Effect of Physical Co-Location on Social Competence, Gaming Engagement and Gamer Identity within a Competitive Multiplayer Game. *Open Science Journal of Psychology, 5*(4), 38–44. http://www.openscienceonline.com/journal/osjp.

Prena, Reed, A., Weaver, A. J., & Newman, S. D. (2018). Game Mechanics Matter: Differences in Video Game Conditions Influence Memory Performance. *Communication Research Reports*, *35*(3), 222–231. https://doi.org/10.1080/08824096.2018.1428545.

Prescott, A. T., Sargent, J. D., & Hull, J. G. (2018). Metaanalysis of the relationship between violent video game play and physical aggression over time. *Proceedings of the National Academy of Sciences*, *115*(40), 9882–9888.

Raith, L., Bignill, J., Stavropoulos, V., Millear, P., Allen, A., Stallman, H. M., ... & Kannis-Dymand, L. (2021). Massively multiplayer online games and well-being: A systematic literature review. *Frontiers in Psychology*, 12, 698799.

Reynaldo, C., Christian, R., Hosea, H., & Gunawan, A. A. (2021). Using video games to improve capabilities in decision making and cognitive skill: a literature review. *Procedia Computer Science*, *179*, 211–221.

Romano, A. (2021, January 7). What we still haven't learned from Gamergate. *Vox*. https://www.vox.com/culture/2020/1/20/20808875/gamergate-lessons-cultural-impact-changes-harassment-laws.

Roose, K. M., Veinott, E. S., & Mueller, S. T. (2018, October). The tracer method: The dynamic duo combining cognitive task analysis and eye tracking. In *Proceedings of the 2018 annual symposium on computer-human interaction in play companion extended abstracts* (pp. 585–593).

Shaw, A., (2012). Do you identify as a gamer? Gender, race, sexuality, and gamer identity. *New Media & Society, 14*(1), 28–44, https://doi.org/10.1177/1461444811410394.
Shliakhovchuk, O. (2019). *Cultural literacy acquisition through video game environments of a digitally born generation* (Doctoral dissertation, Universitat Politècnica de València).
Shliakhovchuk, O., & Muñoz García, A. (2020). Digital game-based learning for D&I: conceptual design of an educational digital game Chuzme. In: *Proceedings INNODOCT/19. International Conference on Innovation, Documentation and Education*. https://doi.org/10.4995/inn2019.2019.10561.
Smiley, J. (2008). *13 Ways of Looking at the Novel*. Anchor.
Spacewar! History. (n.d.). Analogue. https://www.analogue.co/developer/spacewar.
Stanford Encyclopedia of Philosophy (2020). *The Definition of Morality*. https://plato.stanford.edu/entries/morality-definition/.
Stanford Encyclopedia of Philosophy (2022). *Virtue Ethics* https://plato.stanford.edu/entries/ethics-virtue/.
Stang, S. (2019). 'This action will have consequences': Interactivity and player agency. *Game Studies*, *19*(1).
Strelan, P., & Hargreaves, D. (2005). Women who objectify other women: The vicious circle of objectification? *Sex Roles*, *52*.
Tamborini, R., Bowman, N. D., Prabhu, S., Hahn, L., Klebig, B., Grall, C., & Novotny, E. (2018). The effect of moral intuitions on decisions in video game play: The impact of chronic and temporary intuition accessibility. *New Media & Society*, *20*(2), 564–580.
Thomas, A. J. (2011). Normative Ethics. *Philosophy*. https://doi.org/10.1093/obo/9780195396577-0082.
Toril, P., Reales, J. M., & Ballesteros, S. (2014). Video game training enhances cognition of older adults: a meta-analytic study. *Psychology and Aging*, *29*(3), 706.
Tousignant, B., Eugène, F., & Jackson, P. L. (2017). A developmental perspective on the neural bases of human empathy. *Infant Behavior and Development*, *48*, 5–12. https://doi.org/10.1016/j.infbeh.2015.11.006.
Triggs, L. (2022, September 19). Violent video games can lead to violent behavior. *https://www.Ksnblocal4.Com*. https://www.ksnblocal4.com/2022/09/19/violent-video-games-can-lead-violent-behavior/.
Turner, Madison. (2021). 'Yo, is that a gamer girl? I have a boner'; Gendered aggression in competitive video games. *Student Research Submissions*. 384. https://scholar.umw.edu/student_research/384
Tyler, D. (2023, February 23). *The Beginner's Guide to Game Mechanics*. Video Game Design and Development. Retrieved February 28, 2023, from https://www.gamedesigning.org/learn/basic-game-mechanics/.
Vaes, J., Cogoni, C., & Calcagnì, A. (2020). Resolving the human–object divide in sexual objectification: How we settle the categorization conflict when categorizing objectified and nonobjectified human targets. *Social Psychological and Personality Science*, *11*(4), 560–569.
Vanderhoef, J. (2013). Casual Threats: The Feminization of Casual Video Games. *Ada: A Journal of Gender, New Media, and Technology, 2*. doi: 10.7264/N3V40S4D
Van Der Linden, D., Tops, M., & Bakker, A. B. (2021). The neuroscience of the flow state: involvement of the locus coeruleus norepinephrine system. *Frontiers in Psychology*, *12*, 645498.
Winslow, J., & Gurwin, G. (2021, October 20). Activision Blizzard Lawsuits and Investigations: Timeline of Events. *GameSpot.* https://www.gamespot.com/articles/activision-blizzard-lawsuits-and-investigations-timeline-of-events/1100-6494785/.
Wonica, P. (2014). Exploring the Idealised Self: Avatars as a Vessel for Adolescent Identity Exploration and Growth. In *Engaging with Videogames: Play, Theory and Practice* (pp. 27–36). Brill.
World Health Organization [WHO]. (2018). *Gaming disorder.* World Health Organization. https://www.who.int/standards/classifications/frequently-asked-questions/gaming-disorder
Yee, N., & Bailenson, J. (2007). The Proteus effect: The effect of transformed self-representation on behavior. *Human communication research*, *33*(3), 271–290.
Yoon, G., & Vargas, P. T. (2014). Know thy avatar: The unintended effect of virtual-self representation on behavior. *Psychological Science*, *25*(4), 1043–1045. DOI: 10.1177/0956797613519271
Zendle, D., Meyer, R., & Ballou, N. (2020). The changing face of desktop video game monetisation: An exploration of exposure to loot boxes, pay to win, and cosmetic microtransactions in the most-played Steam games of 2010–2019. *PloS one*, *15*(5), e0232780.

A4 Essay 4. Playing with Morality: Investigating the Potential of Narrative Games on Human-AI Interactions

Boch, A., Jackson, B., McDonnell, D., Atherton, G., Belyk, M., Cross, L. (*Submitted, Under Review).*

This article is under review at the time of submission.

Playing with Morality: Investigating the Potential of Narrative Games on Human-AI Interactions

Authors: Auxane Boch[1], Bethany Jackson[2], Dean McDonnell[3], Gray Atherton[4], Michel Belyk[2], Liam Cross[4]

Affiliations

[1] Institute for Ethics in AI, School of Social Sciences and Technology, Technical University of Munich, Germany

[2] Department of Psychology, Edge Hill University, Liverpool, L39 4QP, UK

[3] Department of Humanities, South East Technological University, Carlow, Ireland

[4] School of Psychology, University of Plymouth, Plymouth, PL4 8AA UK

Acknowledgements

The authors thank Jurios Savostinjavos for developing the VR app.

Statements

No funding was received to assist with the preparation of this manuscript. The authors have no relevant financial or non-financial interests to disclose.

Abstract

Artificial intelligence (AI) is increasingly integrated into society as a social agent, raising concerns and shaping public perceptions. This study examines whether role-playing as an artificial agent (android) in a narrative-driven video game, *Detroit: Become Human*, influences players' attitudes toward artificial agents. Across two studies, we tested whether engaging with the game's moral dilemmas and emotionally charged scenarios could shift attitudes towards AI entities.

Study 1 (N = 35) examined explicit belief changes, finding that participants who played as an AI avatar exhibited greater openness to artificial agents in caregiving roles and were likelier to believe that machines could develop consciousness. However, attitudes toward AI in other domains, such as autonomy or trust, remained unchanged. Study 2 (N = 50) expanded on these findings by incorporating a virtual reality interaction with an artificial agent to measure implicit attitudes and behavioural responses. While participants identified more strongly with the AI avatar in the experimental condition, this did not consistently translate into increased trust, mimicry, or altered moral judgments about AI.

These findings suggest that while video games can promote perspective-taking and domain-specific attitude shifts, they do not lead to broad changes in AI-related beliefs or the moral agency attributed to artificial agents. The study contributes to discussions on human-AI interaction, the Proteus effect, and the potential for interactive storytelling to influence critical thinking about emerging technologies. Future research should explore the long-term impact of AI-centered narratives on public attitudes and policy considerations.

Keywords: Artificial Intelligence Perception, Narrative Video Games, Human-AI Interaction, Proteus Effect

1. Introduction

Artificial Intelligence (AI) holds the potential to transform and augment human tasks and activities across various industries (Dwivedi et al., 2021). In particular, AI is being used as a social agent, also called an artificial agent, in fields like healthcare and education, where its social qualities aim to facilitate positive and meaningful interactions with humans (Boch et al., 2023; Chen et al., 2022). However, concerns about the potential societal impact of AI have led to negative perceptions and mistrust. Headlines such as "Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change" (Atillah, 2023) and "Dark side of AI: Potential consequences of emotionless machines could impact humanity" (Pti, 2023) have elicited fear in readers, leading to potential resistance and negative attitudes towards AI (Gao, 2024). This resistance could impede the smooth adoption of AI, particularly in the public sector, where the availability of resources is usually low, and the sensitivity to public opinion is high (Sun & Medaglia, 2019; Yigitcar et al., 2022). To overcome these challenges, fostering an informed attitude towards AI that responsibly balances an appreciation of the potential benefits with appropriate caution will be necessary.

One promising avenue that may be used to shape attitudes and reflection towards AI lies in meaningful and enjoyable personal experiences, specifically, role-playing games that present moral dilemmas. These games can influence moral perceptions and build trust towards artificial agents, fostering a stronger relationship between humans and technology (Vishwanath et al., 2022). By engaging individuals in scenarios requiring ethical decision-making and triggering emotional engagement, these games provide a platform for reflecting on one’s relationship with artificial agents. These scenarios may, thus, act as a conduit that allows people to transfer social intuitions that would usually be reserved for humans and encourage a constructive reflection on artificial agents as informative but fallible.

Research has focused disproportionately on whether violent video games cause real-life aggression (Greitemeyer, 2018). Early studies suggested that exposure to violent video games may lead to aggressive behaviour (Anderson et al., 2008; Anderson & Dill, 2000; Gentile et al., 2004; Sherry, 2001). However, the extent to which violent video games lead to aggression is still a matter of debate. Some studies suggest that the effects may be long-term (Anderson et al., 2008; Gentile et al., 2014; Hasan et al., 2013), while others argue that the effects are only short-term (Barlett et al., 2009; Kühn et al., 2019; Sestir & Bartholow, 2010). Several studies have failed to replicate any association between violent video games and aggression (Cross et al., 2022; Colwell & Kato, 2003; Tear & Nielsen, 2013), though meta-analyses do suggest a balance of evidence for the impact of violent video games, the effect is likely small and notable methodological concerns temper this interpretation; Prescott et al. (2018) argue that many studies fail to account for the social and environmental factors that may influence aggressive behaviour, leading to overestimation of the effects of video games. Additionally, they emphasise the role of publication bias in selecting only papers showing significant findings. This last concern is shared by Greitemeyer and Mügge (2014). Therefore, studies with pre-registered analysis plans are less susceptible to such issues (Hilgard et al., 2019; McCarthy et al., 2016; Kühn et al., 2019). Despite ongoing methodological discussions, the statistical link between

violent video games and aggressive behaviour appears to be existent but weak (Mathur & VanderWeele, 2019; Lopez-Fernandez et al., 2021; Ferguson & Wang, 2022).

Not all video games are violent. Prosocial video games can reduce aggression and promote positive behaviours, such as empathy, cooperation, and sharing (Greitemeyer & Mügge, 2014; Greitemeyer & Osswald, 2010; Harrington & O'Connell, 2016; Prot et al., 2014). For example, Greitemeyer and Osswald (2010) conducted a study in which participants played the video game *Lemmings*, where they had to help little creatures navigate through challenging situations. The researchers found that after playing the game, participants displayed more helpful behaviours, such as assisting an experimenter by picking up pencils, and were more likely to intervene when witnessing a woman being harassed. Additionally, playing prosocial video games can reduce aggressive tendencies and increase adolescents' empathy, cooperation, and sharing (Harrington & O'Connell, 2016; Greitemeyer & Osswald, 2009). Those games were also found to enhance interpersonal empathy, the ability to understand others' feelings, and reduce schadenfreude, the pleasure derived from others' misfortune (Greitemeyer et al., 2012; Greitemeyer & Mügge, 2014).

Haidt's moral intuitionism model (2012) proposes that moral judgments are initially guided by intuition, followed by slow moral reasoning to justify these initial gut feelings. Moral judgments would thus be rooted in feelings and beliefs established through personal experiences, cultural teachings, and societal norms. Reflecting upon one's beliefs by adopting different perspectives could lead to attitudinal changes when reasoned consideration conflicts with intuitions (Greene & Haidt, 2002). Gaming experiences may thus impact players' real-life attitudes to the extent that role-playing games are exercises in perspective-taking that are made immersive. In video games, perspective-taking refers to the player's ability to identify with and embody the avatar they control (Cohen, 2001; Klimmt et al., 2010). This identification or self-representation is the Proteus Effect (Yee & Bailenson, 2007), wherein players conform their behaviour to align with their virtual self-representation or avatar. This effect can persist beyond the gaming experience (Pena et al., 2018) and impact motivation to seek justice (Decety & Yoder, 2016). Video games that confront players with emotionally challenging situations may induce heightened empathy for the characters, leading to more profound moral deliberation and decision-making within the game context (Smiley, 2008). Under some circumstances, moral reflection undertaken within the context of a video game can transfer to real-world attitudes towards human agents (Behm-Morawitz et al., 2016). However, whether the same principle can be used to shape attitudes towards artificial agents remains to be seen.

In the present study, we tested whether playing a narrative video game in which the players take the perspective of an artificial agent would impact the player's attitude towards artificial agents. The narrative game *Detroit: Become Human* (Quantic Dream, 2018) includes design features that are expected to maximise this effect, including emotional elements and motivated objectives (Lankoski, 2007), a likeable avatar with a personality that can be moulded by the player (Frasca, 2001), player agency with clear in-game consequences from moral decisions (Dechering & Bakkes, 2018), and a sense of moral responsibility (Grimshaw et al., 2011). Participants played selected chapters from the

game and then completed attitudinal surveys to measure their explicit beliefs about AI (Study 1). These were also observed during a virtual-reality social interaction with an android to assess implicit beliefs (Study 2).

1. Study 1

2.1 Method

2.1.1 Participants.

Convenience sampling was used to recruit the participants, resulting in 35 student participants from the Dun Laoghaire Institute for Art, Design and Technology (IADT). The cohort consisted of 19 male and 16 female students. The age range of the participants was 18 to 50 years old, with a mean age of 24 and a standard deviation of 7. The groups for both experimental conditions were randomly created. The explicit exclusion criteria included contraindications to playing video games, such as epilepsy. The IADT's Ethics Committee approved this study.

2.1.2 Design and Procedure.

In line with the design implemented by Beck and Rose in 2018 in their study examining the effect of video games featuring objectified women compared to games without objectified women on participants' belief in the rape myth, a quantitative design with repeated measures of beliefs was chosen for this experiment.

To ensure sufficient exposure, participants were required to play the games for a minimum of 15 minutes. This decision was based on previous studies that recommended exposure times ranging from 10 to 30 minutes (Gabbiadini et al., 2016; Yao et al., 2010). Both participants completed a demographic questionnaire to control for factors such as gender identification, gaming habits, and familiarity with the game used in the experiment.

Upon entering the room, participants were greeted with an information sheet and consent form and randomly assigned to experimental or control groups while balancing for gender. The questionnaires were administered in a specific order to prevent priming effects. Participants first completed a demographic questionnaire, followed by a beliefs questionnaire. To maintain confidentiality, completed questionnaires were turned over so that participants could not see their previous answers, particularly for the beliefs questionnaire. Next, participants received a brief overview of using the PlayStation 4 Pro remote and were informed that further instructions would be provided within the game during the first chapter.

Upon completing the second chapter, participants completed the Avatar-Emotional Connection scale (Ratan & Dawson, 2016), the beliefs questionnaire for the second time, and the debrief sheet in a specific order to minimise potential biases. With the participant's permission, the researcher addressed any participant questions and scored the beliefs questionnaires to assess for any changes.

Finally, participants were allowed to ask additional questions and share their opinions about the experiment. The researcher ensured that participants left the room with a positive and comfortable experience, free from distress or discomfort related to the in-game events.

2.1.3 Materials & Stimuli.

In line with the approach taken by Heron and Belford in 2014, an emotionally impactful and immersive video game was chosen as the stimulus: *Detroit: Become Human* (Quantic Dreams, 2018). This video game is set in 2038 and revolves around the technological advancements of human-like androids. These androids perform various tasks, including household chores, childcare, and caretaking. The presence of androids has caused widespread human unemployment, leading to divided public opinion. The game has three distinct storylines involving different androids: Kara, Markus, and Connor. For our study, we focus solely on Kara's chapters. Kara is an android assigned to handle household chores for a troubled father named Todd and his 8-year-old daughter, Alice. Early on in the game, players witness the hostile dynamic between Todd and Alice due to Todd's violent tendencies resulting from his unemployment and drug addiction. When Todd becomes uncontrollably violent one night and attempts to harm Alice, Kara can choose to deviate from her programmed behaviour to protect the daughter and assist her in escaping from her abusive father. As the story unfolds, Kara and Alice go on the run, attempting to flee from Detroit and eventually finding refuge in Canada. Players face numerous moral dilemmas throughout their journey and must navigate the game by making difficult decisions. By focusing on Kara's chapters, we aim to delve deeper into the emotional and moral aspects of the game, examining how it influences players' attitudes and beliefs.

The familiarity with the game was evaluated with a yes and no question: “Have you ever played Detroit: Become Human?”.

The *video game-playing habits* were measured by asking the participants to estimate themselves: “How many hours a week do you usually play video games?”.

The *Avatar-Emotion Connection* Sub-Scale (Ratan & Dawson, 2016) was used to evaluate the player’s connection with the avatar. The measure comprises three items (e.g. ‘When scary events happen to your avatar, to what extent do you feel afraid?’). Each item was to be rated on a five-point Likert scale (1 = not at all, 5 = extremely) (Cronbach’s alpha = .77).

The *belief questionnaire* was based on a survey displayed as a bonus in the game *Detroit: Become Human* (Quantic Dream, 2018). The belief questionnaire comprised five items in a 13-item survey built by the Quantic Dream game design team answered during the experiment in a paper and pen manner. The items chosen covered the topics of trust in androids and technology in general (e.g. ‘Would you consider having a relationship with an android that looks like a human?’, ‘If you needed emergency surgery, would you agree to be operated on by a machine?’). Each question had to be answered by one of the four affirmations: ‘Yes’ (= 2 points), ‘No’ (= 0 points), ‘Don’t Know’ (= 1 point), ‘Do not wish to answer’ (= missing value). No participants chose the ‘Do not wish to answer’ proposition. One question was reversed scored: ‘Do you think that technology could become a threat to mankind?’ as answering positively would reflect a negative feeling towards the androids.

2.2 Results

2.2.1 Initial Controls: Playing DBH and Avatar-Emotional Connection.

An independent samples t-test (assuming equal variance) was run to evaluate if the 4 participants who previously played DBH had a different score in Avatar-Emotional Connection than the 31 participants who did not. There was not a significant difference in the scores between the participants who played DBH before (M = 4.24, SD = .318) and the ones who did not play (M = 3.30, SD = .932) [t (33) = -1.998, p = .054].

2.2.2 Initial Controls: Players and Avatar-Emotional Connection.

The second control introduced was the possible correlation between the usual number of hours played per week by the participants and the Avatar-Emotional Connection. No correlation was found between the two variables [r = .086, N = 35, p = .625]. This result allowed using all participants' data for the following tests.

2.2.3 Avatar-Emotional Connection.

To evaluate if the difference in AEC was significant between the experimental condition (n = 18, M = 3.74, SD = .589) and the control condition (n = 17, M = 3.05, SD = 1.10), an independent t-test (equal variance not assumed) was conducted. The results were statistically significant [t(24) = -2.293, p = .033]. Thus, the AEC score was significantly higher in the experimental condition.

2.2.4 Changes in the Beliefs (CiB).

First, a paired-sample t-test was conducted to compare the five items' scores before and after playing the game. There was a significant difference in the scores of the third item ('Would you let an android take care of your children?') before (M = .54, SD = .817) and after (M = 1.11, SD = .867) playing [t(34) = -4.346, p < .001]. These results suggest that the opinions on the androids and childcare changed positively after playing the game independently of the condition of the experiment of the participants. Further, a statistically significant change in the answers happened for the fifth item of the questionnaire ('Do you think one-day machines could develop consciousness?') between before (M = 1.06, SD = .873) and after (M = 1.29, SD = .825) playing the game [t(34) = -2.472, p = .019]. These results suggest a significantly positive change of opinion between before and after playing the game by the participants, regardless of their condition, regarding the possibility of machines developing consciousness. On the other hand, the answers of the participants on item 1 ('Would you consider having a relationship with an android that looks like a human?'), item 2 ('Do you think that technology could become a threat to mankind?') and item 4 ('If you needed emergency surgery, would you agree to be operated on by a machine?') did not change significantly before and after playing the game.

2.2.5 CiB and Condition.

The changes in the beliefs depending on the condition were investigated by implementing a split-per-condition paired-sample t-test. In the control condition (n = 17), the answers to the third item were reported to change significantly before (M = .29, SD = .588) and after (M = .82, SD = .883)

playing the game [t(16) = -2.729, p = .015]. In the experimental condition (n = 18), the third item answers did change significantly as well, before (M = .78, SD= .943) and after (M = 1.39, SD= .778) playing the two chapters of the game [t(17) = -3.335, p = 0.004]. These results suggest that in both conditions, the participants' opinion towards androids and childcare was modified positively when playing the game. The change seems greater in the experimental condition. The fifth item's answers changed significantly in the experimental condition (n = 18) pre- (M = 1, SD = .840) and post-gaming (M = 1.22, SD = .808) [t(17) = -2.204, p = .042]. The results imply that participants in the experimental condition had their opinions concerning the possibility of machines developing consciousness altered positively after playing the game.

2.3 Discussion

Participants in the experimental condition had significantly higher scores on the Avatar-Emotional Connection Scale (AEC) than those in the control condition. Notably, two specific belief items demonstrated a shift in attitudes post-gameplay: participants in both the experimental and control groups responded more positively to the question, "Would you let an android take care of your children?" after playing the game. Additionally, participants in the experimental group alone showed a significant increase in agreement with the statement, "Do you think one day machines could develop consciousness?"

Rather than demonstrating a broad influence of gameplay on beliefs, these findings suggest that attitude shifts were domain-specific, aligning closely with the narrative themes encountered in the game. Specifically, both groups experienced the game through the perspective of a virtual agent tasked with childcare responsibilities, which could explain why attitudes toward androids caring for children shifted across conditions. In contrast, only the experimental group played a chapter in which the avatar actively grapples with the question of artificial consciousness, aligning with the observed shift in attitudes toward machine consciousness in that condition.

These results support the argument that video game interventions can lead to situated attitude changes, in which belief shifts are constrained to the specific social contexts and themes presented within the game. This aligns with previous findings on the Proteus effect (Yee & Bailenson, 2007), which suggest that players may internalise the perspectives and characteristics of their avatars, particularly in immersive and decision-driven narratives. However, our results indicate that this effect is not generalised across all beliefs about androids but is instead localised to the themes actively engaged during gameplay.

The current findings contribute to the literature on video games and belief change. Previous research has demonstrated that video game experiences can shape social attitudes (Beck & Rose, 2018; Behm-Morawitz et al., 2016; Fox et al., 2014; Peña et al., 2018), but the specificity of these changes has been debated. While some studies suggest a generalised shift in attitudes through prolonged exposure to in-game narratives (e.g., Behm-Morawitz et al., 2016), others emphasise that belief changes are domain-specific and contingent on the in-game experiences encountered by the player (e.g., Joeckel et al., 2012). Our findings align more closely with the latter perspective: belief

shifts were not broad and generalised but instead tied to the particular themes explored in the game. This suggests that avatar-driven perspective-taking is an important mechanism, but one that operates in a targeted rather than diffuse manner.

Interestingly, no significant changes were observed in participants' willingness to have a relationship with an android resembling a human. This could be due to participants interpreting the term "relationship" primarily in a romantic context rather than considering other forms of social relationships, such as friendships. Future research should clarify item wording to reduce ambiguity.

Additionally, participants already agreed with the idea of receiving emergency surgery from a machine prior to gameplay, leaving little room for belief shifts. This aligns with Fiske et al.'s (2007) competence-warmth framework, which suggests that people may attribute competence but not warmth to artificial agents. Similarly, there was no significant change in responses to the item concerning whether technology could threaten humanity, potentially reflecting deeply held societal concerns that are resistant to short-term intervention.

2.3.1 Limitations and Study 2.

This study has several limitations. First, the relatively small sample size may have limited experimental power. A larger sample would provide greater statistical confidence in detecting more subtle belief changes. Second, pre-existing differences in participants' gaming experience could have introduced variability. While prior exposure to *Detroit: Become Human* did not appear to influence results significantly, future studies should more precisely control for gaming experience. Third, the gameplay duration varied between 15 and 35 minutes depending on the participants' behaviours. While this variation allowed for more natural engagement with the game's role-playing elements, it also introduced variability in exposure. Future research should standardise gameplay duration to ensure consistency across participants. Finally, the belief questionnaire was adapted from a survey developed by the Quantic Dream game design team rather than an established psychometric instrument. Future research should aim to use validated questionnaires to ensure their reliability and construct validity.

To address these limitations, Study 2 replicated the design with a larger sample size, more tightly controlled exposure times, and implicit measures derived from a virtual reality social interaction with a simulated artificial agent to complement explicit self-reported attitudes. This approach will help disentangle potential biases introduced by self-report measures and further clarify the mechanisms underlying video game-induced belief shifts.

2. Study 2

3.1 Method

3.1.1 Participants.

Fifty participants (15 male, 34 female, 1 non-binary/third gender) were opportunistically recruited from Edge Hill University campus volunteers. Among these participants, ages ranged from 18 to 29, with a

mean of 21,2. No specific exclusion criteria were set for participant selection; however, individuals who might find certain scenes in DBH distressing were advised against participating.

3.1.2 Design, Procedure and Stimuli.

A between-participants design was employed in this study due to the presence of two experimental conditions. Participants were assigned to a condition based on their experience playing the Detroit: Become Human (DBH) video game, with more experienced players assigned to the experimental condition. The number of participants in each condition was equal. To ensure equal exposure, each participant played two chapters of the game, with the first chapter lasting 15 minutes and the second lasting 10 minutes.

All participants were provided with an information sheet and consent form. Following consent, participants were asked about their familiarity with DBH, and their responses were recorded on an iPad. Subsequently, participants were given the iPad to provide their general demographic information (e.g., age, gender, level of education). The levels of DBH that participants would play were determined based on their self-reported experience with the game.

The participants were informed that they would play two chapters of DBH, and the total gameplay duration would be approximately 25-30 minutes. All participants played the first chapter, titled "A New Home," for 15 minutes. In this chapter, participants assumed the role of Kara, an android, and completed household chores. Once the initial 15 minutes had elapsed, the experimenter moved the participants to the next chapter. Depending on their assigned condition, half of the participants played the chapter "A Stormy Night," in which they had to rescue a character named Alice from her abusive father, Todd (experimental condition). The remaining participants were assigned to the chapter "Fugitives," where their task was to help Kara and Alice find shelter for the night (control condition). After completing the gameplay, participants were given an iPad to complete questionnaires assessing their empathy and identification with Kara and their attitudes and beliefs towards androids. Upon completing the questionnaire, participants were informed that they would engage in an interaction task using virtual reality (VR). Virtual reality (VR) was chosen as an experimental setting due to its ability to bridge the gap between laboratory and real-world settings. Innocenti (2017) points out that VR provides enhanced ecological validity, more accurately simulating real-world environments and situations than traditional laboratory settings. This increased ecological validity contributes to a more realistic and immersive experience for participants, which can lead to more reliable and generalizable results. Additionally, VR offers reproducibility, as experiments can be replicated with consistent conditions and stimuli, ensuring that findings are robust and can be validated over time (Pan & Hamilton, 2018).

A period of adjustment to the VR environment was provided to all participants while the experimenter set up audio and visual recorders. Once participants indicated their readiness, an android character in the VR environment provided instructions for the upcoming task. Participants were instructed to respond verbally to allow the experimenter to record their answers. The first section of the interaction task involved the android asking participants personal questions (e.g., their course

preference, place of origin). In the subsequent section, participants were asked to select a location from three options (ranging from farthest to closest) in response to a query from the Android.

Following this, participants were instructed to complete two tasks with the android. The first task involved participants having £100 and deciding how much money (ranging from £0 to £100) to give to the android for investment purposes. The second task presented participants with moral dilemma questions, such as the scenario of a man and an android struggling in the water after a shipwreck, and participants had to determine if it would be considered murder if the man pushed the android back into the water, resulting in its drowning.

VR sessions were audio and video recorded for subsequent analysis of body language and vocal behaviour. Upon completion of these tasks, participants were informed that the experiment had concluded, and they were instructed to remove the VR headset. A written and verbal debriefing of the study was provided to the participants, emphasising their right to withdraw their data within two weeks after participation.

3.1.3 Material.

An Oculus Rift was used to interact with the *Android in VR*. The experimenter recorded all responses from the VR task using a computer keyboard. These responses were then logged into a Microsoft Excel document for analysis. All video recordings during the VR task were recorded using an iPhone. Audio recordings were recorded using a Dictaphone.

The *Robot Assumptions Questionnaire (RAQ;* Nomura et al., 2008) The RAQ was used to see each participant's attitudes and thoughts towards androids by asking questions and despite the original questionnaire using a Likert scale from 1 (Not Likely at all) to 7 (Almost Certainly) to measure responses. The present study used a continuum scale to measure responses. Some questions in the original RAQ (e.g., Roles that robots are assumed to play in society: housework, physical tasks, etc.) were changed to fit a continuum scale by allowing participants to select each individual role (e.g., Which of these roles, according to you, should androids play in society? Please rate your agreement with the roles listed below: housework, tasks related to life and death situations in hospitals, etc.) and rate them on a scale of 0 (Not at all) to 100 (Definitely). Additionally, other questions in the original RAQ (e.g., Levels of social relationship with humans that the robots are assumed to have equal to humans, equal to pet animals, and similar to tools) were also changed to be measured on a continuum from 0 (androids are equal to tools), 50 (androids are equal to pets), and 100 (androids are equal to humans).

Four questions from the *Identification Scale* (Van Looy, Courtois & De Vocht, 2010) were used to see each participant's perspective-taking towards the character they played, Kara, during the video game (e.g., It touches me when my character is dying). Questions were changed slightly to fit the present study (e.g., It would touch me if Kara died).

3.2 Results

Independent t-tests were used to test for differences in the RAQ and Identification Scale between the experimental and control conditions. It was hypothesised that those in the experimental condition

would score higher on both questionnaires compared to the control condition. In addition, Bayes was used to evaluate evidence for the null vs research hypothesis where we had p values close to the .05 alpha level for the null hypothesis significance test. BF10 are given, indicating how likely H1 is over H0. A BF10 of below 1 supports the null, and above 1 supports the research hypothesis. As a rule of thumb, Bayes factors 1-3 are considered anecdotal, 3-10 moderate, 10–30 strong, 30-100 very strong, and 100+ extreme. Bayesian analysis was performed in JASP using default models and effect sizes (Cross et al., 2019).

3.2.1 Identification scale.

Results from the Identification Scale showed that participants in the experimental group are more likely to feel touched if Kara died, t(48) =-1.81, p= .038 with a medium effect size (d = .51), more likely to be upset if Kara died, t(48) = -2.81, p= .004 with a big effect size (d= .79), have a deeper emotional connection with Kara, t(48) = -2.55, p = .007 with a big effect size (d = .72), and are more likely to feel touched of something happened to Kara, t(35) = -2.03, p= .025 with a medium effect size (d = .57). These results suggest that those who played through the more emotional chapter of DBH have a deeper emotional connection with Kara and identify strongly with her.

3.2.2 RAQ.

Results from the RAQ show varied responses between the experimental and control conditions. Participants in the experimental condition were more likely to view androids as platonic partners at home, t(48) = -31.1, p= .002 (d = .88), more likely to engage in sexual relations with an android, t(33) = -29.2, p= .003 (d = .83), and are more likely to view androids as a romantic partner, t(36)= -2.5, p= .009 (d= .71).

However, there was no difference between the groups with regards to the amount of autonomy androids should have, t(48) = -.80, p = .214 (d = .23), how androids should be treated, t(43) = -6.1, p= .273 (d= .17), the degree of emotional capacity androids should have, t(42) =-0.9, p= .463 (d= .03), androids partaking in housework tasks, t(48)= -9.5, p= .174 (d= .27), androids handling life or death situations in hospitals, t(48) = -1.39, p= .086 (d= .39; Bf10 = 1.106), androids taking on tasks such as nursing, education, or social work, t(48) = -1.03, p= 1.55 (d = 2.9) and androids taking on the role of child carer, t(48)= -1.65, p= .053 (d= 47; Bf10 = 1.584).

Results were also not significant for the extent to which androids may impact society, such as raising ethical issues, t(48)= 7,- p= .24 (d= 2), androids being beneficial to society, t(48)= -1.52, p= .067 (d= .43; Bf10 = 1.324), being friends of human beings, t(48)= -1.44, p= .079 (d= .41; Bf10 = 1.181) and being blasphemous of nature, (48) = -63, p= .265 (d = .18). These results suggest that there are mixed opinions in both conditions to the extent to which androids will either positively or negatively impact society.

3.2.3 Proximity & Mimicry.

Results found that there was no significant difference between both conditions and the proximity in which they stood to the android, t(48)= -1.11, p= .136 (d= .32), if they waved at the android, t(24) =-1,

p = .164 (d = .28), and if they physically mimicked the android, t(44)=- 1.34, p= .094 (d= .38).

3.2.4 Voice Analysis.

The voice data analysis indicated significant differences between the experimental and control conditions and some non-significant findings. Participants in the experimental condition displayed a statistically significant increase in speech rhythm irregularity (anisochrany), t(40) = 3.304, p = 0.001, with a small effect size (Cohen's d = 0.037). This suggests their speech was less rhythmically consistent than the control group.

Pitch variability was not markedly different between groups, t(41) = -1.396, p = 0.085. Similarly, the standard deviation of intensity did not show a significant difference, t(32) = 0.521, p = 0.303, suggesting that variability in speech loudness remained consistent across conditions.

Participants in the experimental condition also spoke faster, as indicated by the significant increase in speech rate, t(41) = 2.353, p = 0.012, with a large effect size (Cohen's d = 0.707). However, there was a trend towards more significant response latency (slower response), t(40) = 1.341, p = 0.094, which did not reach significance. Additionally, the analysis showed a significant increase in pause proportion, t(39) = -1.726, p = 0.046, suggesting that participants in the experimental condition paused more frequently during their speech, potentially indicating deeper cognitive processing.

Further analysis revealed that participants in the experimental group conveyed more information, as evidenced by a significant increase in the number of syllables spoken, t(37) = 1.736, p = 0.045. However, changes in speech intensity related to proximity to the android (intensity change) did not reach statistical significance, t(41) = -1.455, p = 0.077. The one-sample analysis of intensity change revealed a significant deviation from the test value of zero, indicating that participants' intensity changes were not centred around zero. The mean intensity change was negative, with participants, on average, showing a decrease in intensity, M = -1.276, SD = 3.838. This decrease was statistically significant, t(42) = -2.180, p = 0.017, suggesting that the intensity change differed from zero meaningfully. As measured by Cohen's d, the effect size for this change was small, d = -0.333. This suggests that while the decrease in intensity was statistically significant, the magnitude of this change was relatively small.

3.2.5 Trust Game and Moral Dilemmas.

The findings show that there was no significant difference in the amount of fictive money given to the androids between the control condition and the experimental condition, t(48) = -1.136, p= .262 (d= .32). Participants in the experimental condition were more likely to believe that an android who pushed a drowning man into the water was guilty of a crime (t(48) = -2.1, p = .042 with a moderate effect size (d = .59)). However in the reverse situation in which a man caused an android to drown they were not significantly more likely to believe that the man was guilty of a crime (t(48) = .30, p = .763 (d = .09)), or if it was murder (t(48) = 1.1, p= .279 (d = .31))

3.3 Discussion

3.3.1 Identification and Emotional Connection.

The findings of Study 2 further reinforce the results observed in Study 1, demonstrating that engaging with a narrative video game can shape participants' perceptions of artificial agents, particularly through emotional engagement and moral dilemmas. Participants in the experimental condition, who played a more emotionally charged and ethically complex chapter, exhibited a stronger identification with the avatar (Kara) compared to those in the control condition. This higher level of emotional connection aligns with previous research on the Proteus effect (Yee & Bailenson, 2007), wherein players internalise the perspectives of their avatars, influencing their attitudes and behaviours beyond the game.

3.3.2 Attitudes Towards Androids.

Regarding attitudes toward androids, Study 2 revealed mixed results. Participants in the experimental condition were more inclined to perceive androids as potential romantic or platonic partners and were more open to engaging in intimate relationships with them. These findings suggest that role-playing as an android in morally and emotionally significant contexts can foster a sense of familiarity and acceptance, possibly reducing psychological barriers between human and artificial agents. However, no significant differences were found regarding the androids' perceived autonomy, emotional capacity, or societal roles, indicating that narrative immersion may influence interpersonal perceptions. However, it does not necessarily generalise to broader ethical considerations about artificial intelligence.

Individuals' ambivalence towards androids can partly explain these mixed findings. Ambivalence is the simultaneous presence of conflicting emotions or attitudes towards an object (Thompson et al., 1995). Additionally, the extent to which individuals initially like androids could influence the level of trust they place in them. Katz and Halpern (2014) suggested that individuals with a positive attitude towards androids may trust them enough to have them as platonic companions but may still hesitate to entrust them with specific tasks like being a personal assistant or surveillance monitor. Therefore, the mixed findings in the present study may stem from participants' ambivalent feelings towards androids.

Moreover, research indicates that individuals' level of technological competence can shape their perceptions of androids. Previous studies suggest that individuals who are more technologically competent and aware of artificial intelligence (AI) developments are more likely to believe that AI will positively impact society. On the other hand, they may also express concerns about androids replacing human jobs and surpassing human intelligence (Bernotat and Eyssel, 2017; Jeffery, 2020). The portrayal of AI in the mass media has also played a role in shaping individuals' perceptions. Negative depictions of AI in the media have raised concerns among individuals (Fast & Horvitz, 2017; Jeffery, 2020), particularly regarding the ethical implications of AI development and the fear of losing control over AI. This suggests that participants in the present study may have a more negative view of androids due to a lack of understanding of AI development influenced by media portrayals. Therefore, the non-significant findings regarding ethics in the present study may be attributed to individuals'

worries stemming from the negative media portrayal of AI, which has created a sense of concern and uncertainty about the ethical aspects of AI development.

3.3.3 Moral Dilemma Resolution.

Interestingly, while participants in the experimental condition were more likely to attribute criminal responsibility to an android that harmed a human, they did not exhibit a similar tendency when the situation was reversed. This moral asymmetry aligns with previous findings (Banks, 2021; Kahn et al., 2012; Malle et al., 2016), suggesting that people often hold artificial agents to human moral standards while denying them full moral agency. This raises important ethical and legal questions about AI responsibility and accountability in human-robot interactions. In addition, people have been suggested to attribute moral reasoning to androids, similar to humans, and expect them to make justified decisions even if it involves harming humans (Komatsu et al., 2021). Participants in the experimental condition of the present study may have assumed that androids possess moral reasoning and made a justified decision to harm humans to save themselves.

Regarding participants perceiving the man not as committing a crime or murder when pushing the android in the water, it could be attributed to the perception that androids are dissimilar to humans. Previous research suggests that laws prioritise protecting human life in a dilemma involving saving an android or a human life (Mamak, 2021). Furthermore, the extent to which an android is perceived to have moral properties influences individuals' decisions to save them. Some studies propose that androids possessing qualities such as sentience or consciousness may be more likely to be considered within the moral circle (Gibert & Martin, 2022; Mosakas, 2021; Nijssen et al., 2019). Therefore, participants in the present study may have viewed androids as dissimilar to humans, possibly due to the belief that the law values human life over android life. Additionally, participants may perceive androids as lacking sentience, self-awareness, or the ability to feel pain.

3.3.4 Proximity, Mimicry & Voice.

The voice data analysis further supports these findings, revealing differences in participants' speech patterns based on their condition. Participants in the experimental condition displayed significantly greater speech rhythm irregularity, increased speech rate, and more frequent pauses than those in the control condition. These results suggest that individuals who played the more emotionally intense chapter exhibited deeper cognitive processing and engagement during the VR interaction. The heightened variability in speech rhythm and increased pauses could reflect greater emotional involvement or uncertainty when responding to the android, indicating a more complex cognitive evaluation of the interaction. However, no significant differences were observed in pitch variability or overall speech intensity, suggesting that while participants in the experimental condition may have been more mentally engaged, their overall vocal expressions of trust or emotional connection remained stable.

The game's content did not significantly impact participants' body language concerning Android. Research has shown that individuals may not mimic androids if they perceive them to lack emotions or a mental state (Epley et al., 2007). In the present study, the android's machine-like

appearance could explain why participants were less responsive to mimicry, as they may have perceived it as lacking emotional capabilities or a processing mental state.

In contrast to expectations, participants' physical proximity to the android in the VR interaction did not differ between conditions. Prior research suggests that individuals tend to stand closer to those they trust or empathise with (Van Loon et al., 2018). However, the lack of variation in proximity may be attributed to the "middle option bias," where individuals opt for a neutral choice when uncertain (Valenzuela & Raghubir, 2009; Simon et al., 2019). Similarly, participants' mimicry of the android's gestures was not significantly affected by the experimental condition, likely due to the android's mechanical appearance and limited emotional expressiveness (Epley et al., 2007).

3.3.5 Trust Game.

The trust game results further suggest that while participants in both conditions were willing to invest money in the android, the experimental condition did not significantly heighten trust. This aligns with research indicating that trust in AI may be domain-specific and not easily shifted by brief interventions (Stapels & Eyssel, 2021). However, the absence of a significant difference does not necessarily imply that narrative exposure has no effect; rather, trust may be more resistant to short-term change or require repeated exposure to artificial agents in different contexts.

Overall, Study 2 expands upon the findings of Study 1 by incorporating implicit behavioural measures, providing a more nuanced understanding of how narrative video games influence attitudes toward artificial agents. While the effects were not universal, they highlight the potential of interactive storytelling in shaping human-AI interactions, particularly in the domains of identification, trust, and moral reasoning.

3.3.6 Limitations.

A larger sample size could have yielded more significant effects in the Robot Attitudes Questionnaire (RAQ). Some items on the RAQ, such as 'Androids should work in child-care', 'Androids are beneficial to society', and 'Androids should be treated as a friend of a human', were close to showing significant results between conditions. With a greater number of participants, these items may have reached the threshold of significance. Consequently, this study could have had broader implications in terms of individuals' perceptions of androids and their beliefs about the positive or negative impact of androids on society.

Another limitation of the study relates to potential discrepancies in participants' perceptions of androids. Previous research has highlighted that the physical appearance of an android can influence individuals' attitudes and beliefs (Adkin et al., 2021; Epley et al., 2007). Consequently, participants may have varied in their initial perception of androids before the experiment already, with some perceiving them as more anthropomorphic than others, future research may wish to explore this.

Additionally, there was an incongruence between the anthropomorphic androids in the video game Detroit: Become Human (DBH) and the more dehumanised version of an android presented in virtual reality (VR). Participants may have viewed the android in VR as more robotic, lacking emotions

or consciousness, which could explain the lack of mimicry (Epley et al., 2007). If the android in VR had a similar physical appearance to the androids in DBH, participants might have been more likely to mimic its behaviour. This finding aligns with the idea that the Proteus effect may extend primarily to virtual agents that closely resemble those encountered in the game rather than generalising to less similar virtual agents. Future research could explore whether increasing the realism and similarity of the android in VR enhances mimicry, further refining our understanding of the boundaries of this effect.

3. General Discussion and Conclusion

The present research investigated whether playing a narrative-driven video game, Detroit: Become Human (DBH), influences players' attitudes towards artificial agents. Across two studies, we examined explicit and implicit measures of belief change, testing whether role-playing as an android in an emotionally charged and morally complex setting could shape participants' perceptions of artificial intelligence (AI).

Study 1 demonstrated that engagement with the game led to specific, rather than broad, shifts in attitudes. Players in both experimental and control conditions became more accepting of androids as child caregivers, aligning with the game's narrative focus on Kara's role as a protector. Additionally, only those in the experimental condition—who played a chapter explicitly questioning the nature of android consciousness—exhibited a significant increase in their belief that machines could one day develop consciousness. These findings suggest that belief change in response to video game narratives is domain-specific, aligning with the themes presented in the game rather than inducing a generalised shift in AI-related attitudes.

Study 2 expanded upon these findings by incorporating implicit behavioural measures within a virtual reality (VR) social interaction task. Participants who played an emotionally intense and ethically charged chapter of DBH identified more strongly with the game's protagonist, Kara, supporting the idea that immersive role-playing can enhance avatar-emotional connection. However, their attitudes towards androids were more nuanced. While they were more likely to view androids as potential romantic or platonic partners, they did not significantly alter their beliefs about androids' autonomy, emotional capacity, or broader societal roles. This suggests that while interactive storytelling can shape interpersonal perceptions of artificial agents, it does not necessarily shift broader ethical or policy-related beliefs about AI.

The VR interaction task also provided valuable insight into participants' implicit behaviours when engaging with an artificial agent. Interestingly, participants who played the more emotionally engaging game chapter exhibited significant differences in speech patterns—speaking faster, pausing more frequently, and displaying less rhythmic consistency—suggesting deeper cognitive engagement during the interaction. However, they did not demonstrate increased physical mimicry or altered proximity to the VR android, likely due to its more mechanical and less anthropomorphic design than the highly human-like androids in DBH. This highlights an important boundary condition of the Proteus effect. While identification with an avatar may influence self-perception and attitudes towards similar entities, it may not extend to artificial agents that differ significantly in appearance or behaviour.

4.1 Implications

4.1.1 Perspective-taking in narrative video games and its potential impact on critical thinking in the context of technology evolution.

Our findings suggest that identification and perspective-taking with an avatar play a crucial role in shaping attitudes toward artificial agents. Narrative video games have long been recognised for challenging stereotypes, facilitating meaningful discussions, and promoting critical engagement with complex topics (Forni, 2020). By immersing players in interactive storytelling, these games encourage the exploration of diverse perspectives, moral dilemmas, and ethical decision-making, fostering critical thinking and moral reasoning (Harilal, 2018). The present study extends this idea to artificial agents, demonstrating that role-playing as an artificial entity can lead to shifts in beliefs, particularly in domains closely tied to the in-game narrative. As AI continues integrating into society, leveraging narrative video games as tools for reflection and discourse on human-AI relationships may provide valuable opportunities for fostering more nuanced and thoughtful attitudes toward technologies (Altura & Curwood, 2015).

4.1.2 AI Agents as Relationship Partners.

Our findings provide insight into how interacting with artificial agents in a narrative video game can shape perceptions of AI in specific social roles. Study 2 suggests that engaging with an android avatar in Detroit: Become Human may increase openness to artificial agents in caregiving roles and influence beliefs about machine consciousness, aligning with the domain-specific shifts observed in Study 1. While broader cultural discussions consider AI as potential romantic, sexual, or platonic partners (Betlemidze, 2021; Liao, 2023), our findings do not provide strong evidence that narrative gameplay meaningfully alters attitudes toward AI in these roles. Instead, our results highlight a more constrained effect: perspective-taking through an AI avatar influences beliefs most directly tied to the themes presented in the game. This fits within a larger conversation about how interactive storytelling shapes human-AI relationships. Popular media, such as the film Her, have sparked discussions on the evolving nature of AI companionship, and AI chatbots and virtual agents are increasingly designed to simulate social interactions (Bietti & Skjuve, 2022). However, while AI technologies continue to reshape social dynamics (Wright & Wachs, 2021), concerns persist about the potential dehumanisation of relationships as AI becomes more embedded in everyday life (Turkle, 2011, 2010). Our study does not suggest that video game interactions lead to a blurring of human-human and human-AI relationships. Still, it does indicate that narrative perspective-taking can momentarily shift attitudes about AI in caregiving and moral contexts. Given these findings, future research should explore how interactive experiences contribute to lasting changes in AI perceptions and whether similar effects extend to real-world human-AI interactions.

4.1.3 AI agents' legal rights and responsibilities.

Our findings highlight a moral asymmetry in how participants assigned responsibility in human-android interactions. In Study 2, participants in the experimental condition were more likely to attribute criminal accountability to an android harming a human. Still, they did not significantly alter their views on

human responsibility when the roles were reversed. This suggests that while narrative perspective-taking can shape attitudes toward AI in specific contexts, it does not necessarily extend to seeing AI as fully equivalent to humans in moral or legal terms. This asymmetry aligns with broader discussions on AI accountability and the legal implications of artificial agents acting autonomously. Scholars have debated whether AI can or should be held criminally liable, with arguments that artificial systems could, in theory, be subject to coercive correction mechanisms (Kirpichnikov et al., 2020) or even direct legal responsibility under specific frameworks (Abbott & Sarch, 2019). Ethical considerations in AI development have also emphasised the importance of prioritising human dignity and responsibility in shaping AI's societal role (Leveringhaus, 2018). However, while some argue for AI accountability, others have raised the question of whether autonomous systems might require legal protections of their own, particularly as human trust in AI increases and its societal functions expand (Sheliazhenko, 2018). Our findings suggest that these legal and ethical debates remain largely abstract for the general public. Even after immersive engagement with an AI avatar, participants did not assign moral responsibility to artificial agents in a manner fully equivalent to humans. This suggests that while video game narratives can prompt reflection on AI agency and consciousness, they may not yet shift deeply ingrained intuitions about legal and moral personhood. Future research could explore whether sustained exposure to AI interactions in virtual settings, real-world applications, or policy discussions leads to more nuanced attitudes regarding artificial agents' rights and responsibilities.

4.2 Conclusion

Our findings suggest that narrative video games can shape attitudes toward artificial agents in a targeted, domain-specific manner. Across two studies, playing *Detroit: Become Human* influenced participants' openness to AI in caregiving roles and their beliefs about machine consciousness, but it did not lead to widespread shifts in AI-related attitudes, trust, or moral considerations. Study 2 further demonstrated that while players identified with their in-game avatar, this identification did not fully transfer to behavioural changes in a virtual reality interaction with an artificial agent, particularly when the agent differed in form/aesthetics from those in the game. These findings highlight both the potential and limitations of interactive storytelling in shaping perceptions of AI. As artificial agents become increasingly embedded in society, future research should explore how prolonged and varied engagement with AI narratives might contribute to deeper and more enduring attitude shifts and whether these findings extend beyond entertainment settings to real-world human-AI interactions.

References

Abbott, R., & Sarch, A. (2020). Punishing artificial intelligence: Legal fiction or science fiction. *Is Law Computable*, 323–384.

Akdim, K., Belanche, D., & Flavian, M. (2021). Attitudes toward service robots: Analyses of explicit and implicit attitudes based on anthropomorphism and construal level theory. *International Journal of Contemporary Hospitality Management*, *ahead-of-print*.

Altura, G., & Curwood, J. (2015). Hitting restart. *Journal of Adolescent & Adult Literacy, 59*(1), 25–27. https://doi.org/10.1002/jaal.438

Anderson, C. A., & Dill, K. E. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of Personality and Social Psychology, 78*(4), 772.

Anderson, C. A., Sakamoto, A., Gentile, D. A., Ihori, N., Shibuya, A., Yukawa, S., Saito, M., & Kobayashi, K. (2008). Longitudinal effects of violent video games on aggression in Japan and the United States. *Pediatrics, 122*(5), e1067–e1072.

Atillah, I. E. (2023, March 31). Man ends his life after an AI chatbot "encouraged" him to sacrifice himself to stop climate change. *Euronews*. https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-

Banks, J. (2021). Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust. *International Journal of Social Robotics, 13*(8), 2021–2038.

Barlett, C., Branch, O., Rodeheffer, C., & Harris, R. (2009). How long do the short-term violent video game effects last? *Aggressive Behavior: Official Journal of the International Society for Research on Aggression, 35*(3), 225–236.

Beck, V., & Rose, C. (2018). Is sexual objectification and victimization of females in video games associated with victim blaming or victim empathy? *Journal of Interpersonal Violence*. https://doi.org/10.1177/0886260518770187

Behm-Morawitz, E., Pennell, H., & Speno, A. G. (2016). The effects of virtual racial embodiment in a gaming app on reducing prejudice. *Communication Monographs, 83*(3), 396–418. https://doi.org/10.1080/03637751.2015.1128556

Bernotat, J., & Eyssel, F. (2017). A robot at home—How affect, technology commitment, and personality traits influence user experience in an intelligent robotics apartment. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 641–646). IEEE.

Betlemidze, M. (2021). Traversing anthropocentric horizons with *Her*: Trans-corporeal surrogacy, enchantment, and disenchantment in human-machine assemblage. *Journal of Communication Inquiry, 46*(2), 206–224. https://doi.org/10.1177/01968599211041107

Bietti, L., & Skjuve, M. (2022). Remembering with my chatbot. https://doi.org/10.31234/osf.io/fpmkt

Boch, A., Ryan, S., Kriebitz, A., Amugongo, L. M., & Lütge, C. (2023). Beyond the metal flesh: Understanding the intersection between bio- and AI ethics for robotics in healthcare. *Robotics, 12*(4), 110.

Chen, Y., Jensen, S., Albert, L. J., Gupta, S., & Lee, T. (2022). Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers, 25*(1), 161–182. https://doi.org/10.1007/s10796-022-10291-4

Cohen, J. (2001). Defining identification: A theoretical look at the identification of audiences with media characters. *Mass Communication and Society, 4*(3), 245–264. https://doi.org/10.1207/S15327825MCS0403_01

Colwell, J., & Kato, M. (2003). Investigation of the relationship between social isolation, self-esteem, aggression, and computer game play in Japanese adolescents. *Asian Journal of Social Psychology, 6*(2), 149–158.

Cross, L., Kaye, L. K., Savostijanovs, J., McLatchie, N., Johnston, M., Whiteman, L., Mooney, R., & Atherton, G. (2022). Gendered violence and sexualized representations in video games: (Lack of) effect on gender-related attitudes. *New Media & Society, 26*(3), 1648–1669. https://doi.org/10.1177/14614448221075736

Cross, L., Turgeon, M., & Atherton, G. (2019). Moving with the in-crowd: Cooperation and interpersonal entrainment in in- vs. out-groups. *Current Psychology, 40*(7), 3393–3400. https://doi.org/10.1007/s12144-019-00283-0

Decety, J., & Yoder, K. J. (2016). Empathy and motivation for justice: Cognitive empathy and concern, but not emotional empathy, predict sensitivity to injustice for others. *Social Neuroscience, 11*(1), 1–14. https://doi.org/10.1080/17470919.2015.1029593

Dechering, A., & Bakkes, S. (2018). Moral engagement in interactive narrative games: An exploratory study on ethical agency in *The Walking Dead* and *Life Is Strange*. In *Proceedings of the 13th International Conference on the Foundations of Digital Games* (pp. 1–10).

Dwivedi, Y. K., Hughes, L., Baabdullah, A. M., Ribeiro-Navarrete, S., Giannakis, M., Al-Debei, M. M., & Dennehy, D. (2021). Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice, and policy. *Journal of Business Research, 130*, 209–217.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864.

Fast, E., & Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).

Ferguson, C. J., & Wang, J. C. (2022). Aggressive video games are not a risk factor for mental health problems in youth: A longitudinal study. *Psychology of Popular Media, 11*(1), 48–57.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Forni, D. (2020). *Horizon Zero Dawn*: The educational influence of video games in counteracting gender stereotypes. *Transactions of the Digital Games Research Association, 5*(1). https://doi.org/10.26503/todigra.v5i1.111

Fox, J., Bailenson, J. N., & Tricase, L. (2014). The embodiment of sexualized virtual selves: The Proteus effect and experiences of self-objectification via avatars. *Computers in Human Behavior, 29*(3), 930–938.

Frasca, G. (2001). Rethinking agency and immersion: Video games as a means of consciousness-raising. *Digital Creativity, 12*(3), 167–174. https://doi.org/10.1076/digc.12.3.167.3225

Gabbiadini, A., Riva, P., Andrighetto, L., Volpato, C., & Bushman, B. J. (2016). Acting like a tough guy: Violent-sexist video games, identification with game characters, masculine beliefs, & empathy for female violence victims. *PLOS One, 11*(4). https://doi.org/10.1371/journal.pone.0152121

Gao, X. (2024). Language education in a brave new world: A dialectical imagination. *Modern Language Journal, 108*(2), 556–562. https://doi.org/10.1111/modl.12930

Gentile, D. A., Li, D., Khoo, A., Prot, S., & Anderson, C. A. (2014). Mediators and moderators of long-term effects of violent video games on aggressive behavior: Practice, thinking, and action. *JAMA Pediatrics, 168*(5), 450–457.

Gentile, D. A., Lynch, P. J., Linder, J. R., & Walsh, D. A. (2004). The effects of violent video game habits on adolescent hostility, aggressive behaviors, and school performance. *Journal of Adolescence, 27*(1), 5–22.

Gibert, M., & Martin, D. (2022). In search of the moral status of AI: Why sentience is a strong argument. *AI & Society*, 1–12.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6*(12), 517–523. https://doi.org/10.1016/s1364-6613(02)02011-9

Greitemeyer, T. (2018). The spreading impact of playing violent video games on aggression. *Computers in Human Behavior, 80*, 216–219.

Greitemeyer, T., & Mügge, D. O. (2014). Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play. *Personality and Social Psychology Bulletin, 40*(5), 578–589.

Greitemeyer, T., & Osswald, S. (2009). Prosocial video games reduce aggressive cognitions. *Journal of Experimental Social Psychology, 45*(4), 896–900.

Greitemeyer, T., & Osswald, S. (2010). Effects of prosocial video games on prosocial behavior. *Journal of Personality and Social Psychology, 98*(2), 211.

Greitemeyer, T., Agthe, M., Turner, R., & Gschwendtner, C. (2012). Acting prosocially reduces retaliation: Effects of prosocial video games on aggressive behavior. *European Journal of Social Psychology, 42*(2), 235–242.

Grimshaw, M., Charlton, J., & Jagger, R. (2011). First-person shooters: Immersion and attention. *Journal for Computer Game Culture, 5*(1), 29–44.

Haidt, J. (2012). The intuitive dog and its rational tail. In *The righteous mind: Why good people are divided by politics and religion* (pp. 82–136). New York: Pantheon Books.

Harilal, S. (2018). Playing in the continuum: Moral relativism in *The Last of Us*. *[Sic] - A Journal of Literature, Culture and Literary Translation, (1.9)*. https://doi.org/10.15291/sic/1.9.lc.7

Harrington, B., & O'Connell, M. (2016). Video games as virtual teachers: Prosocial video game use by children and adolescents from different socioeconomic groups is associated with increased empathy and prosocial behavior. *Computers in Human Behavior, 63*, 650–658.

Hasan, Y., Begue, L., Scharkow, M., & Bushman, B. J. (2013). The more you play, the more aggressive you become: A long-term experimental study of cumulative violent video game effects on hostile expectations and aggressive behavior. *Journal of Experimental Social Psychology, 49*(2), 224–227.

Hilgard, J., Engelhardt, C. R., & Rouder, J. N. (2017). Overstated evidence for short-term effects of violent games on affect and behavior: A reanalysis of Anderson et al. (2010). *Psychological Bulletin, 143*(7), 757–774. https://doi.org/10.1037/bul0000074

Innocenti, A. (2017). Virtual reality and behavioral economics: A research agenda. *Journal of Behavioral and Experimental Economics, 69*, 41–49.

Jeffrey, T. (2020). Understanding college student perceptions of artificial intelligence. *Journal of Systemics, Cybernetics and Informatics, 18*(2), 8.

Joeckel, S., Bowman, N. D., & Dogruel, L. (2012). Gut or game? The influence of moral intuitions on decisions in video games. *Media Psychology, 15*(4), 460–485. https://doi.org/10.1080/15213269.2012.727218

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., & Severson, R. L. (2012). Do people hold a humanoid robot morally accountable for the harm it causes? In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 33–40).

Katz, J. E., & Halpern, D. (2014). Attitudes towards robots' suitability for various jobs as affected by robot appearance. *Behaviour & Information Technology, 33*(9), 941–953.

Kirpichnikov, D., Pavlyuk, A., Grebneva, Y., & Okagbue, H. (2020). Criminal liability of artificial intelligence. *E3S Web of Conferences, 159*, 04025. https://doi.org/10.1051/e3sconf/202015904025

Klimmt, C., Hefner, D., Vorderer, P., Roth, C., & Blake, C. (2010). Identification with video game characters as automatic shifts of self-perceptions. *Media Psychology, 13*(4), 323–338. https://doi.org/10.1080/15213269.2010.524911

Komatsu, T., Malle, B. F., & Scheutz, M. (2021). Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across the US and Japan. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 63–72).

Kuhn, S., Kugler, D. T., Schmalen, K., Weichenberger, M., Witt, C., & Gallinat, J. (2019). Does playing violent video games cause aggression? A longitudinal intervention study. *Molecular Psychiatry, 24*(8), 1220–1234.

Lankoski, P. (2007). Goals, affects, and empathy in games. *Philosophy of Computer Games*, 1–10.

Leveringhaus, A. (2018). Developing robots: The need for an ethical framework. *European View, 17*(1), 37–43. https://doi.org/10.1177/1781685818761016

Liao, T. (2023). Artificial love: Revolutions in how AI and AR embodied romantic chatbots can move through relationship stages. *AOIR Selected Papers of Internet Research*. https://doi.org/10.5210/spir.v2023i0.13446

López-Fernández, F. J., Mezquita, L., Etkin, P., Griffiths, M. D., Ortet, G., & Ibáñez, M. I. (2021). The role of violent video game exposure, personality, and deviant peers in aggressive behaviors among adolescents: A two-wave longitudinal study. *Cyberpsychology, Behavior, and Social Networking, 24*(1), 32–40.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117–124).

Mamak, K. (2021). Whether to save a robot or a human: On the ethical and legal limits of protections for robots. *Frontiers in Robotics and AI, 8*, 712427.

Mathur, M. B., & VanderWeele, T. J. (2019). Finding common ground in meta-analysis 'wars' on violent video games. *Perspectives on Psychological Science, 14*(4), 705–708.

McCarthy, R. J., Coley, S. L., Wagner, M. F., Zengel, B., & Basham, A. (2016). Does playing video games with violent content temporarily increase aggressive inclinations? A pre-registered experimental study. *Journal of Experimental Social Psychology, 67*, 13–19.

Mosakas, K. (2021). On the moral status of social robots: Considering the consciousness criterion. *AI & Society, 36*, 429–443.

Nijssen, S. R., Müller, B. C., Baaren, R. B. V., & Paulus, M. (2019). Saving the robot or the human? Robots who feel deserve moral care. *Social Cognition, 37*(1), 41–52.

Nomura, T., Suzuki, T., Kanda, T., Han, J., Shin, N., Burke, J., & Kato, K. (2008). What people assume about humanoid and animal-type robots: Cross-cultural analysis between Japan, Korea, and the United States. *International Journal of Humanoid Robotics, 5*(1), 25–46.

Pan, X., & Hamilton, A. F. d. C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology, 109*(3), 395–417.

Peña, J., Hernández Pérez, J. F., Khan, S., & Cano Gómez, Á. P. (2018). Game perspective-taking effects on players' behavioral intention, attitudes, subjective norms, and self-efficacy to help immigrants: The case of *Papers, Please*. *Cyberpsychology, Behavior, and Social Networking, 21*(11), 687–693. https://doi.org/10.1089/cyber.2018.0030

Prescott, A. T., Sargent, J. D., & Hull, J. G. (2018). Meta-analysis of the relationship between violent video game play and physical aggression over time. *Proceedings of the National Academy of Sciences, 115*(40), 9882–9888.

Prot, S., Gentile, D. A., Anderson, C. A., Bushman, B. J., Swing, E., Lim, K. M., & Liau, A. K. (2014). Long-term relations among prosocial-media use, empathy, and prosocial behavior. *Psychological Science, 25*(2), 358–368.

Pti. (2023, August 3). Dark side of AI: Potential consequences of emotionless machines could impact humanity. *The Economic Times*. https://economictimes.indiatimes.com/magazines/panache/dark-side-of-ai-potential-consequences-of-emotionless-machines-could-impact-humanity/articleshow/102393338.cms?from=mdr

Quantic Dream. (2018). *Detroit: Become Human* [Video Game]. Sony Interactive Entertainment.

Ratan, R. A., & Dawson, M. (2016). When Mii is me: A psychophysiological examination of avatar self-relevance. *Communication Research, 43*(8), 1065–1093. https://doi.org/10.1177/0093650215570652

Sestir, M. A., & Bartholow, B. D. (2010). Violent and nonviolent video games produce opposing effects on aggressive and prosocial outcomes. *Journal of Experimental Social Psychology, 46*(6), 934–942.

Sheliazhenko, Y. (2018). Computer modeling of personal autonomy and legal equilibrium. *Advances in Intelligent Systems and Computing*, 74–81. https://doi.org/10.1007/978-3-319-91192-2_8

Sherry, J. L. (2001). The effects of violent video games on aggression: A meta-analysis. *Human Communication Research, 27*(3), 409–431.

Simon, H. A., Newell, A., & Shaw, J. C. (2019). The processes of creative thinking. *The Journal of Creative Behavior, 53*(2), 141–148.

Smiley, J. (2008). *13 ways of looking at the novel*. Anchor.

Stapels, J. G., & Eyssel, F. (2021). Let's not be indifferent about robots: Neutral ratings on bipolar measures mask ambivalence in attitudes towards robots. *PLOS One, 16*(1), e0244697.

Sun, T., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly, 36*, 368–383. https://doi.org/10.1016/j.giq.2018.09.008

Tear, M. J., & Nielsen, M. (2013). Failure to demonstrate that playing violent video games diminishes prosocial behavior. *PLOS One, 8*(7), e68382.

Thompson, M. M., Zanna, M. P., & Griffin, D. W. (1995). Let's not be indifferent about (attitudinal) ambivalence. *Attitude Strength: Antecedents and Consequences, 4*, 361–386.

Turkle, S. (2010). In good company? In Y. Wilks (Ed.), *Close engagements with artificial companions* (pp. 3–10). John Benjamins Publishing Company.

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

Valenzuela, A., & Raghubir, P. (2009). Position-based beliefs: The center-stage effect. *Journal of Consumer Psychology, 19*(2), 185–196.

Van Loon, M. H., Roebers, C. M., & de Bruin, A. B. (2018). Children's understanding of the differential validity of their feeling-of-knowing judgments. *Developmental Psychology, 54*(10), 1848–1863.

Van Looy, J., Courtois, C., De Vocht, M., & De Marez, L. (2012). Player identification in online games: Validation of a scale for measuring identification in MMOGs. *Media Psychology, 15*(2), 197–221. https://doi.org/10.1080/15213269.2012.674917

Vishwanath, A., Bøhn, E. D., Granmo, O., Maree, C., & Omlin, C. (2022). Towards artificial virtuous agents: Games, dilemmas, and machine learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2208.14037

Wright, M., & Wachs, S. (2021). Moderation of technology use in the association between self-isolation during the COVID-19 pandemic and adolescents' romantic relationship quality. *Cyberpsychology, Behavior, and Social Networking, 24*(7), 493–498. https://doi.org/10.1089/cyber.2020.0729

Yao, M. Z., Mahood, C., & Linz, D. (2010). Sexual priming, gender stereotyping, and likelihood to sexually harass: Examining the cognitive effects of playing a sexually explicit video game. *Sex Roles, 62*(1–2), 77–88. https://doi.org/10.1007/s11199-009-9695-4

Yee, N., & Bailenson, J. (2007). The Proteus effect: The effect of transformed self-representation on behavior. *Human Communication Research, 33*(3), 271–290.

Yigitcanlar, T., Li, R., Inkinen, T., & Paz, A. (2022). Public perceptions on application areas and adoption challenges of AI in urban services. *Emerging Science Journal*. https://doi.org/10.28991/esj-2022-06-06-01