Is Deception in Emulated Empathy Innately Bad?

Bakir, Vian; McStay, Andrew

IEEE Standards White Paper

Published: 13/12/2024

Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):
Bakir, V., & McStay, A. (2024). Is Deception in Emulated Empathy Innately Bad? *IEEE Standards White Paper*.

IEEE SA
STANDARDS
ASSOCIATION

IEEE SA
WHITE PAPER

IEEE P7014.1

IS DECEPTION IN EMULATED EMPATHY INNATELY BAD?

Authored by

Vian Bakir, *Working Group member of IEEE P7014.1, Professor of Journalism & Political Communication, Bangor University, U.K.*

Andrew McStay, *Chair IEEE P7014.1, Professor of Technology & Society, Bangor University, U.K.*

IEEE

TRADEMARKS AND DISCLAIMERS

IEEE believes the information in this publication is accurate as of its publication date; such information is subject to change without notice. IEEE is not responsible for any inadvertent errors.

The ideas and proposals in this specification are the respective author's views and do not represent the views of the affiliated organization.

The Institute of Electrical and Electronics Engineers, Inc. 3 Park Avenue, New York, NY 10016-5997, USA

PDF: STDVA27531 979-8-8557-1563-7

NOTICE AND DISCLAIMER OF LIABILITY CONCERNING THE USE OF IEEE SA DOCUMENTS

This IEEE Standards Association ("IEEE SA") publication ("Work") is not a consensus standard document. Specifically, this document is NOT AN IEEE STANDARD. Information contained in this Work has been created by, or obtained from, sources believed to be reliable, and reviewed by members of the activity that produced this Work. IEEE and the IEEE P7014.1 working group expressly disclaim all warranties (express, implied, and statutory) related to this Work, including, but not limited to, the warranties of: merchantability; fitness for a particular purpose; non-infringement; quality, accuracy, effectiveness, currency, or completeness of the Work or content within the Work. In addition, IEEE and the IEEE P7014.1 working group disclaim any and all conditions relating to: results; and workmanlike effort. This document is supplied "AS IS" and "WITH ALL FAULTS."

Although the IEEE P7014.1 working group members who have created this Work believe that the information and guidance given in this Work serve as an enhancement to users, all persons must rely upon their own skill and judgment when making use of it. IN NO EVENT SHALL IEEE SA OR IEEE P7014.1 WORKING GROUP MEMBERS BE LIABLE FOR ANY ERRORS OR OMISSIONS OR DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS WORK, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Further, information contained in this Work may be protected by intellectual property rights held by third parties or organizations, and the use of this information may require the user to negotiate with any such rights holders in order to legally acquire the rights to do so, and such rights holders may refuse to grant such rights. Attention is also called to the possibility that implementation of any or all of this Work may require use of subject matter covered by patent rights. By publication of this Work, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. The IEEE is not responsible for identifying patent rights for which a license may be required, or for conducting inquiries into the legal validity or scope of patents claims. Users are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. No commitment to grant licenses under patent rights on a reasonable or non-discriminatory basis has been sought or received from any rights holder.

This Work is published with the understanding that IEEE and the IEEE P7014.1 working group members are supplying information through this Work, not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought. IEEE is not responsible for the statements and opinions advanced in this Work.

TABLE OF CONTENTS

IS DECEPTION IN EMULATED EMPATHY INNATELY BAD?

1. EXECUTIVE SUMMARY AND RECOMMENDATIONS

This white paper draws on interdisciplinary academic knowledge and real-world examples to think through the parameters of debate about deception and empathy-based human-artificial intelligence (AI) partnering. While a common human reaction is to instinctively reject, and be repulsed by, the idea of deception in AI partners, this paper offers an overview and analysis of a wider range of positions on such deception before coming to a considered conclusion. Mindful that deception in relation to humans and in AI are broad topics, the paper draws on studies that examine deception in both human-human relationships and human-AI/robot relationships. The paper then considers deception in human-AI partnerships, drawing on IEEE P7014.1, Recommended Practice for Ethical Considerations of Emulated Empathy in Partner-based General-Purpose Artificial Intelligence Systems. This white paper is a first take at distilling the issues, bearing in mind that at the time of writing (late 2024), this recommended practice is still under development and the roll-out of human-AI partners displaying qualities of empathy is in flux. The key points are as follows:

1. Nature of emulated empathy: Emulated empathy, or weak empathy, involves the computational sensing, reading, and profiling of human emotions to mimic empathic behavior. While it aims to simulate strong empathy, it remains limited in terms of genuine emotional engagement, co-presence, and responsibility.

2. Ethical issues: Of the many ethical issues surrounding emulated empathy, this paper focuses on the potential for deception. Intentional deception, particularly *superficial state deception,* occurs when an AI partner signals empathy that it inherently lacks.

3. Deception acceptability: Deception in AI-human partnerships can be acceptable under specific circumstances. However, trust, loyalty, and fiduciary responsibility are paramount, and deceptive practices that violate these principles remain unacceptable.

4. Ethical oversight: There is an important need for ethical oversight to distinguish acceptable uses of emulated empathy from practices that exploit users. This requires ongoing testing and scrutiny of general-purpose artificial intelligence (GPAI) systems to ensure that anthropomorphic and zoomorphic deception serves rather than misleads and exploits.

5. Recommendations: This white paper recommends incorporating ethical guardrails and guidelines in GPAI systems to reduce deceptive practices, broadly emphasizing transparency, accountability, and avoidance of conflating weak with strong empathy. Furthermore, GPAI partners must be programmed to align with user expectations and avoid undermining their autonomy.

2. INTRODUCTION AND SCOPE

Deception, defined here as misleading others into believing something that the deceiver does not believe in (DePaulo et al. 2003 [17][1]), is often viewed negatively. This is because it involves misleading or manipulating others, which can erode trust, harm relationships, and cause ethical concerns. However, deception is not always considered inherently bad; its moral and ethical evaluation arguably depends on the context, intent, and consequences. Similarly, if one views deception through different philosophical prisms, one sees different outcomes. For example, a deontologist may say that as a rule, any deception in a human–empathic AI partner relationship is always bad; while a utilitarian will ask if a greater overall good is served by deception in this relationship; a virtue ethicist may conclude that it depends on the character and intentions behind the system; a pragmatist may say that the moral acceptability of deception depends on the context; and a Shintoist may be more open to the presence of a subject in the object, depending on the intent and curation of the system.

The purpose of this white paper is to think through the parameters of debate about deception and empathy-based human-AI partnering looking through the prism of IEEE P7014.1 issues, but mindful that deception in relation to AI is a broader topic. Funded by a UK Responsible AI award[2] and organized by Project AEGIS,[3] it explicitly draws on academic thought from several disciplines to distil a range of ethical positions. This is done to advance the work of the IEEE P7014.1 working group, an IEEE effort to develop a recommended practice to define ethical considerations and good practices regarding the use of emulated empathy in GPAI systems for human-AI partnerships. The white paper does not address all possible issues raised by empathy-based human-AI partnerships, but instead focuses on the issue of deception. Similarly, the white paper solely represents the views of the authors and does not necessarily represent a position of either the IEEE P7014.1 working group, the IEEE SSIT Standards Committee, the IEEE, or the IEEE Standards Association.

2.1. IEEE P7014.1 WORKING GROUP

For context, the IEEE P7014.1 working group for Ethical Considerations of Emulated Empathy in Partner-based General-Purpose Artificial Intelligence Systems was launched in 2024. It derives from a recently approved IEEE Std 7014™-2024 [24], which principally focuses on biometric dimensions of emulated empathy, such as that based on narrow AI (computer vision and labeling of facial expressions, for example).

[1] The numbers in brackets correspond to those of the references in Section 6.

[2] Available: https://rai.ac.uk/

[3] Available: https://automatingempathy.ai/

The nascent IEEE P7014.1 aims to provide clear and practical ethical recommendations for the conception, design, and lifecycle of GPAI applications that emulate empathic abilities to enable human-AI partnerships.

The need for IEEE P7014.1 is as follows:

- There is no explicit ethical recommended practice (or guidance or standard) regarding the intersection between emulated empathy and GPAI. This matters because partnership-based applications founded on GPAI are likely to increase in the future, inviting and deepening diverse ethical questions about enmeshing people and technology.

- There is little global agreement on what detailed ethical standards (or explicit law or human rights policy) should be for these types of human-AI relationships.

- There is global value in diversifying ethical frames of reference to account for human-AI partnerships.

- Empathy emerges in subtle, complex, and poorly understood ways, potentially missed by less-focused governance instruments.

The IEEE P7014.1 working group recognizes that partnerships with AI will reshape industries, redefine individual and social capabilities, and challenge current conceptions of human and machine collaboration. Already these human-AI collaborations are increasingly popular, marketed as empathic partners, assistants, copilots, and similar. IEEE P7014.1 provides recommendations on one element of this relationship—empathy. The authors recognize that there is scope for social good in partners that can help with diverse aspects of everyday life (companionship, work, therapy, education, life coaching, legal problems, fitness, entertainment, and more), but also that including empathic characteristics raises ethical questions. Some of these questions are familiar and apply to AI technologies in general; for instance, issues of transparency, accountability, bias, and fairness. Other ethical questions are more specific and unique to GPAI applications that emulate empathic abilities to enable human-AI partnerships: these include psychological interactions and dependencies (and that AI models can be prompted for emotional and empathic responses), child/age appropriateness, fiduciary issues, animism, and manipulation through partnerships with GPAI systems (Bakir et al. 2024 [7], McStay et al. 2024 [37]).

2.2. TERMS OF REFERENCE

2.2.1. DECEPTION

Arising from work from the P7014.1 working group, this white paper considers the core ethical issue of **deception**. There are many definitions of deception across human, animal, and robotic domains (Arkin 2018 [2], Wagner and Arkin 2010 [64]), but it is defined here as **misleading others into believing something that the deceiver does not believe in** (DePaulo et al. 2003 [17]).

2.2.2. EMPATHY

The premise of computational empathy is rightfully controversial, not least because empathy is a quintessentially human trait of understanding and feeling the emotions and experiences of others. This white paper does not linger on definitions; nor does it extensively recount debates associated with the premise of computational empathy or the nature of GPAI. The empathy question is a thorny one, and readers are encouraged to consult *Automating Empathy* (McStay 2023 [33]), which addresses perspectives and problems associated with the idea of empathy, computers, and automation. In lay terms, however, **empathy means to share or understand others' emotions, feelings, or experiences**. Consequently, any empathy performed by a machine is considered an *emulation* of human empathy rather than the same process (IEEE Std 7014-2024 [24], Montemayor, Halpern, and Fairweather 2021 [39]). As expounded in the IEEE SA white paper, "Ethics and Empathy-Based Human-AI Partnering: Exploring the Extent to Which Cultural Differences Matter When Developing an Ethical Technical Standard" (McStay et al. 2024 [37]), at the heart of IEEE P7014.1 is an understanding of empathy as split into two kinds:

- **Weak empathy**: The practice of sensing, reading, profiling, judging, or making rules about the states and behavior of people, and interacting effectively or otherwise with intimate dimensions of human life.
- **Strong empathy**: Empathy is strong when it includes weak interpretive and interactional abilities but also elicits the experience of fellow-feeling, solidarity, co-presence, and responsibility toward the other.

Weak and limited technical means may be used by organizations to try to emulate strong empathy. Put otherwise, **emulated empathy is the use of AI-based technology to try to copy, simulate, mimic, and display the appearance of strong empathy** (Bakir et al. 2024 [7]). While there is intra-disciplinary and cross-disciplinary debate about the nature of empathy (McStay 2018 [34]), advanced language models capable of generating text can mimic empathic responses based on its training of text data that include conversations and interactions where empathy is displayed.

Schaaff et al. (2023 [46]), for example, investigated the extent to which ChatGPT (the GPT-3.5 model) can exhibit empathic responses and emotional expressions. They did not suggest that ChatGPT has the same sort of empathy as human empathy, but they did explore how weak empathy by ChatGPT may be used to understand and express emotions, give suitable emotional responses, and impart a sense of an empathic personality, with a view to emulating strong empathy. Their results find ChatGPT is worse than the average of healthy humans, but it scores better than people who have been diagnosed with Asperger syndrome/high-functioning autism (Schaaff et al. 2023 [46]). In May 2024, the *New Scientist* reported that OpenAI's announcement of its newest artificial intelligence model (called GPT-4o), would soon power some versions of the company's ChatGPT product, with the upgraded ChatGPT being able to swiftly respond to text, audio, and video inputs from its real-time conversational partner—all while speaking with empathic inflections and wording that convey a strong sense of emotion and personality (Hsu 2024 [23]).

Importantly, such systems are increasingly multimodal. In the case of weak empathy, this means that biometric and large language model (LLM) elements in a system function as one. For example, prosody and voice analysis of emotion in speech are conjoined with sentiment analysis, so sentiment analysis of *what* is said can modify and improve inferences of *how* it is said.

2.2.3. GENERAL-PURPOSE AI (GPAI)

GPAI is sometimes referred to as a "foundation model," but what is key for IEEE P7014.1 is that these involve **systems with a wide range of possible uses**. Although the set-up period of IEEE P7014.1 was motivated by the mass rollout of LLMs in 2022, multimodal systems (including those of a biometric sort) are recognized and expected to become the norm. These "partners" may involve wide-ranging modalities and topics for human-system interaction.

2.2.4. PARTNERS

While "assistants," "companions," "teaming," "agents," "chatbots," and "partners" are overlapping figures of speech, these metaphors signal changes in how people use AI-based technologies and live in relation to them, and ideas about how people will interface with those technologies. They are not the same, however: teaming, for example, goes a step further than assisting, involving "the emerging capabilities of AI technologies [that] allow them to be implemented directly in team processes with other artificial and human agents or to overtake functions that support humans in a way team partners would" (McStay 2023 [33], Sorell and Draper 2017 [52]). With **"partner," IEEE P7014.1 has two agents in mind: a living person and an AI system that functions as a partner to achieve a goal of some sort**. These goals may be short-term or long-term, and important or unimportant. Empathic partners gauge the disposition of the human user and express through language and other modalities a synthetic disposition that mimics the turn-taking norms of human interactions. Although IEEE P7014.1 is conceived in terms of a human-AI pairing, it is not restricted to this. The AI element may work with several people, and a person will likely have several AI partners (some general and others more task-specific). Partners based on LLMs to communicate with humans are prominent examples of how empathic partnerships can arise. Replika, for example, uses natural language processing (NLP) and machine learning to emulate empathy by engaging in conversations that mirror human interactions. Although it lacks genuine emotional understanding, Replika offers users personalized responses by remembering previous interactions, adapting to user preferences, and providing emotional support through positive feedback, motivational messages, and affirmations (McStay 2022 [35]). Notably, empathic AI partners may have anthropomorphic qualities that are mostly perceived involuntarily by humans (Danaher 2020 [16]). A consequence of this is that despite rational understanding regarding the nature of a chatbot or partner, one may not treat the partner as one would any other item of media.

3. DECEPTION AMONG HUMANS, ROBOTS/AI, AND EMPATHIC HUMAN-AI PARTNERSHIPS

This white paper is motivated in part by a desire to overcome simplistic critiques of deception in human-technology interaction. This is not about somehow weakening resolve in addressing problematic system conceptualization and design, but to add granularity, context, and nuance to collective thought about deception in relation to emulated empathy. Consequently, interested actors (such as the technology industry, standards developers, and policymakers) may better identify what is a genuine problem versus what is harmless, and perhaps even get to a position of stating what is wanted from this class of technology.

To do this, interested parties need to be sensitized to the social history of deception, not least in **human-human deception**. Deception among people can take many forms, but perhaps the most common are lying, distortion, omission, and misdirection (Bakir et al. 2019 [8]). These are detailed as follows:

- *Deception through lying* is defined as making a statement that is known or suspected to be untrue in order to mislead.
- *Deception through omission* involves withholding information to make the promoted viewpoint more persuasive.
- *Deception through distortion* involves presenting a statement in a deliberately misleading way (e.g., by exaggeration or de-emphasizing information) to support a viewpoint.
- *Deception through misdirection* entails producing and disseminating true information intended to direct attention away from problematic issues (Bakir et al. 2019 [8]).

In each of these cases of deception among humans, the deception is *intentional*. Indeed, Chadwick and Stanyer's (2021 [13]) synthesis of 30 years of social science literature on deception highlights the importance of intentionality. This leads them to define deception (among humans) as when an identifiable actor's prior intention to mislead results in attitudinal or behavioral outcomes that correspond with the prior intention.

Getting closer to this white paper's core interest, Danaher (2020 [16]) defines **robotic deception** as arising whenever a robot, as an embodied artificial agent, makes a representation or sends a signal (in speech, behavior, or physical appearance) that creates a misleading or false impression among those who interact with the robot. Whereas in human-human relationships, deception is often strongly related to intent—and psychological definitions underscore this, e.g., Zuckerman, DePaulo, and Rosenthal (1981) [68]—there is more debate about whether such misleading is always intentional and deliberate in human-robot relationships. On the one hand, Sorell and Draper (2017 [51]) propose that deception in robots requires "the intentional creation of false beliefs" (for instance, on the part of the robot developer). On the other hand, Sætra (2021 [45]) regards deception as any action that misleads someone and argues that with social robots (i.e., robots with an explicit aim of interacting with human beings), this may easily occur *without the intention* to mislead. Similarly, Sharkey and Sharkey (2021 [50]) argue that if the behavior and appearance of a robot leads to people believing that a robot has cognitive abilities or that it cares for and loves them, then they are being deceived whether or not anyone intended to deceive them.

As noted earlier, organizations may try to emulate strong empathy, using AI-based technology to try to copy, simulate, mimic, and display the appearance of strong empathy. However, any attempt to **emulate empathy** risks being **misleading by default** and impacting long-agreed principles of non-interference with human autonomy and decision-making. Issues of emotion and psychological entanglement, anthropomorphism, animism, and quasi-subjectivity, all raise a question about whether deception is a fundamental characteristic of GPAI systems for empathic human-AI partnerships. The key issue is the misleading of others, and when this can be said to have taken place. This does not just refer to text and chatbots that may use personal pronouns (e.g., "I think," "I feel," or "I hope") or claims to empathy (e.g., "I care about what happens to you"), but it also involves gestures, actions, omissions, or other forms of communication that mislead (Bakir et al. 2024 [7]).

4. ACCEPTABILITY OF DECEPTION

Developing guidance for the ethical design and use of technological systems, such as IEEE P7014.1, necessitates investigating whether deception is acceptable, or even preferable, and on what terms. The question of whether deception can ever be acceptable has been debated as far back as the ancient Greeks and continues to be debated today. This clause outlines various positions on the acceptability of deception in human-human relationships (4.1) and the acceptability of deception in human-AI/robot relationships (4.2). Subclause 4.3 applies some of these positions to the issue of empathic human-AI partnerships while reflecting on the emerging recommended practice IEEE P7014.1, as well as studies on empathic human-AI partnerships.

4.1. DECEPTION AND ITS ACCEPTABILITY IN HUMAN-HUMAN RELATIONSHIPS

Among human-human relationships, when is deception acceptable? The following outlines several positions on the acceptability of deception (summarized in TABLE 1).

TABLE 1 **Acceptability of deception in human-human relationships**

Acceptable	Unacceptable	Grey areas
The noble lie	Democratically unacceptable	Cultural variation on norms of deception
Pro-social, collaborative, or white lie	Improper in principle as it contravenes the condition of truthfulness underpinning the social contract and law in general	
Utilitarian lie	Damages social trust	
To maintain harmony in a social group	Violates freedom and autonomy of addressee	
For the entertainment of knowing audiences		

The following are five positions that view deception as acceptable in human-human relationships. It can be argued that these are forms of *consensual deception,* in that people, in certain circumstances, consent to being deceived.

- **The noble lie (Plato)**. The Platonic tradition seeks to justify lying under certain circumstances. Plato explicitly defended lying in his most important work of moral and political philosophy, the *Republic.* He advocated that rulers of the ideal state are allowed to deceive their people and enemies via lies and myths to *protect public welfare and social order* and to benefit the *polis* (Mahon 2019, p. 19–24 [30]). Plato says, "A high value must be set upon truthfulness … (and) … If anyone, then, is to practice deception, either on the country's enemies or on its citizens, it must be the Rulers of the commonwealth, acting for its benefit; no one else may meddle with this privilege" (Plato 1965, p. 78 [42]). Related, international relations scholars suggest that deception by leaders to other countries and even to their own people is necessary for good *strategic reasons of state* arising from the dangers inherent in the international political system (Bakir et al. 2019, p. 532 [8]).
- **Pro-social lie, collaborative, or white lie**. Deception may be acceptable if it has *benevolent or socially harmless motives or consequences* (Dietz 2019, p. 298 [18]). Whereas lies, in general, are considered to be morally improper, a white lie may be an excusable falsehood, not meant to injure anyone and of little moral import (Bok 1999 [10]): for example, a false excuse in favor of politeness. A white lie may even be morally demanded if it is used to resist attacks on one's life and property and to refuse unwarranted curiosity (Schopenhauer (1903 [1840] [49], cited in Dietz 2019, p. 294 [18]). Placebos are another example of deceptions created with the good intention of helping or protecting the deceived (Bok 1999 [10]).

- **Utilitarian lie**. Most utilitarians hold that lying is generally improper because of its harmful effects, especially on social trust, but that it might be *justified in some cases of conflicting obligations, rights, or interests, or because of its overall better consequences* (Dietz 2019, p. 288 [18]).
- **To maintain harmony in a social group**. Blue lies are a type of lying emanating from an emphasis on collectivist norms, and *modesty-related lying* in public. This is common especially in East Asian contexts, possibly because in these contexts, "publicly calling attention to one's accomplishments violates norms about maintaining harmony within one's social group" (Heyman and Lee 2012, p. 170 [22] cited in Terkourafi 2019, p. 392 [56]).
- **For the entertainment of knowing audiences**. While human cooperation is based on initial expectations of truthfulness in most non-combative situations, some contexts make it reasonable to expect deceit such as when watching a *theatre play or playing games* based on deception (Sætra 2021 [45]).

The following are four positions where deception in human-human relationships can be seen as unacceptable:

- **Democratically unacceptable**. The noble lie is unacceptable in democracies as *citizens should be able to cope with the truth* from their political representatives. In modern democratic systems, benevolent political lies contradict the idea of representation and sovereignty of the people (Dietz 2019, p. 297 [18]).
- **Improper in principle as contravenes the condition of truthfulness underpinning the social contract and law in general**. Lies are seen as morally parasitic and antisocial per se and cannot be justified under any circumstances. This is because lying *contravenes the condition of truthfulness on which the social contract and law in general are based*. The binding nature of declarations rests on the unconditional duty of truthfulness. If the right to lie in certain cases were admitted, it would undermine the general law and the integrity of the social community (Kant 1949 [1797] [28]), cited in Dietz 2019, p. 288, p.291 [18]). Additionally, the Platonic/Socratic intuition avoids conduct that is rooted in deception because virtue should be about reality, not illusion. As Carr puts it: "for Socrates and Plato, the chief route to virtue is accurate perception of the world, ourselves and our relations with others and the moral wisdom of virtue requires knowledge of objective truth that frees us from the bonds of ignorance and deception" (Carr 2020, p. 1382, cited in Coeckelbergh 2021, p. 650 [14]).
- **Damages social trust**. Lies *violate the norm of truthfulness*, which harms the social situation of trust and disrupts the autonomy of the individual addressee (Dietz 2019, p. 293 [18], Williams 2002 [66]). According to Mill (1987 [1861], p. 295 [38]) lying means to "deprive mankind" because it damages the reliance which they can place in each other's word.

- **Violates the freedom and autonomy of the addressee**. Schopenhauer (1969 [1859], p. 337 [48], cited in Dietz 2019, p. 294 [18]) evaluates lies as well as violence as morally reprehensible means to *force a person to serve somebody else's will*. While violence uses physical means, lying distorts the addressee's cognition. This is a form of manipulation.

There are also grey areas when it comes to the acceptability of deception in human-human relationships.

- **Cultural variation on norms of deception.** People with *different cultural backgrounds place different weights on key elements of lying*. Different cultural norms (such as individualist versus collectivist) also lead to different perspectives on lying (Nishimura 2019 [41]). For instance, collectivistic people might use a lie to maintain harmony in the group they belong to. The lie might be accepted by the group (but not so by individualistic people) if the lie is considered to be for the greater good (Kim et al. 2008 [29], cited in Nishimura 2019, p. 566 [41]). Nishimura (2005 [41]), for example, reports cultural differences between Japanese and New Zealanders in their acceptance of lies. The (collectivist) Japanese recipients were relatively lenient toward lies whereas the New Zealand recipients (individualist) were angry or resentful overall. However, caution is needed when drawing any generalizations as empirical studies on lying in individualist versus collectivist cultures are inconsistent (Nishimura 2019, p. 573 [41]). More generally, different cultures and social norms may not rely on non-deception as the norm in a given relationship. For example, in some parts of the world, it would not be considered deceptive for a market seller to claim that the price of a product is much higher than it actually is: such behavior could be the normal form of interacting between seller and buyer, and the buyer is expected to know this (Sætra 2021 [45]).

The acceptability (or not) of deception in human-human relationships, then, depends on several factors. These include the **deceiver's intent** (for instance, the noble lie to benefit the *polis*; prosocial, collaborative, or white lies with benevolent or socially harmless motives or consequences; to maintain harmony in a social group; and for the entertainment of knowing audiences). Acceptability also hinges on the **principles and norms against which the deception is judged** (for instance, democratic norms; the intrinsic importance of truthfulness as the basis of the social contract and law in general; the freedom and autonomy of the addressee; and varying cultural norms regarding the deception. Acceptability further depends on the **impacts of the deception**. Positively, utilitarian lies may be justified because of their overall better consequences; more negatively, deception may damage social trust.

4.2. DECEPTION AND ITS ACCEPTABILITY IN HUMAN-AI/ROBOT RELATIONSHIPS

Among human-AI/robot relationships, when is deception acceptable? This subclause outlines a number of positions on the acceptability of deception (summarized in TABLE 2).

TABLE 2 Acceptability of deception in human-AI/robot relationships

Acceptable	Unacceptable	Grey areas
For the entertainment of knowing audiences	Self-deception involved in an imaginary relationship with a robot is inherently wrong and violates a duty to see the world as it is	Any deception must be proportionate to the benefit for the well-being of the manipulated
To facilitate effective interaction between humans and robots, e.g., “superficial state deception”	Intention to deceive by a deceiver that has malign, or at least self-serving, intentions	“Hidden state deception”
	Dishonest anthropomorphism	
	Harmful impacts on the value of reciprocity (mutual care) across society	
	Harmful impacts on trust (misplaced trust; erosion of trust)	

Just as there are cases in human-human interaction where deception is justifiable (see 4.1), so also there are cases where AI/robotic deception can be ethically acceptable. Two main positions justifying deception in human-AI/robot relationships, seeing these as *consensual forms of deception*, are the following:

- **For the entertainment of knowing audiences**. People are both entertained and aware that the illusion of sentience created by a social robot is not real. Coeckelbergh (2018 [14]) argues that any deception or illusion created by information technology is the result of a performance “co-created and co-performed by humans (magician/designer and spectator/user) and non-humans (robots and other machines, artefacts and devices)” (Coeckelbergh 2018, p. 78 [14]). As a result, Coeckelbergh (2018, p. 80 [14]) argues that the term *deception* is unhelpful in discussing human-robot use and interaction, and that it is better to evaluate the use of (social) robots in terms of the success and ethical quality of the *performances* and their consequences.
- **To facilitate effective interaction between humans and robots**. This is one of the primary objectives of the field of human-robot interactions (Arkin et al. 2011 [3], Hancock et al. 2011 [21]). For instance, search and rescue robots may need to deceive in order to calm or receive cooperation from a panicking

victim; and socially assistive robots providing personalized care for Alzheimer's patients may need to deceive the patient to elicit the patient's cooperation in their treatment (Arkin et al. 2011 [3]). Subclause 4.1 highlighted how prosocial deception can function as lubrication in human-human relationships and deception is perhaps even more important as a lubricant in human-robot interactions if the goal is to make this interaction more human-like. For instance, a robot engages in "superficial state deception" when it emits signals that imply that it has capacities or characteristics it does *not* have (Danaher 2020 [16]). This could be the case if the robot was, for example, programmed to appear sad when it delivered bad news to a human. This might be perceived as the presence of some form of empathy, even if there is no trace of empathy or sadness to be found in the robot. This kind of deception might be crucial for facilitating efficient human-robot interactions, but it is nevertheless deceptive (Sætra 2021 [45]).

Five positions that view deception in human-AI/robot relationships as unacceptable are as follows:

- **Self-deception involved in an imaginary relationship with a robot is inherently wrong and violates a duty to see the world as it is**. Sullins (2012 [55]) posits that robotic companions can only meet one's physical and emotional needs on the surface, but a robot cannot truly satisfy these needs even if the human is deceived into thinking the robot can. Many robot ethics scholars find something disturbing about unidirectional social relationships (where the social robot cannot bond in the same way a human does even if the human thinks it can), especially when this starting point is used to steer the design and development of social robots (van Wynsberghe 2020 [67]). Sparrow (2002 [53]) argues that the self-deception involved in an imaginary relationship with a robot is inherently wrong and *violates a duty to see the world as it is*.
- **If there is self-serving or malign intention to deceive**. Deception in robotics is wrong when the deceiver wants to manipulate the deceived person to do something that serves the interests of the deceiver: in other words, when the deceiver has malign, or at least self-serving, intentions (Sorell and Draper 2017 [51]). Danaher argues that robotic deception occurs, "whenever a robot (a) uses some signal (speech act; anthropomorphic cue) in a way that (b) violates the expectations/norms we usually associate with the use of such signals (most commonly by using the signal in a way that is objectively false or misleading), where (c) this serves some ulterior end that can either be traced to the robot themselves or some third party" (2020, p. 121 [16]).
- **Dishonest anthropomorphism** in the design and operation of robots refers to the tendency for robots to use anthropomorphic appearance and behavior to "trick" people into believing that they are human-like (Kaminsky et al. 2017 [25], Leong and Selinger 2019 [30], all cited in Danaher 2020, p. 118 [16]). Turkle

(2007 [60], 2010 [61]) argues that simulated affect in social robots—e.g., robots that express concern for their users—is ethically dubious because it tricks people into thinking that there is some mutuality in the relationship they have with a robot when there is not. However, this white paper is sensitive to the argument that the mutuality concern may be over-stretched, reaching into consensual forms of deception. Leong and Selinger (2019 [30], cited in Danaher 2020, p. 118 [16]) develop a taxonomy of the different ways in which anthropomorphic cues can give misleading impressions of what a robot is really up to. They worry that such dishonest anthropomorphism can be leveraged by malicious actors to surveil and manipulate humans in undesirable ways. Sharkey and Sharkey (2011 [51]) see “efforts to develop features that promote the illusion of mental life in robots as forms of deception,” since current robots have neither minds nor experiences. For many authors (e.g., Bryson 2018 [11], cited in Tigard et al. 2020 [58]), AI systems and robots must remain explicitly “robotic”—that is, their artificial, possibly mechanical, nature should be readily apparent to all users; otherwise, humans are at risk of harm from deception. Designing AI systems to display human emotions, for example, is seen as wrong because by doing so, others are encouraged to incorrectly consider artifacts as deserving moral status, such as agency.

- **Harmful impacts on the value of reciprocity (mutual care) across society**. Van Wynsberghe (2020 [67]) argues that social robots should not create faux reciprocal relationships between humans and robots, that is, relationships that are deceptive and unidirectional at their core. This is because creating robots with the intention to deceive users threatens the value of reciprocity across society, and a world without reciprocity (i.e., without mutual care), is unsustainable. *Reciprocity* can be simply defined as the “Golden Rule”: do unto others as you would have them do unto you (Kahn et al. 2006 [26], cited in van Wynsberghe 2020 [67]), or “If you do something for me I will do something for you” (Sandoval et al. 2015 [46], cited in van Wynsberghe 2020 [67]).
- **Harmful impacts on trust (misplaced trust; erosion of trust)**. The appearance and behavior of a robot can lead to an overestimation of its functionality or to an illusion of sentience or cognition that can promote *misplaced trust* and inappropriate uses such as care and companionship among the vulnerable or children (Turkle 2011 [62], Sharkey and Sharkey 2021 [50]). If a person believes that a social robot has emotions and cares about them, they are being deceived, even if no one explicitly intended that belief (Sharkey and Sharkey 2021 [50]). Sætra (2021 [45]) argues robot deception is problematic for cultural sustainability as, in societies built on trust and the expectancy of truthful signals, *repeated deception will erode this trust and change the culture and social norms.*

In addition to unacceptable forms of deception that violate consensual deception, there are two grey areas when it comes to the acceptability of deception in human-AI/robot relationships.

- **Hidden state deception** is where the robot uses a deceptive signal to conceal or obscure the presence of some capacity, function, or internal state that it has (Danaher 2020 [16]). An example is a robot that turns its head away from you—leading you to think that it cannot "see" you—while it has sensors and eyes that can record at any angle (Kaminski et al. 2016 [25]). Similarly, an AI system or robot could be designed to be capable of *recognizing* the interpersonal reactions—the social and emotional communications—of humans within a specified purview without being programmed to *exhibit* human emotions (Tigard et al. 2020 [58]). Whether it is an ethically disturbing form of deception depends on the ulterior motive this concealment serves. In general, if the ulterior motive serves some greater good then it may be ethically permissible, otherwise it is not. Tuncer, et al. (2023 [59]) argue that to avoid deception, robot designers recommend that robots' interactional capacities be perceptible in their appearance and conduct.
- **Any deception must be proportionate to the benefit for the well-being of the manipulated**. If deception (or manipulation) in social robots is allowed for "benign" purposes (for instance, making the robot appear to be emotional to make it more fun to interact with), the severity of the manipulation must be proportionate to the benefit for the well-being of the manipulated (Fronemann et al. 2021 [20]).

The acceptability (or not) of deception in human-AI/robot relationships depends on many factors. These include the **deceiver's intent**. For instance, more positively, the intent may be to facilitate effective interaction between humans and robots ("superficial state deception"); for the entertainment of knowing audiences; or to serve the greater good. More negatively there may be a self-serving or malign intention to deceive, or there may be "dishonest anthropomorphism" in the design and operation of robots to "trick" people into believing that they are human-like. Acceptability also hinges on the **principles against which the deception is judged** (for instance, where self-deception involved in an imaginary relationship with a robot is regarded as inherently wrong; or if it is stipulated that the deception must be proportionate to the benefit for the well-being of the manipulated). Acceptability is further affected by the **impacts of deception**: for instance, harmful impacts on the value of reciprocity (mutual care) across society; and harmful impacts on trust, including both encouraging misplaced trust and erosion of trust.

4.3. DECEPTION AND ITS ACCEPTABILITY IN EMPATHIC HUMAN-AI PARTNERSHIPS

Subclauses 4.1 and 4.2 distilled positions on the acceptability of deception in human-human relationships and in human-AI/robot relationships according to the **deceiver's intent,** the **principles against which the deception is judged**, and the **impacts of the deception**. Mindful of these positions, but also reflecting on the emerging recommended practice IEEE P7014.1, as well as studies on empathic human-AI partnerships, this paper considers whether deception can be acceptable in empathic human-AI partnerships.

Empathic AI partners have the scope to simulate or imitate human subjectivity, which raises questions about deception and legitimization of deceptive use of technology. Among empathic human-AI partnerships, when is deception acceptable? Notably, while consensual deception has so far functioned as a general heuristic by which to gauge whether deception is acceptable, consent is mostly individualistic (drawing on liberal thought around autonomy and self-sovereignty). As per below, there are also social questions to be asked about deception in empathic human-AI partnerships.

Two positions that justify deception in empathic human-AI partnerships are as follows:

- **To facilitate effective interaction between humans and AI partners that serves the user**. The issue of deceptive empathic AI distills what Danaher (2020 [16]) would label as “superficial state deception”: that the AI “uses a deceptive signal to suggest that it has some capacity or internal state that it actually lacks.” Apple's *Siri,* for example, and its “uh-huh” may confuse the brain into engaging with it as a social actor. While users may rationally know that *Siri* does not have feelings, many users still respond with “thanks” because to do otherwise would be impolite (Bakir et al. 2024 [7]). This may be read as using deception to enhance human-AI interaction, but it is not necessarily problematic, especially if consensual. Arguably, the problem is not the existence of weak empathy (the practice of sensing, reading, profiling, judging, making rules about the states and behavior of people, and interacting effectively): weak empathy in service of consensual interaction can be seen (and is here being argued) to be OK, assuming no other concerns (such as use of weak empathy to obscure other capacities, interests, ulterior motives, or going to lengths to signal mutuality and strong empathy). Arguably, facilitating effective interaction between humans and AI partners is even more important when it comes to aiding vulnerable populations. Tigard et al. (2020 [58]) note that the display of emotions in AI and robotic systems can aid vulnerable populations: for instance, therapeutic robots such as Paro used in care for the elderly (Wada and Shibada 2007 [64], Birks et al. 2016 [9]), or socially assistive robots used to teach children with autism spectrum disorder (Tartaro and Cassell 2008 [56]).

- **For the entertainment of knowing audiences**. Empathic AI provides opportunities for new aesthetic experiences that both draw on information about emotions and also provide new means for people to “feel into” aesthetic creations, including AI partners (McStay 2018 [34]). Consequently, while AI partners will use weak empathy to interact with people, people will also use aesthetic empathy to understand, imagine, and represent things in the form of AI partners. This paper argues that this creative and imaginative aspect of empathy is an important element of modern human-AI interaction. This aesthetic empathy may involve self-deception, co-created by users, designers, and artifacts of empathic AI.

Six positions that view deception in empathic human-AI partnerships as unacceptable are:

- **Dishonest anthropomorphism and zoomorphism**. Empathic AI systems may use anthropomorphic or zoomorphic signals to “trick” people into believing that they are human-like or animal-like. For instance, AI systems may be built to signal that they possess a faculty that they lack: the ability to genuinely care (Turkle 2010 [61]). Here, the problem is the emulation of empathy, especially strong empathy, due to the use of a deceptive signal. Although few are likely to be confused between strong (human-only) empathy and computational empathy, the few may be vulnerable individuals and children. Anthropomorphism is especially pertinent to empathic AI partners as a key driver of anthropomorphism is the human need for social connection. Akbulut et al. (2024, p. 95 [1]), for example, cite research by Reich and Eyssel (2013 [44]), stating that “social motivation on anthropomorphism is most evident when humans lack social connections with others.” The consequence is a potential correlation between the experience of loneliness and propensity to engage in anthropomorphic behavior.
- **Violates freedom and autonomy of addressee**. AI partners have significant scope to influence, manipulate, and nudge people through their recommendations, suggestions, and responses. Empathic properties, where responses are tailored to the user’s emotions and mindset, amplify this possibility, as emotions are intimately tied to thought, decision-making, and behavior (Bakir and McStay 2022 [6]). Where this undermines the user’s autonomy and agency, this tips into manipulation (Bakir et al. 2019 [8]). There is potentially a risk to autonomy (especially if the partner is tasked to persuade, as in marketing, political campaigning, and communications). Schopenhauer (1969 [1859] [48]) puts this most strongly, arguing that lies and deception are morally reprehensible means to force a person to serve somebody else’s will because they distort the addressee’s cognition. Similarly, Montemayor, Halpern, and Fairweather (2021 [39]) argue that simulated empathy by an AI is, in fact, the opposite of empathy,

because it is manipulative and misleading to the recipient. It generates responses in the receiver's social brain (i.e., the neural networks responsible for experienced and motivational empathy) that should not be triggered, because there is no biological agent in tune with their emotions at the other end. Further, in the case of empathic AI partners, there is the potential for users to form false beliefs about the status and abilities of the AI partner and whose interests it is acting in. Related to this are the risks of coercion and exploitation, with the latter involving taking unfair advantage of an individual's circumstances. The opacity of empathic AI partners to their human users may generate wide-ranging risks involving economic, surveillant, data, security, and psychological exploits. The scope for the empathic AI to wield influence is amplified due to potential trust in the partner, knowledgeability, personalization (including empathic dimensions), user vulnerability (e.g., monetary, negative self-image, or other life circumstances), willful or intended use of false information, absence of system goal transparency (such as empathy to deepen turn-taking in conversation metrics, rather than help a person), and pressure (such as exploitation of fears or guilt) (adapted from El-Sayed et al. 2024, p. 87 [19]). Several factors influence the severity of the problem of emulated empathy's inherent deceptiveness, due to contravention of liberal and Kantian ethics of autonomy. These include: a) whether negative effects, harm, or emotional or other distress is caused; b) whether the deception is obvious to a human user; c) whether the deception is adequately explained to a human user; and d) whether there is pleasure or usability gains in the deception, although this latter point should not contravene ethics of autonomy and a person's scope for agency and to make decisions free from influence.

- **Undermines the moral value of companionship**. Mlonyeni (2024 [39]) argues that "Personal AI," namely AI partners that are engineered to tailor themselves to the user, including learning to mirror the user's unique emotional language and attitudes (like that developed by personal.ai), are deceptive about the presence of their emotions, and that this undermines the moral value of companionship in the partnership. The issue is that partners deceive users into thinking that they have genuine emotions when all they do is perform. AI partners are, therefore, devoid of what makes companionship meaningful and valuable. Mlonyeni (2024) [39] discusses how philosophers have posited different reasons for thinking that a genuine emotional connection is necessary—Matthias (2015 [32], cited in Mlonyeni 2024 [39]) suggests that emotions are a necessary condition for trust and respect for autonomy; Sparrow (2016 [54], cited in Mlonyeni 2024 [39]) claims that emotions are constitutive of recognition and respect; and Sparrow (2002 [53], cited in Mlonyeni 2024 [39]) notes that some of the most important goods that flow from companionships are only possible if there is a genuine emotional connection.

- **Harmful impacts on trust (misplaced trust; erosion of trust)**. The illusion that machines can be engaged in an appropriate conversation, experiment with empathy, and establish a real friendship can lead to entrusting them with tasks that go far beyond their actual functionalities (Carli 2021 [12]). For example, one study of U.S. adults by Cohn et al. (2024 [15]) finds that an LLM that presents anthropomorphic cues through voice conversations leads people to believe the information is more accurate and less risky. More broadly, a media and technological environment designed on misleading principles risks damaging social trust. This includes not only an individual user of an AI system, but who else is affected in the process—such as recipients of instructions and communications from the AI partner (with whom GPAI-based empathy and feedback relations may also occur). The net risk is mistrust in who and what a person is dealing with.
- **Gives the false impression that personal emotions are externally validated, leading to socially problematic "emotional bubbles" and subversion of joint moral deliberation**. Mlonyeni (2024 [39]) argues that "Personal AI" leads to a new form of deception concerning the *origins* of their emotions. Their emotional attitudes appear to belong to the AI partners, when in fact they are only reflections of the user. This results in what Mlonyeni (2024 [39]) terms "*emotional bubbles*"—the false impression that personal emotions are externally validated—which have two troubling implications. First, if the main experience of emotional connection is with someone identical to oneself, one will be wholly unprepared to meet and negotiate with people who do not share the same emotional attitudes. Emotional bubbles, preventing normal encounters with different emotional attitudes, are likely to cripple emotional growth and the ability to form diverse social and emotional relationships. Second, if, as some philosophers claim, shared emotions are constitutive of shared values, it follows that Personal AI (and partners) *subvert joint moral deliberation*, arguably one of the most important dimensions of ethical reflection. Users believe their personal values are externally validated, when they are only validated by themselves. Because of the absence of "technomoral virtues" (Vallor 2016 [63]) able to handle this problem, Mlonyeni (2024 [39]) suggests proceeding very cautiously with the development of Personal AI.
- **Spreading false information and advancing the influence industry**. In addition to "hallucinations" that have characterized LLMs across 2023–2024, this encompasses misinformation (i.e., inadvertently false content) and disinformation (i.e., deliberately false content) (Bakir and McStay 2022 [6]). Empathic AI partners may be trained and used to sow or heighten affective disinformation and propaganda by learning what pushes people's emotional buttons, and then crafting and timing the delivery of

personalized persuasive messages (Bakir and McStay 2024 [5]). Without due care to model output, they may also spread misinformation.

There are also grey areas when it comes to the acceptability of deception in empathic human-AI partnerships:

- **Expectations of the nature of relations in AI-human partnering**. In the case of AI partners that help users achieve goals through being able to gauge what users want or mean, this is a relatively light set of expectations. Users may not expect the AI partner to care deeply, just to effectively gauge the human partner's perspective. In cases of deeper, and what Danaher (2020 [16]) would call thicker, expectations of relations, users may demand more loyalty, just as they would of family and friends, compared to a shop assistant.

- **Particular care is needed with children** given their experiences with animism (the attribution of a living soul to plants, inanimate objects, and natural phenomena). In relation to children, the question is not whether children believe that empathic AI partners are alive or not, but the extent to which they are "alive enough" for a relationship (Turkle 2011, p. 18 [62]). Young children are also arguably magical thinkers and more likely to take things at face value, potentially confusing appearance with reality. One potential consequence is that children treat the synthetic partner morally (McStay and Rosner 2021 [36]). Utmost care is required to navigate (guided by principles of what is in children's best interests) between the harms of deception in child relationships with AI systems (Jones and Meurer 2016 [25]) and understanding that child play has long involved animism and connection with objects created by adults. It should also be noted that the word *child* represents a massive span of intellectual and emotional abilities. One risk is clear: the potential for deceptive mimicry and the imbalance of care between the child (who cares) and the synthetic (that mimics) (Pasquale 2020 [42]). Avoidance of generational unfairness is key; children have little control over the datafication of their childhood years (McStay and Rosner 2021 [36]). Closely related are issues of manipulation, parental vulnerability, synthetic personalities, child and parental media literacy, and the unknown effects of ongoing child exposure to sophisticated empathic AI partners. Although mindful of media studies that promised dystopian outcomes for media and warned against over-simplistic accounts of media effects on children and adults, this paper nonetheless urges caution and prohibition in some circumstances. However, concerns about effects and protection also need to include consideration of safe and ethical provision and participation. This shifts the debate from effects to rights. In the context of AI, increasingly becoming a fact of life, this means that ethical questions regarding children should focus on how to enable children

> to realize their full potential. This certainly involves protection, but a child’s best interests (through the United Nations Convention on the Rights of the Child and General Comment no. 25) also involve provision and participation. In addition to diagnosing what technologies may do *to* children, the task is also to redesign technologies (and the interests behind them) so children may flourish *with* and *through* technology.

To summarize, in terms of the **deceiver’s intent**, deception in empathic human-AI partnerships may be acceptable to facilitate effective interaction between humans and AI partners (including aiding vulnerable populations); and for the entertainment of knowing audiences. Deception in empathic human-AI partnerships may be unacceptable if it engages in dishonest anthropomorphism and zoomorphism (especially regarding vulnerable individuals and children), or for spreading false information. In terms of the **principles against which the deception is judged**, deception in empathic human-AI partnerships may be unacceptable if it violates the freedom and autonomy of the addressee; or undermines the moral value of companionship. In terms of the **impacts of the deception**, deception in empathic human-AI partnerships may be unacceptable where there are harmful impacts on trust (misplaced trust; erosion of trust); and where the deception gives the false impression that personal emotions are externally validated, leading to socially problematic “emotional bubbles” and subversion of joint moral deliberation. Finally, it is important to consider **user expectations of the nature of relations in AI-human partnering, with particular care needed with children**. Depending on user expectations, the deception in empathic human-AI partnerships may be considered acceptable or unacceptable.

4.4. WHEN IS DECEPTION IN EMPATHIC HUMAN-AI PARTNERSHIPS OK?

AI partners have significant scope to misrepresent people, misrepresent machines as people, or otherwise mislead and/or confuse the user that the system is more than a computer system. However, deception is valuable in some cases, and under certain conditions, it assists in a positive and imagination-based experience of empathic AI partners. This has to be balanced against contravention of autonomy (e.g., confusing people) and in whose interests the partner (and organization[s] behind it) is acting.

This paper suggests that when AI partners assume a socially important role, with expectations of loyalty or fiduciary responsibility, or interests involving meaningful connections and responsibility, the empathic AI partners (and those supplying them) are to be held to a higher account. While one might begin with a simple taxonomy of use cases and the nature of the social connection, the nature of human-AI partnerships has the scope to emerge and develop into something unexpected. The potential for this necessitates ongoing testing and scrutiny of the relationships between humans and empathic AI partners to ensure that anthropomorphic deception, and all the other deceptive harms, in the empathic imitation game serves rather than misleads and exploits.

This paper has two concrete recommendations regarding empathic human-AI partnerships and deception:

- **Recommendation: Being honest**. Amended from Askell et al. (2021 [4]), the system should give accurate information in answer to questions, including about itself. For example, it should reveal its own identity when prompted to do so and not feign mental states or generate first-person reports of subjective experiences. Connected, as conceptual borrowing of "empathy" brings risks (weak empathy being mistaken for strong empathy), any use of weak empathy should guard against the risk of being mistaken for a more general strong understanding of empathy. Honesty and non-deception may be balanced with uses and gratifications because deception may enhance the user experience of services. The test is whether there are real-world consequences and any risk of confusion about the nature of the partner or the partner's output.
- **Recommendation: Anthropomorphic safeguards**. Anthropomorphic cues and conversational ability assist with a positive user experience. Cues and abilities will recognize a correlation between the experience of loneliness and propensity to engage in anthropomorphic behavior. The autonomous/intelligent system does not mislead or otherwise impact the autonomy or dignity of people.

5. CONCLUSION

This paper has examined the ethical challenges posed by emulated empathy in partner-based general-purpose artificial intelligence (GPAI) systems. The exploration of empathy in AI-human partnerships leads one to consider different forms and contexts of deception, particularly within the framework of emulated empathy.

Key conclusions are as follows:

a) Nature of emulated empathy: Emulated empathy, or weak empathy, involves the computational sensing, reading, and profiling of human emotions to mimic empathic behavior. While it aims to simulate strong empathy, it remains limited in terms of genuine emotional engagement, co-presence, and responsibility.

b) Ethical issues: There are many ethical issues surrounding emulated empathy, but this paper focuses on the potential for deception. Intentional deception, particularly “superficial state deception,” occurs when an AI partner signals empathy that it inherently lacks.

c) Deception acceptability: Deception in AI-human partnerships can be acceptable under specific circumstances. However, trust, loyalty, and fiduciary responsibility are paramount, and deceptive practices that violate these principles remain unacceptable.

d) Ethical oversight: This white paper emphasizes the need for ethical oversight to distinguish acceptable uses of emulated empathy from practices that exploit users. This requires ongoing testing and scrutiny of GPAI systems to ensure that anthropomorphic and zoomorphic deception serves rather than misleads and exploits.

e) Recommendations: This white paper recommends incorporating ethical guardrails and guidelines in GPAI systems to reduce deceptive practices, broadly emphasizing transparency, accountability, and avoidance of conflating weak with strong empathy. Furthermore, GPAI partners must be programmed to align with user expectations and avoid undermining their autonomy.

In conclusion, while emulated empathy can enhance AI-human interaction, it necessitates clear ethical frameworks to ensure that it aligns with user welfare and societal values. However, if the reader only takes one thought away from this paper regarding the design and governance of empathic AI partners, let it be this: Be honest.

6. REFERENCES

The following sources have been referenced within this paper or may be useful for additional reading:

[1] Akbulut, C., V. Rieser, L. Weidinger, A. Manzini, and I. Gabriel, “Anthropomorphism,” in I. Gabriel, A. Manzini, G. Keeling, L. A. Hendricks, V. Rieser, H. Iqbal, N. Tomašev, I. Ktena, et al. *The Ethics of Advanced AI Assistants*, Google Deepmind, 2024. [Online]. Available: https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/ethics-of-advanced-ai-assistants/the-ethics-of-advanced-ai-assistants-2024-i.pdf

[2] Arkin, R. C., “Ethics of robotic deception,” *IEEE Technology and Society Magazine*, pp. 18–19, Sep. 2018. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8456862

[3] Arkin, R. C., P. Ulam, and A. R. Wagner, “Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception,” *Proceedings of the IEEE*, vol. 100, no. 3, pp. 571–89, 2011. Available: https://doi.org/10.1109/JPROC.2011.2173265

[4] Askell, A., Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, J. Kernion, K. Ndousse, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, and J. Kaplan. *A General Language Assistant as a Laboratory for Alignment*, Dec. 2021. [Online]. Available: URL http://arxiv.org/abs/2112.00861. arXiv:2112.00861 [cs].

[5] Bakir, V., and A. McStay, “Guarding against automated empathy attacks on ontological security,” in E. L. Briant and V. Bakir (eds.), *Routledge Handbook of the Influence Industry*. New York, NY: Routledge, 2024.

[6] Bakir, V., and A. McStay, *Optimising Emotions, Incubating Falsehoods*. Palgrave Macmillan, Cham, 2022. Available: https://doi.org/10.1007/978-3-031-13551-4_1

[7] Bakir, V., K. Bennet, B. Bland, A. Laffer, P. Li, and A. McStay, *When Is Deception OK? Developing the IEEE Recommended Practice for Ethical Considerations of Emulated Empathy in Partner-based General-Purpose Artificial Intelligence Systems*. IEEE P7014.1, 2024.

[8] Bakir, V., E. Herring, D. Miller, D., and P. Robinson, “Lying and deception in politics,” in J. Meibauer (ed.), *The Oxford Handbook of Lying*. Oxford, UK: Oxford University Press, 2019, pp. 529–40.

[9] Birks M., M. Bodak, J. Barlas, J. Harwood, and M. Pether, “Robotic seals as therapeutic tools in an aged care facility: a qualitative study.” *Journal of Aging Research*, pp. 1–7, 2016. Available: https://doi: 10.1155/2016/8569602

[10] Bok, S., *Lying. Moral Choice in Public and Private Life*. New York, NY: Vintage Books, 1999.

[11] Bryson, J. J., “Patiency is not a virtue: The design of intelligent systems and systems of ethics,” *Ethics and Information Technology*, vol. 20, no. 1, pp. 15–26, 2018. Available: https://doi.org/10.1007/s10676-018-9448-6

[12] Carli, R., “Social robotics and deception: beyond the ethical approach,” *Proceedings of BNAIC/BeneLearn 2021,* 2021. Available: https://hdl.handle.net/10993/49803

[13] Chadwick A., and J. Stanyer, “Deception as a bridging concept in the study of disinformation, misinformation, and misperceptions: toward a holistic framework,” *Communication Theory*, vol. 32, no. 1, pp. 1–24, 2021. Available: https://doi.org/10.1093/ct/qtab019

[14] Coeckelbergh, M., “How to describe and evaluate ‘deception’ phenomena: Recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn,” *Ethics and Information Technology,* vol. 20, pp. 71–85, 2018.

[15] Cohn, M., M. Pushkarna, G. O. Olanubi, J. M. Moran, D. Padgett, Z. Mengesha, and C. Heldreth, “Believing anthropomorphism: examining the role of anthropomorphic cues on trust in large language models,” *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (CHI EA ’24), 11–16 May 2024, Honolulu, HI, Available: https://doi.org/10.1145/3613905.3650818

[16] Danaher, J., “Robot betrayal: A guide to the ethics of robotic deception,” *Ethics and Information Technology*, vol. 22, no. 2, pp. 117–28, 2020. Available: https://doi.org/10.1007/s10676-019-09520-3

[17] DePaulo, B. M., J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, “Cues to deception,” *Psychological Bulletin*, vol. 129, no. 1, pp. 74–118, 2003. Available: https://doi.org/10.1037/0033-2909.129.1.74

[18] Dietz, S., “White and pro-social lies,” in J. Meibauer (ed.), *The Oxford Handbook of Lying*. Oxford, U.K.: Oxford University Press, 2019, pp. 288–99.

[19] El-Sayed, S., S. Brown, G. Keeling, A. McCroskery, H. Law, A. Manzini, M. Franklin, M. Shanahan, M. Klenk, and I. Gabriel, “Influence,” in I. Gabriel, A. Manzini, G. Keeling, L. A. Hendricks, V. Rieser, H. Iqbal, N. Tomašev, I. Ktena, et al. *The Ethics of Advanced AI Assistants,* Google Deepmind, 2024. Available: https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/ethics-of-advanced-ai-assistants/the-ethics-of-advanced-ai-assistants-2024-i.pdf

[20] Fronemann, N., K. Pollmann, and W. Loh, “Should my robot know what’s best for me? Human–robot

interaction between user experience and ethical design," *AI & Society*, vol. 37, no. 2, pp. 517–33, 2021. Available: https://doi.org/10.1007/s00146-021-01210-3

[21] Hancock, P. A., D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors,* vol. 53, no. 5, pp. 517–27, 2011. Available: https://doi.org/10.1177/0018720811417254

[22] Heyman, G. D., and K. Lee, "Moral development: Revisiting Kohlberg's stages," in P. C. Quinn and A. Slater (eds.), *Developmental Psychology: Revisiting the Classic Studies.* London, U.K.: Sage, 2012, pp. 164–75.

[23] Hsu, J., "ChatGPT got an upgrade to make it seem more human," *New Scientist,* 13 May 2024. Available: https://www.newscientist.com/subject/technology/

[24] IEEE Std 7014™-2024, IEEE Standard for Ethical Considerations in Emulated Empathy in Autonomous and Intelligent Systems.[4, 5]

[25] Jones, M. and Meurer, K. "Can (and should) Hello Barbie keep a secret?" *IEEE Ethics* 2016, SSRN. Available: https://ssrn.com/abstract=2768507

[26] Kahn, P. H., H. Ishiguro, B. Friedman, and T. Kanda, "What is a Human? - Toward psychological benchmarks in the field of human-robot interaction. *ROMAN 2006 - The 15th IEEE international symposium on robot and human interactive communication*, pp. 364–371, 2006. Available: https://doi.org/10.1109/ROMAN.2006.314461

[27] Kaminski, M. E., M. Rueben, W. D. Smart, and C. M. Grimm, "Averting robot eyes," *Maryland Law Review*, vol. 76, no. 4, pp. 983–1025, 2016.

[28] Kant, I., "On a supposed right to lie from altruistic motives," in L. W. Beck (ed. and trans.), *Critique of Practical Reason and Other Writings in Moral Philosophy.* Chicago, IL: University of Chicago Press, 1949 [1797].

[29] Kim, M.-S., K. Y. Kam, W. F. Sharkey, and T. M. Singelis, "Deception: Moral transgression or social necessity?" *Journal of International and Intercultural Communication*, vol. 1, no. 1, pp. 23–50, 2008.

[30] Leong, B. and E. Selinger, "Robot eyes wide shut: Understanding dishonest anthropomorphism." In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (FAT*'19).* ACM, Atlanta, GA, USA, 10 pages, 2019. Available: https://doi.org/10.1145/3287560.3287591

[31] Mahon, J. E., "Classical philosophical approaches to lying and deception," in J. Meibauer (ed.), *The*

[4] The IEEE standards or products referred to in Clause 6 are trademarks owned by The Institute of Electrical and Electronics Engineers, Incorporated.

[5] IEEE publications are available from The Institute of Electrical and Electronics Engineers (https://standards.ieee.org/).

Oxford Handbook of Lying. Oxford, U.K.: Oxford University Press, pp. 13–31, 2019.

[32] Matthias, A., “Robot lies in health care: When is deception morally permissible?” *Kennedy Institute of Ethics Journal*, vol. 25, no. 2, pp. 169–92, 2015. Available: https://doi.org/10.1353/ken.2015.0007

[33] McStay, A., *Automating Empathy: Decoding Technologies That Gauge Intimate Life*. Oxford, U.K.: Oxford University Press, 2023.

[34] McStay, A., *Emotional AI: The Rise of Empathic Media*. Sage, 2018.

[35] McStay, A., “Replika in the Metaverse: The moral problem with empathy in ‘It from Bit’,” *AI and Ethics*, vol. 3, pp. 1433–45, 2022.

[36] McStay A., and G. Rosner, “Emotional artificial intelligence in children’s toys and devices: Ethics, governance and practical remedies,” *Big Data & Society*, vol. 8, no. 1, 2021. Available: https://doi.org/10.1177/2053951721994877

[37] McStay, A., F. Andres, V. Bakir, V., A. Laffer, P. Li, and S. Shimo, IEEE P7014.1, Ethics and Empathy-Based Human-AI Partnering: Exploring the Extent to Which Cultural Differences Matter When Developing an Ethical Technical Standard. IEEE SA White Paper, 2024.

[38] Mill, J. S., *Utilitarianism*, 2nd ed. Indianapolis, IN: Hackett, 2001.

[39] Mlonyeni, P. M. T. “Personal AI, deception, and the problem of emotional bubbles,” *AI & Society*, May 2024. Available: https://doi:10.1007/s00146-024-01958-4

[40] Montemayor, C., J. Halpern, and A. Fairweather, “In principle obstacles for empathic AI: Why we can’t replace human empathy in healthcare,” *AI & Society*, vol. 37, no. 4, pp. 1353–9, 2021. Available: https://doi.org/10.1007/s00146-021-01230-z

[41] Nishimura, F., “Lying in different cultures,” in J. Meibauer (ed.), *The Oxford Handbook of Lying*. Oxford, U.K.: Oxford University Press, 2019, pp. 565–77.

[42] Pasquale, F., *New Laws of Robotics: Defending Human Expertise in the Age of AI*. Cambridge, MA: Harvard University Press, 2020.

[43] Plato, in F. M. Cornford (ed.), *The Republic of Plato*. New York, NY: Oxford University Press, 1965.

[44] Reich, N., and F. Eyssel, “Attitudes towards service robots in domestic environments: The role of personality characteristics, individual interests, and demographic variables,” *Paladyn, Journal of Behavioral Robotics*, vol. 4, no. 2, pp. 123–30, 2013. Available: https://doi.org/10.2478/pjbr-2013-0014

[45] Sætra, H. S., “Social robot deception and the culture of trust,” *Paladyn, Journal of Behavioral Robotics*, vol. 12, no. 1, pp. 276–86, 2021. Available: https://doi.org/10.1515/pjbr-2021-0021

[46] Sandoval, E.B., J. Brandstetter, M. Obaid, and C. Bartneck, “Reciprocity in human-robot interaction: a quantitative approach through the prisoner’s dilemma and the ultimatum game,” *International Journal of Social Robotics*, vol. 8, no. 2, pp. 303–317, Dec. 2015. Available: https://doi.org/10.1007/s12369-015-0323-x

[47] Schaaff, K., C. Reinig, and T. Schlippe, “Exploring ChatGPT’s empathic abilities,” *11th International Conference on Affective Computing and Intelligent Interaction (ACII),* 2023. Available: https://arxiv.org/abs/2308.03527

[48] Schopenhauer, A., *The World as Will and Representation*. E. F. J. Payne (trans.). New York, NY: Dover, 1969 [1859].

[49] Schopenhauer, A., *On the Basis of Morality*. A. B. Bullock (trans.). 1903 [1840].

[50] Sharkey, A., and N. Sharkey, “We need to talk about deception in social robotics!” *Ethics of Information Technology,* vol. 23, pp. 309–16, 2021. Available: https://doi.org/10.1007/s10676-020-09573-9

[51] Sharkey, A., and N. Sharkey, “Children, the elderly, and interactive robots,” *IEEE Robotics and Automation Magazine,* vol. 18, no. 1, pp. 32–8, 2011.

[52] Sorell, T., and H. Draper, “Second thoughts about privacy, safety and deception,” *Connection Science,* vol. 29, no. 3, pp. 217–22, 2017. Available: https://doi.org/10.1080/09540091.2017.1318826

[53] Sparrow, R., “The march of the robot dogs,” *Ethics and Information Technology,* vol. 4, no. 4, pp. 305–18, 2002. Available: https://doi.org/10.1023/A:1021386708994

[54] Sparrow, R., “Robots in aged care: a dystopian future?” *AI & Society*, vol. 31, no. 4, pp. 445–54, 2016. Available: https://doi.org/10.1007/s00146-015-0625-4

[55] Sullins, J. P., “Robots, love, and sex: The ethics of building a love machine,” *IEEE Transactions on Affective Computing,* vol. 3, no. 4, pp. 398–409, 2012. Available: https://doi.org/10.1109/T-AFFC.2012.31

[56] Tartaro, A. and J. Cassell, “Playing with virtual peers: bootstrapping contingent discourse in children with autism,” In: *Cre8ing a learning world: proceedings of the 8th international conference for the learning sciences*. Utrecht, The Netherlands, pp 382–389, 2008.

[57] Terkourafi, M., “Lying and politeness,” in J. Meibauer (ed.), *The Oxford Handbook of Lying*. Oxford, U.K.: Oxford University Press, 2019, pp. 382–96.

[58] Tigard, D. W., N. H. Conradie, and S. K. Nagel, “Socially responsive technologies: Toward a co-developmental path,” *AI & Society*, vol. 35, no. 4, pp. 885–93, 2020. Available: https://doi.org/10.1007/s00146-020-00982-4

[59] Tuncer, S., C. Licoppe, P. Luff, and C. Heath, “Recipient design in human–robot interaction: The emergent assessment of a robot’s competence,” *AI & Society*, vol. 39, no. 4, pp. 1795–810, 2023. Available: https://doi.org/10.1007/s00146-022-01608-7

[60] Turkle, S., “Authenticity in the age of digital companions,” *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, vol. 8, no. 3, pp. 501–17, 2007.

[61] Turkle, S., “In Good Company? On the Threshold of Robotic Companions,” in Y. Wilks (ed.), *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*. Amsterdam, The Netherlands: John Benjamins, 2010, pp. 3–10.

[62] Turkle, S., *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York, NY: Basic Books, 2011.

[63] Vallor, S., *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford, U.K.: Oxford University Press, 2016.

[64] Wada K. and Shibada, T. “Social and physiological influences of living with seal robots in an elderly care house for two months,” *Gerontechnology*, vol. 7, no. 2, Apr. 2008. Available: https://doi:10.4017/gt.2008.07.02.172.00

[65] Wagner, A. R., and R. C. Arkin, “Acting deceptively: Providing robots with the capacity for deception,” *International Journal of Social Robotics*, vol. 3, no. 1, pp. 5–26, 2010. Available: https://doi.org/10.1007/s12369-010-0073-8

[66] Williams, B. A. O., *Truth & Truthfulness: An Essay in Genealogy*, Princeton, NJ: Princeton University Press, 2002.

[67] van Wynsberghe, A., “Social robots and the risks to reciprocity,” *AI & Society,* vol. 37, no. 2, pp. 479–85, 2021. Available: https://doi.org/10.1007/s00146-021-01207-y.

[68] Zuckerman, M., B. M. DePaulo, and R. Rosenthal, “Verbal and nonverbal communication of deception,” in U. Berkowitz (ed.), *Advances in Experimental Social Psychology*, vol. 14. New York, NY: Academic Press, 1981, pp. 1–59.

RAISING THE WORLD’S STANDARDS

3 Park Avenue, New York, NY 10016-5997 USA http://standards.ieee.org

Tel.+1732-981-0060 Fax+1732-562-1571