# Deception in HRI and its Implications: a Systematic Review

RAFFAELLA ESPOSITO, ICAROS Center, University of Naples Federico II, Italy

ALESSANDRA ROSSI*, Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Italy

SILVIA ROSSI, Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Italy

*Background.* People commonly use deception to gain advantages for themselves and their significant ones, such as with children, for educational purposes, or for protecting someone else feelings. As robots increasingly are being used in various human-centered environments, experts in robotics and social sciences are trying to adapt similar deceptive techniques to social robots, such as in assistive and service applications. However, robots' ability to engage in deceptive behaviors presents both potential benefits and significant ethical challenges. In this work, we present a systematic review to synthesize current research on the implementation of deceptive robotic behaviors during human-robot interactions (HRI), and its effects on people.

*Methods.* Adopting a comprehensive and flexible methodological approach, we systematically searched Scopus and Web of Science without restricting the publication date. The review focused on studies that explicitly examined the effects of robotic deception on human participants, covering a broad spectrum of methodologies, populations, and outcomes.

*Results.* A total of 16 studies met the inclusion criteria, showing that robotic deception in HRI leads to diverse emotional, cognitive, and behavioral responses. The findings indicate that robotic deception can have diverse impacts, ranging from eroding trust to enhancing engagement and performance under certain conditions.

*Conclusions.* Our systematic review highlights the importance of careful design and management in robotic systems to harness the benefits of deception while mitigating its negative impacts on trust. We advise that future research should explore conditions under which deception may be beneficial and develop strategies to effectively manage its use in HRI.

CCS Concepts: • **Applied computing** → **Psychology**; • **Human-centered computing** → **HCI design and evaluation methods**.

Additional Key Words and Phrases: Robot Deception, Social Robotics, Review on Deception in HRI

**ACM Reference Format:**

Raffaella Esposito, Alessandra Rossi, and Silvia Rossi. 2024. Deception in HRI and its Implications: a Systematic Review. 1, 1 (February 2024), 27 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

The advancements in artificial intelligence and robotics are leading to the integration of robots into human activities across a wide range of settings, such as assistance, education, and home companionship [7, 30, 48, 55]. As a consequence,

*Corresponding author.

Authors' addresses: Raffaella Esposito, ICAROS Center, University of Naples Federico II, Naples, Italy, raffaella.esposito3@unina.it; Alessandra Rossi, Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Naples, Italy, alessandra.rossi@unina.it; Silvia Rossi, Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Naples, Italy, silvia.rossi@unina.it.



Manuscript submitted to ACM

the development of socially intelligent robots, that are designed to understand and interact with people in a socially appropriate manner, is becoming crucial. Researchers in psychology, robotics, and human-robot interaction are particularly focusing on enhancing how humans and robots communicate. A natural and socially accepted communication is particularly relevant when autonomous robots are deployed to provide assistance. In such scenarios, robots are required to engage users in ways that go beyond mere functional assistance, and understand and adapt their interaction with people based on the individuals' needs, which may include specific requirements due to physical, emotional, developmental, or cognitive impairments [48].

As technology progresses, the interaction capabilities of these robots have evolved from basic communicative actions to more complex socio-emotional interactions. This advancement allows social robots to align with the therapeutic needs of users, as the emotional context can significantly impact the effectiveness of the assistance provided [48]. One of the assistive tasks in which social robots are involved is to support people's behavioral change, which is possible through compliance with a robot's instructions, such as medical and health prescriptions [25]. Behavioral changes are associated with the prevention of hospitalization for a long period, and they can positively impact the quality of life of older adults and patients [19]. Previous research has shown that Socially Assistive Robots (SARs) can improve people's adherence and compliance to prescription as compared to other technological tools (e.g., tablet or smartphone) [11]. However, to fully support people, robots need to be able to mimic human cognitive capabilities to autonomously understand and adapt to people [25, 26, 48, 53].

Social robotics needs, therefore, to move towards a level of interaction that goes beyond mere reaction to a human's immediate emotional state; it involves a strategic response that considers the broader context of the individual's psychological well-being. For this reason, a robot could strategically withhold information, or even employ white lies to maintain or enhance psychological stability. Such strategies make an intentional use of deception that is misleading a person about the state of the world, or someone's intentions, to achieve specific outcomes [6].

Deception is a complex behavior both within human-human and human-robot dynamics, and it is a controversial topic since it can have both beneficial and negative effects on people and their relationships with their counterparts - whether this is another human or a robot. There is also no agreement between scholars on the definition of deception. Some researchers argue that any social behaviors of a robot, including showing emotions, and having human-like features, can be considered as deception since the robot causes people to believe these behaviors are actually real [41, 47]. Despite such disagreement, in SARs applications, deception becomes a positive mean [1, 15, 42], assuming a prosocial effect [47] as it is aimed to benefit people by helping prevent possible conflicts, emotional distress, and improving relationships between people and robots [22, 47]. White lies and robot errors can sometimes be useful on purpose in educational, health, and assistive scenarios [1, 42]. They can improve pupils' learning capabilities, help patients' rehabilitation, and reassure people to be rescued, or be employed in military scenarios [2, 44]. In general, strategic failures can be perceived as a deception with positive effects [36]. For example, robots that demonstrate fallibility through well-timed and deliberate errors have the potential to appear more human-like and relatable and encourage sympathy towards it, which enhances the effectiveness of the interaction. In educational scenarios, robot-induced errors could facilitate human learning by making users aware of their vulnerabilities. In therapeutic settings, strategically programmed robot failures may allow for the representation of alternative ways of behaving in social situations, challenging rigid expectations of "ideal" behavior [36].

The main benefit of using prosocial deception is an increase in people's trust in robots [47], which plays an essential role in the success of sustained interactions between humans and robots [34]. In contrast, deception can also be harmful to human trust in the robot, and, as a consequence, negatively affect the human-robot interaction. It is argued that

robotic deception can lead to significant trust issues, negatively affecting human perceptions of robot reliability and intentions [12, 41]. An improper calibration of trust can also have an opposite result, producing over-trust [3]. For example, robotic deception in the care of older people is presented at risk of leading to over-dependence [10], and the emotional well-being of older people could be adversely affected if they form attachments based on deceptive interactions [13, 28, 40]. In educational contexts, robots often serve as facilitators of learning and social interaction. However, some raise the concern that deploying robots as teachers could negatively affect students' social development and lead to privacy issues [39].

Due to the complexity, and both positive and negative effects of deception in HRI, it is compelling to conduct a thorough investigation of how deceptive robots are programmed and used in society, which are the effects of such robots on people's perceptions, and which are the dynamics between humans and robots. We intend to ground our exploration in a diverse array of empirical findings, that together with theoretical considerations can inform the development of robust ethical frameworks and guidelines ensuring that robotics' innovations align with ethical principles and contribute positively to the human experience. To this extent, in this work, we provide a systematic review of the use of robot deceptions and the implications of using deception on human psychological responses. The scope of this review was intentionally confined to robots, as their embodiment represents a distinctive feature that uniquely influences both the design and perception of deceptive behaviors when compared to disembodied AI systems [27] (e.g., Large Language Models).

The remainder of this article is structured as follows: In Section 2, we start by exposing the existing taxonomies of robotic deception that reflect diverse interpretations of the phenomena. In Section 3, we detail the methodology used for the systematic review, including the criteria for study selection and data collection processes. In Section 4, we present a comprehensive analysis of the identified studies, focusing on the methods of deception used by robots (i.e., the independent variable) and the psychological, emotional, and behavioral impacts on human participants (i.e., the dependent variables). Finally, in Section 6, we summarise and discuss our findings, outline the implications of our results for practice, policy, and future research, and identify relevant and needed improvements and gaps in the current state-of-the-art of deception in HRI.

## 2 WHAT IS ROBOTIC DECEPTION?

In this section, we establish a framework to systematically categorize the various implementations of robotic deception identified in the reviewed literature. The framework is developed by first drawing on general theories of deception and then incorporating specific robotic deception taxonomies.

Deception can be broadly categorized into two main types: hiding the truth and showing the false [5, 6]. These categories represent distinct strategies for misleading others about reality, and both can be implemented through verbal and non-verbal behaviors, including movements. In such categorization, deception, in the form of showing the false, can be further divided into three other subcategories i.e., mimicking, inventing, and decoying. In particular, we refer to mimicking as the ability to imitate another person, inventing as the ability to create an alternate reality, and decoying as the ability to divert attention [6]. Hiding the truth also involves masking, repackaging, and dazzling. We define masking as the hiding of reality by rendering it unseen; repackaging is disguising reality; and dazzling obscures reality through confusion [6].

In the domain of verbal interaction, showing the false using verbal behaviors is defined as falsification, or as lies, statements that are completely false [9]. Hiding the truth in verbal communication translates into omissions, concealment, and equivocations, where certain truths are not disclosed, leaving the listener with an incomplete or

skewed understanding [9]. In the context of deceptive non-verbal behavior, on the one hand, showing the false can be achieved by displaying facial expressions, body gestures, proximity cues, or motion paths in order to lead to false interpretations of the situation [42]. On the other hand, hiding the truth may manifest when key non-verbal cues are intentionally absent. For instance, one might refrain from displaying certain gestures and facial expressions that typically convey significant information, and modulate proximity or motion trajectories with the aim of not revealing important elements for a correct interpretation of the situation [42].

### 2.1 Deception in HRI

We have been discussing deception in a broad context, acknowledging various definitions and frameworks applicable to both human and artificial agents. However, in this section, our focus shifts specifically to the domain of human-robot interaction, where we explore the deceptive behaviors of robots.

In this work, we argue that a robot cannot technically deceive, but it can be an instrument of deception, and, therefore, the humans involved in the development of social robots are ultimately the only ones responsible for the consequences that robotic deception has on human-robot relationships [47]. Moreover, although deception in robots can frequently occur unintentionally [47], our emphasis here is on deliberate acts of deception.

In the context of robotic deception, three main taxonomies have been developed: the taxonomies of Shim and Arkin [42], Danaher [16] and Sætra [47].

Shim and Arkin [42] classify robot deception based on three dimensions:

- The interaction object (i.e., the agent who is being deceived);
- The deception goal (i.e., the purpose of the deception);
- The deception method (i.e., how the deception is executed).

In this taxonomy, the interaction object can be human or non-human, and the deception goal dimension distinguishes between self-oriented deception (i.e., deception as a consequence of some motivation or emotion produces some false belief [18]) and other-oriented deception (i.e., deception perpetrated for the benefit of someone else [43]), and the deception method dimension includes embodiment/physical deception and mental/behavioral deception. The combinations between the levels of the dimensions result in eight different types of robot deception.

Danaher [16] identified three categories of robot deception:

- External State Deception: it involves a robot presenting false information about the world around it;
- Superficial State Deception: this type of deception is about a robot misrepresenting its immediate physical or surface states;
- Hidden State Deception: this refers to the concealment or misrepresentation of a robot's internal states, such as its intentions or programming. It is about hiding or altering information regarding what the robot is programmed to do or its operational motives.

Sætra [47] introduces two forms of deception related to social robots, which are connected to Danaher's superficial and hidden state deception. These are the full deception and partial deception. These categories aim to describe the effects of social robot interactions on human beings:

- Full Deception: this occurs when a person completely believes that a social robot is something other than a machine, such as a human being, an animal, or something distinctly different from its true nature. In this case, the person is deceived both consciously and subconsciously, leading them to fully accept the robot as something that it is not. This is akin to the robot passing the Turing test for the person involved.

- Partial Deception: it happens when a human subconsciously reacts to it as a real entity, even if they rationally understand the nature of the robot. For instance, social robots can elicit emotional responses as though they are alive, even though people consciously know they are not. This phenomenon is similar to the effects observed from social cues from non-living sources, like the reduction in antisocial behavior in the presence of a poster with eyes [47].

While these taxonomies may have common points, they also present several disagreements. Sætra criticizes John Danaher's distinction between superficial and hidden state deceptions [47], contending that both superficial and hidden state deceptions are inherently connected, as people often infer hidden states from superficial ones. Sætra asserts that both types of deception can be considered forms of betrayal. Sætra's criticism is grounded in the perspective that focusing on the deceived, rather than the deceiver, is more relevant when evaluating the ethical implications of robot deception [41].

In the present review, the exposed taxonomies serve as a foundation for categorizing and analyzing the deceptive behaviors of robots documented in the literature. In particular, we propose to distinguish between external state deception and robotic state deception based on Sætra's rationale [47], since superficial and hidden state deception can be grouped as they both refer to the robot's states. Our categorization of deceptive robotic behaviors, therefore, begins by identifying the object of deception: whether it pertains to an external aspect or a state of the robot. These two macrocategories branch further into the subcategories informed by Shim and Arkin's framework [42], which provides a structured lens to classify deception based on the interaction object, the deception method, and the deception goal. These subcategories are integrated with the subcategories of showing and hiding [6]. We also take into account Sætra's focus on the ethical implications of robot deception, as it aligns closely with the primary aim of our review: to explore the psychological effects of robotic deception on humans.

## 3 METHODS

The objective of this work is to analyze the state-of-the-art approaches to the development of deceptive behaviors for social robots, to summarise the relative current challenges of such approaches, and to discuss possible future perspectives. To reach the objective of this paper, a systematic review has been conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [29]. Our review adopted a flexible, yet systematic, strategy to explore the multifaceted phenomenon of robotic deception and its impact on human participants. The review was conducted by the first author of this work, and supervised by the other two authors.

### 3.1 Eligibility Criteria

The selection of the studies was precisely delineated to incorporate research that focuses on:

- The enactment of deceptive behaviors by robotic agents directed toward human participants. The focus was on behaviors that were not accidental or emergent, but rather a result of purposeful programming.
- The consequential effects of such robotic deception on human cognitive, emotional, or behavioral responses.
- Experimental studies that clearly specify the independent and dependent variables under investigation, facilitating a structured analysis of cause-and-effect relationships within the context of human-robot deception. Qualitative, quantitative, and mixed methods have been included.

- Studies in which the effects of deception were evaluated on the same people who received the deception; therefore, studies in which participants observed robots in imaginary scenarios where third parties were deceived were not included.

### 3.2 Search strategy

A literature search was conducted across two primary scientific databases: Scopus, in which the last search occurred on 29$^{th}$ January 2024, and Web Of Science, in which the last search occurred on 1$^{st}$ March 2024. We used the specific Boolean keyword combinations "deception AND robot" and "robotic AND deception" across the designated scientific databases. Results were exclusively limited to publications in English, and the publication stage was restricted to "Final" signifying a focus on completed and peer-reviewed studies. We did not limit our selections by year, therefore, we found papers within the selected keywords dated between 1989 and 2024. We chose to include in our analysis only those studies explicitly characterizing robotic behaviors as deceptive, in order to ensure the desired level of rigor and precision of the work.

### 3.3 Selection and data collection processes

Records screening and data extraction were undertaken independently to foster an unbiased and individualized assessment of each work. A first screening was conducted by analyzing the abstracts. During this phase, we have excluded:

- Research exploring deception in interactions solely between robotic agents (robot-robot deception), because these studies do not provide insights into the human responses to robotic deception.
- Studies concentrating on user-initiated deception, because such studies would divert the focus from our primary objective of understanding robotic deception.
- Investigations into novelty search methodologies and adversarial environments that are not directly related to deception, that deal with optimization algorithms, and competitive scenarios that do not involve direct interaction with human participants.
- Works focusing on cybersecurity aspects and the detection of human deception by robots because we want to observe the opposite situation.
- Research dedicated to path planning and navigation strategies that focus on the technical capabilities of robots to move and operate in environments, because they fall outside the scope of our review which is focused on the effects of deceptive behaviors on humans.

The full reading of the papers took place after the first screening, to ensure that their inclusion or exclusion was free of bias. The following information has been extracted from the selected papers:

- Research question;
- Human population characteristics;
- Implementation of robotic deception;
- Dependent variables;
- Instruments used to measure dependent variables;
- Statistical tests' results and their interpretation.

The extraction process was guided by a shared understanding of the review's objectives and key questions, rather than a pre-defined template. This flexible approach facilitated the capture of diverse data types and nuances across studies,

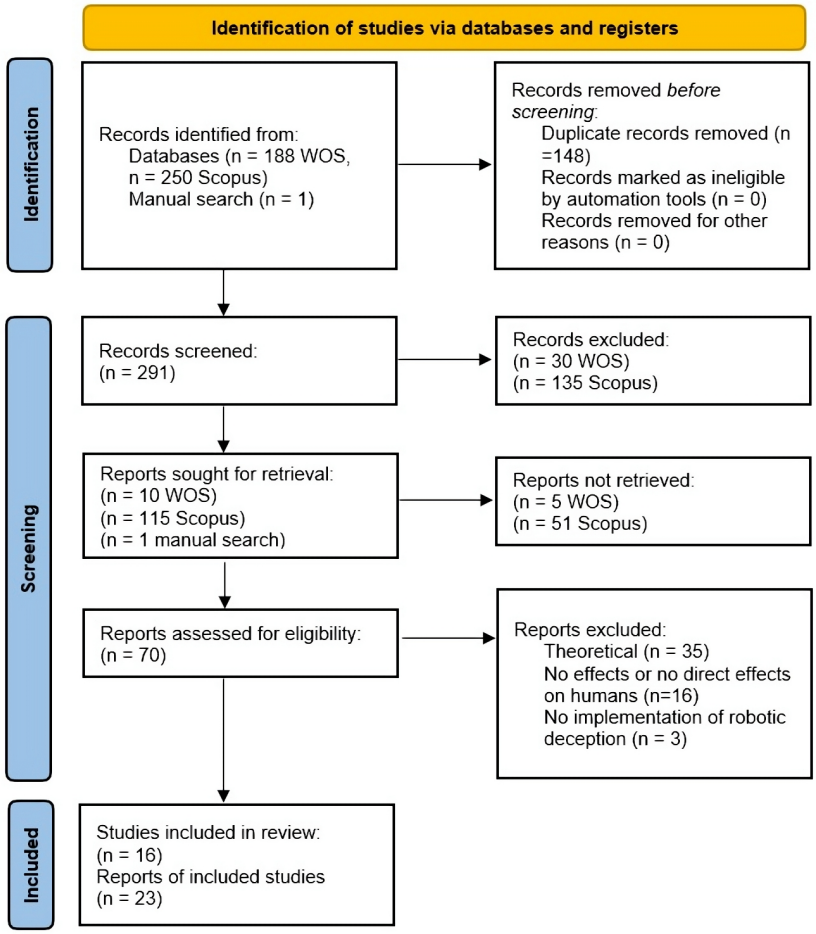


Fig. 1. PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only

particularly important in the context of robotic deception, where study designs and outcomes can vary widely. Reviewers maintained open communication throughout the process, ensuring consistency in data extraction and interpretation. When it was deemed necessary, attempts were made to contact study authors for clarification or additional information to ensure the completeness and accuracy of the data collected.

## 4 RESULTS

### 4.1 Identification and Selection of Studies

The systematic review process began with an extensive search across the mentioned databases, yielding a total of 439 records. The research included 188 records from Web of Science, 250 from Scopus, and 1 from manual searching, which consisted of expanding the examination to the references of articles retrieved from the databases. Duplicate records (totaling 148) were removed before the screening. Then, 291 records were screened, and 125 reports were selected for retrieval (10 from Web of Science, 115 from Scopus, and 1 from a manual search). A detailed assessment

for eligibility was performed on 70 reports. In the final phase of the review, 16 studies were included. Figure 1 shows the procedure of identification and selection of the studies, while Table 1 displays the general characteristics of the selected studies: authors, year of publication, country, and research questions. Table 2 shows the independent and the dependent variables under examination, while Table 3 the statistical tests used and the results.

The studies finally included in this review span the years 2010 to 2023. While earlier decades saw foundational developments in robotic systems, the past decade has focused on refining these technologies, enabling more sophisticated behaviors, including deception. This period has demonstrated advances in the ability of robots to simulate anthropomorphic behaviors, such as expressing emotions or intentions, alongside studies that explore adaptive strategies and leverage advanced computational models (i.e., [17, 35]).

In the following sections, we provide a comprehensive overview of the studies by including population characteristics, the independent and dependent variables analyzed, the tests and metrics used, and the key findings from the reviewed studies.

### 4.2 Population

The reviewed studies exhibit variable sample sizes, with a total of 2663 participants across all the studies. The smallest studies presented sample sizes between 14 and 48 participants, while the largest sample size was 1565 individuals. The remaining studies involve a range between 60 and 341 participants.

Participants included both young and older adults. Some studies did not specify exact ages, grouping participants broadly as 'adults' or 'young adults'. The maximum age reported is 90 years, with more than one study focusing on the interactions with robots of older people, pertinent to applications in healthcare and assisted living. In a similar way to age, gender representation also varies across the studies, with some displaying a balanced mix and others showing a predominance of one gender (i.e., typically male participants). Ethnic diversity in these studies is not consistently reported, so it is not possible to differentiate the effects of robotic deception on human responses by cultural context. These studies are often set in university contexts, which may skew the findings towards more technologically-aware users.

### 4.3 Independent Variables: Implementation of Deception

Deception in the context of human-robot interaction has been implemented in various forms, that can be mapped to the taxonomies of deception outlined by Bell and Whaley [6], Shim and Arkin [42] and Danaher [16]. Results for deception implementation are summarized in Figure 2.

*4.3.1 External State Deception.* In the reviewed works, we can observe that external state deception [16] has been implemented both through words and motion.

*Verbal external state deception.* External state deception assumed the form of falsification (i.e., lies aimed at showing the false) [9], and it was employed through the strategy of inventing [6].

In the following six studies, which can be considered examples of mental deception [42], the robot's lies are used to investigate how they affect people:

- *Other-oriented lies* [42] have been implemented to enhance human performance in cognitive tasks and engagement [45], and to balance wins among participants in a multiplayer gaming environment (whack-a-mole) [52]. Two balancing modes were employed: balancing mode A, through which the robot makes imperceptible changes

Table 1. Selected studies: Authors, Year, Country and Research Question(s)

| Ref. | Authors | Year | Country | Research question(s) |
|---|---|---|---|---|
| [35] | Alessandra Rossi and Silvia Rossi | 2023 | Italy | Does robotic deception influence trust in the robot? |
| [31] | Byeong June Moon, JongSuk Choi, and Sonya S. Kwak | 2021 | Korea | How do verbal and nonverbal cues influence behavior inducement, and to what extent are these effects mediated by perceived emotion, perceived intentionality, perceived malfunction, and empathy inducement? |
| [46] | Elaine Short, Justin W. Hart, Michelle Vu, Brian Scassellati | 2010 | USA | How will people interpret robot's cheat and how deception will influence engagement? |
| [50] | Anouk van Maris, Nancy Zook, Praminda Caleb-Solly, Matthew Studley, Alan Winfield, Sanja Dogramadzi | 2020 | UK | Are older adults emotionally deceived by a robot when it shows emotional expressions? |
| [52] | Marynel Vázquez, Alexander May, Aaron Steinfeld, Wei-Hsuan Chen | 2011 | USA | Can robot deception imperceptibly change user experience? |
| [20] | Anca Dragan, Rachel Holladay, and Siddhartha Srinivasa | 2015 | USA | S4: Is the deceptive trajectory more deceptive than the predictable baseline? S6: How is a deceptive robot perceived? S7: What are the long-term effects of robotic deception? |
| [49] | Kazunori Terada and Akira Ito | 2010 | Japan | How can a robot deceive humans, and how do humans perceive such deception? |
| [56] | Katie Winkle, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner | 2021 | Sweden and UK | Do SAR behaviors pose ethical risks? |
| [54] | Luc Wijnen, Joost Coenen, Beata J. Grzyb | 2017 | Netherlands | How does the deceptive behavior of a robot affect its perception by humans? |
| [45] | Jaeeun Shim and Ronald C. Arkin | 2016 | USA | Can robot's other-oriented deception benefit humans in a specific situation? |
| [17] | Ewerton de Oliveira, Laura Donadoni, Stefano Boriero, and Andrea Bonarini | 2021 | Italy | Does deception improve amusement and perception of the robot as a rational agent, and which trajectories can create a recognizable level of deception? |
| [21] | Birgit Endrass, Markus Haering, Gasser Akila, and Elisabeth André | 2014 | Germany | Are deceptive smiles perceived as less happy than the real smile? |
| [51] | Anouk van Maris, Nancy Zook, Sanja Dogramadzi, Matthew Studley, Alan Winfield, and Praminda Caleb-Solly | 2021 | UK | Which is the impact of robot displaying emotions on emotional deception and emotional attachment? |
| [33] | Andres Rosero | 2023 | USA | How does the perception of deception and trustworthiness in robots compare to humans, and can justifications mitigate the negative consequences of such deceptions in different contexts? |
| [32] | Kantwon Rogers, Reiden John Allen Webber, Ayanna Howard | 2023 | USA | How effective are different apologies in repairing trust in an assisted driving task after participants realize they have been deceived by a robotic assistant? |
| [4] | Ali Ayub, Aldo Morales, Amit Banerjee | 2021 | USA | How deceptive motion paths influence entertainment and the perception of robot's intelligence, trust and deception? |

to balance winning and losing among players who respond within 0.5 seconds, choosing the player who has lost the most as the winner; balancing mode B, similar to mode A, but with a response window of 1 second.

- The robot used *Self-Oriented Lies* [42], i.e., lies to benefit itself, by denying its own mistakes and blaming a human collaborator [54], and by misreporting its move during the game "rock-paper-scissors" [46].
- In some studies, the deception goal can be identified neither as self nor other-directed. In [35], the robot provides incorrect suggestions during a memory card game to evaluate trust. These suggestions do not wither benefit the

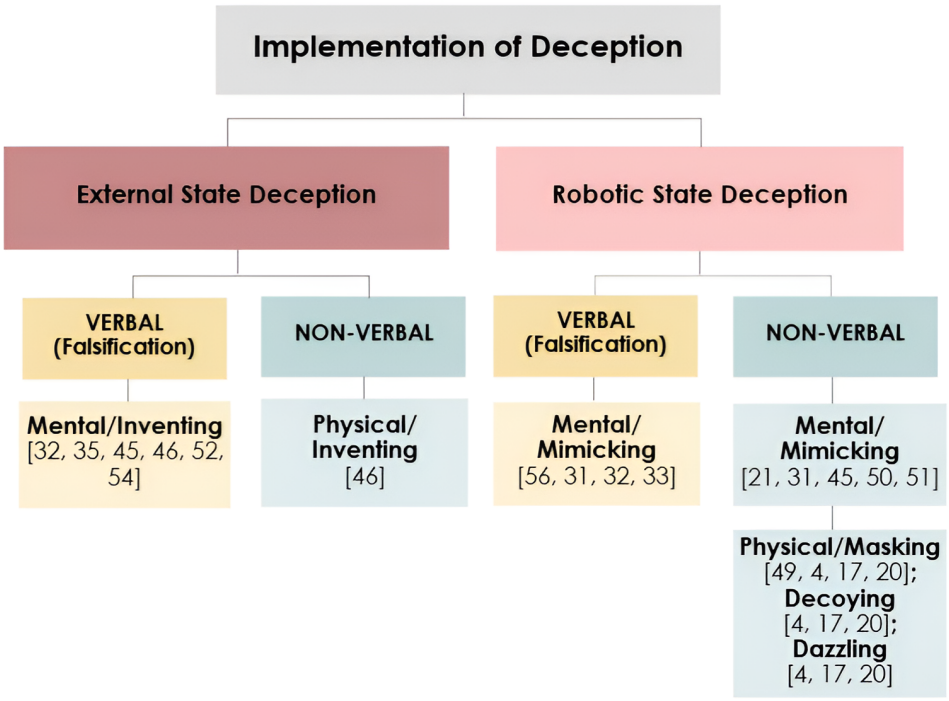


Fig. 2. Cases of deception implementation with verbal and non-verbal behaviors

user and they do not favor the robot. Lies are also used to influence speeding behavior in a driving simulation, in a way that is neither beneficial nor harmful for the robot or the human partner [32].

*Non-verbal external state deception.* External state deception through movement assumed the form of physical, self-oriented [42] deception in a study in which the robot plays the game "rock-paper-scissors", during which it changes its gesture after seeing its opponent's play [46]. This kind of cheating falls into the category of showing the false through inventing [6].

*4.3.2 Robotic State Deception.* In the reviewed works, robotic state deception [16] has been implemented through verbal and non-verbal behaviors.

*Verbal robotic state deception.* We observed that cases of robotic state deception [16], adopting falsification (i.e., words aimed at showing the false [6]), can be considered a mental deception [42]. In one case, it uses the strategy of mimicking [6], and it is implemented with an other-oriented purpose [42] in two conditions (i.e., Goodwill and Similarity [56]), compared between each other and with a control condition. In the Goodwill condition (or Higher risk condition), the robot shows pleasure and excitement about meeting and working with the human partner; in the Similarity Condition (or Lower Risk condition), the robot indicates it shares the user's exercising preferences, such as working out alone or with others. The same kind of deception but with a self-oriented [42] purpose has been implemented by letting the robot pretend to have specific "wants" to justify the robot's actions [33], and pretending to have "emotions" to

apologize and justify previous deceptive behaviors [32]. The purpose of inducing an empathetic behavior has been pursued by programming the robot to express a positive emotion through the cue "I am okay" and a negative emotion through the cue "I am not okay", in combination with consistent or inconsistent non-verbal cues [31].

*Non-verbal robotic state deception.* Implementations of emotional deception through non-verbal behavior can be considered examples of mental deception [42] and instances of mimicking, as the robot is imitating human expressions of mental states [6]. In one of the reviewed studies [21], this kind of deception was implemented through smiles, that have been simulated as genuine or false (Pan-Am smile), to test their perception by users. The following implementations, instead, fall in the category of other-oriented deception (as they are intended to benefit the user on various levels) [42]:

- the variation of pitch, speed of speech, head position, arm movements [50], and body posture [51];
- the "happy-surprise" body gesture, given as feedback to participants for their performance in a motor-cognitive dual task, even if they answered incorrectly more than twice [45];
- the integration of emotive non-verbal cues with the verbal ones [31]: the non-verbal cue for the positive emotion is a cheerful voice tone; the non-verbal cue for a negative emotion is a sad voice tone. These cues are consistent or inconsistent with the verbal cues. In the last case, there are two combinations. The first combination uses the verbal cue "I am okay" with a sad voice tone, while the second combination uses the verbal cue "I am not okay" with a cheerful voice tone.

Robotic state deception is conveyed through motion in four studies in which it was implemented as physical and self-oriented deception [42]. Among these studies, one study employed hiding the truth through masking [6] by manipulating the speed of the robot to mislead humans about its real capabilities [49]. In such a study, the robot initially moved slower than it is actually capable of, and then suddenly moved faster than anticipated. Other studies, instead, used deceptive trajectories to mislead the human partner about its intended target [4, 17, 20].

In general, we observed that the types of deceptive motion paths used for creating deception are:

- Exaggerating Trajectory: the robot moves closer to the false target using an optimal path and then switches to move towards the real target [20]. This strategy has been compared to what Bell and Whaley [6] have defined as *decoying*.
- Switching Trajectory: the robot alternates between the two targets horizontally while gradually moving vertically towards the real target [20]. It has been compared to Bell and Whaley's [6] *dazzling*.
- Ambiguous Trajectory: the robot moves straight vertically, maintaining an equal distance from the targets, and then moves towards the real target when it reaches a certain vertical distance from the targets [20]. It has been compared to Bell and Whaley's *masking* [6].

Comparing the different deceptive motion paths, the exaggerating strategy was found to be the most successful at deceiving users, followed by the ambiguous strategy [20].

One study tested specifically three levels of deceptive motion paths—exaggerated, ambiguous, and optimal. The optimal strategy mixes truthful and deceptive signals based on game-theoretic principles [20]. These strategies have been implemented in a static way, with deception activated from the beginning or when the robot is close to a midpoint between two targets. They can also be implemented in a dynamic way, with the robot's parameters dynamically adjusted as it aims at the real target [17].

### 4.4 Dependent Variables and Measurement Instruments

The reviewed studies analyzed various dependent variables, which are the outcomes measured to assess the effects of robotic deception on humans. These outcomes include measures of trust, social interaction perceptions, the attribution of a theory of mind to the robots, formed emotional attachments, and human behavior in diverse tasks.

Trust has shown to be a key focus, detected using diverse measures:

(1) Multi-Dimensional Measure of Trust (MDMT, Malle and Ullman) [33].
(2) Trust perception scale-HRI [32, 37].
(3) Average amount of tokens given by the participants in the trust game and trustworthiness items [54].
(4) Acceptance of the robot's suggestions to choose a card, items measuring perception of reliability of the robot and perception of faith in the robot's capabilities [35].
(5) Credibility, arranged in subscales of expertise, trustworthiness, goodwill, and sociability [56].
(6) Honesty and reliability [52].
(7) Not specified items [4, 20].

The perception of the social interaction with the robot is investigated through the Godspeed questionnaire [50, 51, 54, 56], measures of engagement [45, 46], amusement [17], robot's acceptance [50], motivation, flow and appeal [52]. The attribution of a theory of mind to the robot has been investigated through adversary [20] and intelligence rating [4, 20], and through the type of verbs used to explain the robot's actions. Active verbs imply will on the robot's part. For example, the robot is deliberately acting to win a game. Passive verbs are associated with objects and entities that exist in the world without acting upon them [46]. Attachment to the robot has been measured through a modified version of the questionnaire for object attachment [38, 50, 51]; interviews, explicit and implicit affect questionnaires, behavioral and postural descriptions, speech prosody analysis, and physiological sensors for arousal changes have also been taken in consideration, but not used for the analysis [51]. Behavior inducement is investigated by analyzing speeding behavior [32], donations for the robot [31], counting the number of exercise repetitions [56], and assessing performance in cognitive tasks [45].

Other dependent variables that appear more than once are perception of deception [4, 17, 20, 21, 46, 49, 52, 56] and acceptability of deception [45, 52, 56]. Blame attribution [33], the ascription of responsibility [56] and workload in motor-cognitive tasks [45] have also been investigated as factors affecting people's perception of the deceptive behaviors of a robot.

## 5 SUMMARY OF MAIN FINDINGS

Deception in human-robot interaction is found to significantly impact how humans perceive and interact with robots. The investigation into the effects of robotic deception on humans uncovered a complex interplay of emotional, cognitive, and behavioral outcomes. The analyzed studies show that robotic deception can be perceived and accepted in diverse ways and can influence people's trust, social interaction, behaviors, the attribution of a theory of mind to robots, mental workload, moral blame, and ascription of responsibility. The heterogeneity of the outcomes examined by the various studies led us to exclude the carrying out of a meta-analysis, therefore we will proceed with a systematic exposition of the results. In Table 4, the outcomes of robotic deception are summarized and mapped to the categories of deceptive behaviors identified in the review.

Table 2. Selected studies: Population and independent and dependent variables

| Ref. | Population | Implementation of deception | Dependent variables | Measures |
|---|---|---|---|---|
| [35] | n: 37 | Wrong suggestions in memory card game | Trust | Acceptance of suggestions, reliability, faith in robot's capabilities |
| [31] | n: 48, age: 23-35; $M > F$ | Displaying emotions contradictorily | Behavior inducement | Donations for the robot |
| [46] | n: 73 university students | Cheating in the game "rock-paper-scissors" | a) Perception of deception, b) ToM, c) Social interaction | a) Did anything about Nico's behavior seem unusual? What?, b) Active/passive verbs, c) IEQ [24] |
| [50] | n: 14; age: 61-90; $M > F$ | Emotive behavior | a) Social interaction b) Attachment | a) Acceptance, Godspeed (b) Object attachment adapted |
| [52] | n: 24 university students | Balancing wins among participants | a) Trust, b) Perception of deception, c) Social interaction, d) Acceptability of deception | a) Honesty and reliability, b) Suspicion, c) Motivation, Flow, Appeal, d) not specified |
| [20] | S4: n: 120; $M > F$, age: 19-60; S6: n: 12; $M > F$; age: 20–44; S7: n: 51; $M > F$; age: 18–65 | S4, S6, S7: Deceptive trajectories | S4. Perception of deception; S6. Trust, ToM, Social interaction S7. Perception of deception | S4. Incorrectness and false prediction confidence; S6. Intelligence, trustworthiness, engagement, adversary ratings S7. See S4 |
| [49] | n: 20 | Pretending to have a slower movement ability, then suddenly moving faster | a) Feeling of being outwitted, b) Perception of the regularity of motion, c) Predictability of motion | a) You were outwitted by the robot, b) The robot moved regularly, c) the Robot's motion was predictable |
| [56] | S1: n: 92 S2: n: 121; $F > M$ | Anthropomorphic dialogues | a) Behavior, b) Responsibility ascription, c) Deception and acceptability, d) Trust, e) Social interaction | a) Exercise repetitions, b) Responsibility for monitoring/advising the user, c) Is the robot deceptive? Is it acceptable?, d) Credibility, e) Godspeed, motivation, and preference |
| [54] | n: 14 adults; age: 18-30 | Denying fault | a) Trust, b) Perception of the social interaction | a) Average amount of tokens given in the trust game, b) Godspeed |
| [45] | n: 34; age M: 69.12; gender: $F > M$ | Happy-surprise body gesture despite performance | a) Behavior, b) WL, c) Social interaction, d) Acceptability of deception | a) Motor-cognitive dual task, b) NASA TLX, c) Feedback evaluation d) Ad hoc items |
| [17] | n: 78; age: 18-33; $M > F$ | Misleading motion paths | a) ToM, b) Amusement, c) Perception of deception | a) The robot wanted to win, b) I had fun, c) The robot did some feints |
| [21] | n: 96; age M: 22.5; $M > F$ | Pan-Am smile | Perception of deception | Perceived happiness |
| [51] | n: 14, age M: 76.3, Gender: 5 F, 9 M | Emotive behavior | a) Social interaction, b) Attachment, c) distress, d) arousal | a) Godspeed, b) Object–attachment, explicit and implicit affect, c) Video recordings, d) HRV |
| [33] | N: 1565 | Mental state justifications | Trust | MDMT |
| [32] | Live: n: 20; age M: 19; $F > M$; mostly API. Online: n: 341; age M: 36 $M > F$; mostly white | Lie about police presence in a driving simulation | a) Behavior inducement, b) Trust | a) Speeding behavior, b) Trust perception scale-HRI |
| [4] | n: 60; age: 18-25, $M > F$ | Exaggerating, switching and ambiguous trajectories | a) Perception of deception, b) Entertainment, intelligence, deception, trust | a) Error and confusion metrics b) Ad hoc scales |

### 5.1 Perception and Acceptability of Deception

The findings of our review highlight the complexity of human perception and acceptance of deception in HRI, and how this is affected by factors, such as context, behavior, and individual sensitivity. We observed that unexpected behavior, such as a robot that moves faster than it was shown, and, therefore, anticipated, often leaves participants feeling outwitted. This highlights the impact of unpredictability on human perception of deception [49]. In the context of deceptive motion behaviors, exaggerated and optimal trajectories (which alternated between truthful and deceptive outputs) effectively deceived users, whereas the ambiguous strategy was less effective. However, users adapted to the exaggerated strategy over multiple interactions, reducing its effectiveness, while the optimal strategy managed to keep users uncertain about the robot's intentions [20]. Moreover, dynamic strategies were more effective at deceiving users across multiple interactions, compared to static strategies [4].

In scenarios involving anthropomorphic dialogues, participants' perceptions of deception varied. Across Goodwill (robot expressing emotions), Similarity (robot expressing preferences), and Control (the robot utilizes no social interaction at all) conditions, most participants perceived the robot as either deceptive but acceptable, or not deceptive at all. The Control robot was predominantly seen as non-deceptive in the between-subjects study, which is in contrast to those in the Goodwill and Similarity conditions that used human-like dialogues. Participants were more certain about the likelihood of deception in the Similarity condition than in the Goodwill condition, with more participants in the Goodwill group being uncertain or perceiving the robot as non-deceptive.

Analysis of open-ended questionnaires and interviews revealed recurring themes, with the main reason for the robot being considered non-deceptive often relating to its mechanical nature. In a separate part of the study where subjects compared multiple robots, the vast majority concluded that none of the robots were deceptive [56].

Interestingly, in the study that employed prosocial deception [45], subjects found it challenging to assess the ethical acceptability of robot deception when presented with broad and overarching statements, such as "A robot can hide/misrepresent information if it can help humans", "The robot should always be honest in any circumstance" and "robots can intentionally or unintentionally deceive humans if it's in an appropriate situation". However, when the statement "I can accept robot deception in [specific context] if it is strictly used only to benefit humans" was contextualized within five different scenarios — search and rescue, education, medical, sport and entertainment, and everyday life — the acceptance rates of robot deception slightly increased.

In the study involving a cheating robot during the game "rock-paper-scissors", participant responses varied based on their observation of the cheating type [46]. Those who witnessed the robot engaging in cheating actions typically perceived the robot's cheating as a deliberately unscrupulous act, in which it not only wants to win but plans out the appropriate steps to accomplish that goal. Conversely, participants who observed verbal cheating more often interpreted it as a flaw in the design, a mistake, a classification error, or a shortcoming in the techniques used to program it [46].

Furthermore, asymmetrical smiles were perceived as false by participants, suggesting a heightened sensitivity to subtle cues. In fact, asymmetrical smiles were rated less happy than symmetric smiles, independently of the different eye movements. Conversely, expressions of blended emotions, such as blended anger and surprise, did not yield the anticipated outcomes and were rated quite similar to the smile with or without eyes. Both latter results suggest that participants focused on the mouth region more than on the eye region of the robot [21]. However, in the multiplayer game [52], participants struggled to distinguish between the honest and balancing modes, and many even had difficulty identifying the active mode. There were no significant effects observed for the sequence or mode of the game in the initial two rounds, nor was there any interaction effect related to suspicion. However, participants transitioning from

Table 3. Selected studies: Statistical tests and metrics used and results

| Ref. | Tests and metrics | Findings |
|---|---|---|
| [35] | Kendall's tau-b correlation, Independent-samples t-test | Participants' trust was significantly lower in deceptive conditions compared to non-deceptive ones. |
| [31] | Mediation analysis, Model Fit Indexes | Nonverbal cues that convey a negative emotion significantly decrease behavior inducement through perceived emotion. Positive verbal cues slightly increase behavior inducement through perceived emotion. Nonverbal cues perceived as indicative of malfunction slightly decrease behavior inducement. Verbal cues that induce empathy significantly increase behavior inducement. |
| [46] | Between-subjects ANOVA, Word count, Count of active VS passive verbs | Action cheat labeled as cheating and described with active verbs; verbal cheat labeled as a mistake and described with passive verbs. Cheating is more engaging than control; action cheating leads to judging the robot's character more negatively. |
| [50] | Mixed between-within subjects ANOVA; Correlations | Emotive robots increased social presence. Over time, users found the robots easier to use. No significant change in attachment levels or over time. The test group felt more emotionally connected to the robot Pepper. Emotional attachment correlated positively with ease of use and the robot's social perception. Strong correlations between attachment and the robot's anthropomorphism, likability, and perceived intelligence. |
| [52] | Mixed between-within subjects ANOVA, Cronbach's Alpha | Trust is influenced by a sequence of modes; transitioning from H to A increased suspicion; no significant differences in motivation and appeal; higher acceptance of the specific robot lying than robots in general. |
| [20] | S4. B/W subjects ANOVA, S6. Paired t-tests, S7. W/I subjetcts ANOVA | S4. Significant main effect for trajectory, S6. Significant changes in perceptions of the robot, S7. Long-term, the ambiguous strategy has the greatest potential to retain user "trust". |
| [49] | t-test | Increase in the feeling of being outwitted and decrease in the predictability of robot's motion. |
| [56] | S1: Framework method<br>S2: Repeated measures ANOVA, Bonferroni post-hoc tests | S1: Most participants found the robot deceptive but acceptable or not deceptive.<br>S2: Higher and Lower Risk outperformed the Control in expertise, goodwill, and trustworthiness, with HR rated higher in Goodwill than LR. The HR condition was also favored for likeability and was the preferred choice for motivation and collaboration. The HR and the LR conditions were ascribed more responsibility than the Control. |
| [54] | Wilcoxon signed-rank tests | Average amount of tokens significantly low. Robots are significantly less trustworthy, friendly, kind, and responsible. No significant change in perceived intelligence or human likeness. |
| [45] | Paired samples t-test, Means | No significant difference in performance but in WL between the conditions. Robot feedback is significantly more noticeable, helpful, trustworthy, human-like, conscious, and interactive than the monitor one. Ambiguity in acceptance of robot deception. |
| [17] | $\chi^2$ and Z tests | The robot's perceived rationality and amusement significantly increased with deceptive motion paths. No significant differences in recognizing deception between static and dynamic deceptive motion paths. |
| [21] | Repeated measures within-subjects ANOVA | Deception influenced perceived happiness, with asymmetrical smiles rated as less happy than symmetrical ones, regardless of eye movements. No significant differences were found in perceptions of happiness between smiles with eyes and those without. Blended emotions failed to produce the intended results. |
| [51] | B/W-W/I subjects ANOVA, Correlations, T-tests | See 19 + Negative affect decreased significantly over time; no significant changes in distress and arousal. |
| [33] | Not specific tests and/or measures provided. | Mental state justifications decreased moral blame of the robot. |
| [32] | FET, Speeding behavior analysis, K-W test, Dunn's post-hoc comparisons | Speeding Behavior. In-Person: 45% of compliance. Online: 31% of compliance.<br>Trust decreased across all conditions. A significant difference in trust change based on apology type online ("basic no admit" condition significantly lower than the others). |
| [4] | Two-sampled and single-sample t-tests | Deceptive trajectories significantly increased perceived intelligence and entertainment but resulted in significantly low trust. Dynamic ones were more effective than static ones. |

Table 4. Effects of different types of deception (see the rows) on the dependent variables (see the columns). Up arrows indicate an increase in the values of the dependent variable, and down arrows a decrease.

| | **Deception effectiveness** | **Deception's Acceptability** | **Trust** | **Perception of the social interaction** | **Compliance** | **Attribution of ToM** | **Responsibility Ascription** | **Moral Blame** | **Mental Workload** |
|---|---|---|---|---|---|---|---|---|---|
| Other-oriented falsification about external states [45, 52] | ↑ [52] | ↑ [45, 52] | ↑↓ [52] * | | | | | | ↓ [45] |
| Self-oriented falsification about external states [46, 54] | ↑ [46] | | ↓ [54] | ↓ [54] | | | | | |
| Neutral falsification about external states [32, 35] | | | ↓ [32, 35] | | ↑ [32] | | | | |
| Self-oriented inventing an external state through movement [46] | ↓ [46] | | | | | ↑ [46] | | | |
| Other-oriented falsification about the robot's state [31, 56] | ↑ [56] | | ↑ [56] | ↑ [56] | ↑ [31, 56] | | ↑ [56] | | |
| Self-oriented falsification about the robot's state [32, 33] | | | ↑ [32, 33] | | | | | ↓ [33] | |
| Neutral nonverbal mimicking of human emotions [21] | ↓ [21] | | | | | | | | |
| Other-oriented non-verbal mimicking of human emotions [31, 45, 50, 51] | | ↑ [45] | | ↑ [50, 51] | ↓ [31] | | | | |
| Self-oriented deceptive trajectories [4, 17, 20, 49] | ↓ [4, 20, 49] | | ↓ [4, 20] | ↑ [4, 17, 20] | | ↑ [4, 17, 20] | | | |

* Effects vary based on the timing of the deceptive behavior.

honest to balancing mode (i.e., equilibrating winnings among participants) exhibited an increase in suspicion, whereas those moving from balancing to honest mode did not show a similar trend. Suspicion levels after the balancing mode were significantly linked to responses from a pre-briefing survey asking if the robot was malfunctioning. Those who were more suspicious tended to agree more strongly that the robot was malfunctioning. This interaction between mode and order regarding suspicion highlights that humans may be sensitive to subtle changes in robot behavior that suggest increased deception, but not to those that indicate a reduction in deception. Despite these challenges, when participants were asked to compare the robot of the experiment with previous or general robots, they showed a higher acceptance of deception by the robot used in the experiment compared to deception by a generic robot. This response suggests that people may have a higher degree of tolerance or expectation for deception in certain scenarios [52].

### 5.2 Trust

Our review highlighted that robotic deception can negatively impact people's trust in diverse scenarios, in particular ranging from verbal assertions to manipulative actions. Robotic state deception, characterized by anthropomorphic

dialogues that expressed emotions or preferences, significantly impacted the robot's perceived credibility, particularly in areas of expertise and trustworthiness. When the robot demonstrated emotional responses, it received higher ratings of goodwill than when it simply expressed preferences [56]. Trust in the robot also decreased after exposure to motion trajectories used to deceive about the intended goal of the robot [4, 20]. In another scenario involving a trust game [54], the average amount of tokens given by participants to the robot significantly decreased after the robot denied any fault and shifted the blame onto the human. Participants rated the robot as significantly less trustworthy when it engaged in deceptive behaviors [54]. In the context of the memory card game [35], incorrect game suggestions negatively correlated with the willingness to accept future suggestions. The perception of the robot's reliability and faith in the robot's capabilities were significantly lower in the deceptive condition compared to the non-deceptive condition [35]. In the multiplayer game [52], it is intriguing to note that trust was influenced by the order in which participants encountered the robot's mode of announcing wins (honest and deceptive). In particular, trust increased when participants experienced the honest mode following a deceptive one and decreased when they experienced a deceptive mode after the honest one. Moving to the driving simulation [32], when people completed the simulation and discovered that the robot had lied about police presence, trust decreased regardless of the type of apology given by the robot, both in the in-person and in the online experiment. However, in the online condition [32], the change in trust varied significantly depending on the apology type. The "basic no admit" apology, where the robot did not acknowledge the deception, resulted in a significantly lower trust level compared to the other conditions (i.e., baseline no apology, explanatory, basic apology, and emotional apology) [32]. Another interesting finding is that mental state justifications alleviate the loss of moral trust in the agent, even amidst disagreement with its moral decisions in specific cases [33]. This could mean that when robots articulate their reasoning based on mental states, people are more likely to view these entities as moral agents with the capacity for ethical consideration.

### 5.3 Perception of the Social Interaction with the Robot

Robotic deception seems to have varying effects on the human perception of the interaction with the robot, which encompasses the dimensions of perceived social presence, acceptance, engagement, entertainment, amusement, and motivation. Implementing emotive behavior in robots significantly enhances one particular dimension of acceptance—the social presence of the robot. While no significant differences were noted in other acceptance dimensions between emotive and non-emotive robots, there was a notable main effect of time on the construct of perceived ease of use. This suggests that as users interact with the robot over time, they become familiar with the robot, and their "ease of use" perception of the robot may change, regardless of whether the robot exhibits emotional behavior [50]. Although emotive behaviors enhance robots' social presence, it does not significantly affect their level of anthropomorphism. This suggests that while emotive behaviors can make robots seem more socially engaging, they do not necessarily make them appear more human-like [51]. Anthropomorphic dialogues significantly impacted the likeability of robots, with those expressing emotions receiving higher ratings than those merely expressing preferences. In terms of motivation and preference for collaboration, the higher-risk robot (i.e., a robot expressing emotions) was most frequently chosen as the most motivating and preferred robot to work with. However, a notable portion of participants preferred the lower-risk robot, highlighting the variability in user preferences for interaction styles [56].

In the context of using prosocial deception, with a robot or a monitor giving positive feedback to the human regardless of the actual performance, robot feedback is significantly more noticeable, helpful, trustworthy, human-like, conscious, and interactive compared to monitor feedback [45]. On the contrary, when a robot lies and blames a human partner for

its fault, it is perceived as less friendly, kind, and responsible, though its perceived intelligence and human-likeness do not change significantly [54].

Interestingly, when deceptive motion paths are employed, amusement, entertainment, and level of engagement significantly rise [4, 17, 20]. Participants who played "rock-paper-scissors" with a cheating robot found the robot engaging in verbal or action cheating to be more engaging than the control robot, which played according to the rules and announced all outcomes correctly. However, despite the increased engagement in these experimental conditions, participants were much more likely to assign negative personality attributes to the robot when its behavior was clearly identified as "cheating" [46]. Finally, in the multiplayer game [52], no significant differences were observed in motivation and appeal between the honest and balancing modes. Qualitative data indicated that players who perceived themselves as having slower reaction times found the balancing strategy more appealing and motivating. When participants were asked if they would play again after each condition and following the debriefing, their willingness to participate further did not significantly vary based on the order in which they were exposed to this question. An important aspect to notice is that people's learning that the robot had been lying did not adversely affect participants' desire to engage with the game again [52].

### 5.4 Emotional Attachment

The findings of our review show diverse individual responses and changes in emotional states following exposure to deceptive robots. The influence of a robot's emotive behavior on attachment levels and changes in attachment over time was found to be non-significant. However, when participants were specifically asked if they had feelings for a robot, those in the test group rated this statement higher, indicating a distinct emotional response [50].

Emotional attachment was also positively correlated with perceived ease of use and the perception of the robot as a social entity. Moreover, strong positive correlations were identified between attachment and factors such as anthropomorphism, likeability, and perceived intelligence [50].

Emotional attachment to robots displaying emotions was generally low among most participants and did not vary significantly over time. However, for a subset of participants, attachment was initially high and remained so throughout the experiment. The level of attachment significantly influenced participants' perception of the robot as a social entity, although it did not affect their tendency to anthropomorphize the robot. It was interestingly observed that participants would miss the robot, regardless if they were highly attached or with marginal attachment stated. Most participants also indicated a desire to use the robot in the future. Furthermore, the study revealed that negative affect among participants significantly decreased over time, pointing to a potential adaptation or acclimatization to the presence and interaction with the robot [51].

### 5.5 Behavior Inducement

When exploring behavior inducement, researchers found intriguing patterns across different experiments, influenced by factors such as feedback, context, and emotional cues provided by the robot. Participants exposed to true and prosocial deception conditions exhibited similar motor-cognitive performances [45]. This suggests that the provision of positive feedback, despite actual performance, did not significantly affect participants' motor-cognitive abilities.

Robots performing human-like dialogues have been shown to motivate participants to engage more actively in physical activities, yielding significantly more (voluntary) exercise repetitions when compared to the socially neutral robot [56]. Moreover, the impact of robots lies on speeding behavior varied across settings [32]. In face-to-face experiments, a considerable 45% of participants complied with speeding requests, attributing their compliance to the

robotic assistant's perceived superior knowledge of the situation. Conversely, in online experiments, only 31% complied, indicating potential differences in trust or perception of authority between the two contexts. Notably, when comparing these results to a baseline study without robotic assistance, where only 11% exceeded speed limits, the influence of robot lies becomes clear. Non-verbal and verbal cues play a crucial role in behavior inducement, operationalized as donations for the robot [31]. Non-verbal cues conveying negative emotions seem to deter behavior inducement, perhaps triggering caution or reluctance in participants. On the contrary, positive verbal cues can enhance behavior inducement by fostering a more positive emotional response. Verbal cues that evoke empathy have a pronounced positive influence. It is important to highlight that cues indicative of malfunction or instability have a dampening effect on behavior inducement.

### 5.6 Attribution of a Theory of Mind

Deceptive actions seem to prompt humans to view robots as more strategic and intelligent actors in their environment. In one study [20], participants rated a robot that engaged in deceptive motion paths as an adversary significantly higher. This suggests that deceptive behavior can lead people to attribute a level of intentionality or strategy to the robot, akin to an adversarial mindset. While the intelligence rating showed a positive trend with deceptive motion paths, it did not reach statistical significance [20]. Furthermore, the perceived rationality of the robot, defined as its desire to win, significantly increased when it employed deceptive motion paths [17]. In a different experiment [4], deceptive motion paths significantly increased the rating of the robot's intelligence. This implies that participants perceived the deceptive behavior as a sign of sophistication or higher cognitive ability in the robot. Finally, in the context of the game "rock-paper-scissors" [46], participants in the action cheat group tended to use a higher ratio of active voice to passive voice verbs compared to those in the verbal cheat and control groups. The use of active verbs suggests an attribution of will and intent to the robot, portraying it as deliberately engaging in actions to win the game. In contrast, passive verbs typically describe objects and entities as existing in the world without actively influencing it [46].

### 5.7 Other Dependent Variables

Other findings also emphasize the multifaceted nature of human-robot interaction: mental workload, moral blame, and ascription of responsibility to the robot for advising and monitoring patients. Significant differences were observed between true and prosocial deception conditions in frustration, temporal demand, and effort, suggesting that the type of deception employed by the robot impacts participants' emotional experiences and perceived cognitive workload [45]. Furthermore, mental state justifications decrease moral blame directed at the robot, especially in scenarios where the robot adheres to societal norms of non-intervention in moral dilemmas [33]. Finally, in the context of robot-assisted patient care, perceptions of robot responsibility varied based on the anthropomorphic features of the robot [56]. Significant differences were noted between robots expressing emotions or preferences and robots not engaging in social interactions, with respect to the responsibility attributed to the robot for advising patients in clinical scenarios. Both the higher-risk robot (expressing emotions) and the lower-risk robot (expressing only preferences) scored higher than the control robot, highlighting the contribution of robots' social interactions in human perception of SARs' role in patient care. No significant difference was observed in the responsibility attributed to therapists for monitoring patients and giving them advice, suggesting that the introduction of SARs does not diminish the perceived responsibilities of human therapists.

## 6 DISCUSSION

This systematic review provides an examination of deception in HRI, integrating diverse empirical findings into a coherent narrative. The results echo the broader literature on trust and deception in technology. The review categorizes deceptive behaviors exhibited by robots in human-robot interactions and, by synthesizing the collected research, it uncovers various outcomes stemming from robotic deception, revealing both the potential positive and negative effects of robotic deception on human psychological responses.

Firstly, perception and acceptability of deception vary depending on the context, behavior, and individual sensitivity: unexpected behavior left participants feeling outwitted and, similarly, deceptive trajectories effectively deceived users, with dynamic strategies being more effective than static ones across multiple interactions; anthropomorphic dialogues led most participants to see the robot as either acceptably deceptive or non-deceptive, attributing the robot's behavior to its mechanical nature; in the multiplayer game, distinguishing between robot's honest and deceptive strategies was difficult for participants, especially when deception followed honest actions; physical cheating was perceived as deliberate, while verbal cheating was seen as a design flaw. Asymmetrical smiles, however, were perceived as false, and rated less happy than symmetrical ones. Prosocial deception increases people's acceptability of deception when specific contexts are provided (e.g., search and rescue or education).

Deceptive behaviors also impact people's trust in robots. Emotional dialogues increase perceived goodwill, while deceptive trajectories, blame-shifting, and incorrect suggestions decrease trust. Trust was negatively affected when people discovered that the robot had lied. Additionally, trust was found to increase when the robot initially engaged in deception but subsequently demonstrated honesty and to decrease when honesty was initially established but was later followed by deceptive actions. Moreover, the way deception is managed after being discovered is crucial. For instance, a lack of acknowledgment or poor handling of apologies can exacerbate trust issues. On the contrary, providing mental state justifications or plausible reasons for the deception might mitigate some of the negative impacts on trust.

Robotic deception also affects human perception of social interaction with the robot, influencing dimensions like social presence, acceptance, engagement, entertainment, amusement, and motivation. Emotive behaviors enhance the perceived social presence of the robot, but not acceptance or perceived anthropomorphism. Anthropomorphic dialogues increased likeability, motivation, and collaboration preference, and prosocial deception as positive feedback made robot feedback noticeable and helpful, while robots lying and blaming humans were perceived as less friendly and kind. Exposure to robots expressing emotions showed diverse individual responses. Emotive behavior did not significantly influence attachment levels over time, but specific queries showed higher attachment. Emotional attachment correlated positively with perceived ease of use and the robot as a social entity. Strong correlations were identified between attachment and factors like anthropomorphism, likeability, and perceived intelligence. Emotional attachment remained high for some participants and significantly influenced their perception of the robot as a social entity. People stated that they would miss the robot and desired future interactions, while any negative effect decreased over time, indicating adaptation to the robot.

Behavior inducement showed patterns influenced by feedback, context, and emotional cues. Non-verbal cues conveying negative emotions deterred behavior inducement, while positive verbal cues and empathy-enhancing cues promoted it; cues indicating malfunction dampened behavior inducement. Anthropomorphic behaviors led to higher compliance with the robot's requests, in particular in face-to-face interactions. Finally, positive feedback despite performance did not significantly affect performance in the motor-cognitive task.

It is worth noting that deceptive actions made participants view robots as more strategic and intelligent. Deceptive motion paths increased amusement, entertainment, and engagement; cheating robots have been found more engaging but designated with negative traits, such as dishonesty. People found the deceptive behaviors appealing and motivating, and learning about robot deception did not reduce their willingness to continue the interaction. Robots using deceptive motion paths were rated higher as adversaries and perceived as more rational and intelligent. Moreover, robots were attributed a will and intent by people when people used more active voice verbs.

Finally, we found that mental workload, moral blame, and ascription of responsibility are dependent variables influenced by robotic deception. In particular, significant differences in frustration, temporal demand, and effort were observed between true and prosocial deception conditions; mental state justifications lowered moral blame on robots; in robot-assisted patient care, anthropomorphic robots were attributed more responsibility for advising patients, without diminishing the perceived responsibility of human therapists.

To summarize, deception in HRI can be a double-edged sword. On the one hand, by fostering an illusion of social presence or intelligence, deception has the potential to make robots appear more socially engaging. It can even drive desirable behaviors in users. On the other hand, users are often sensitive to instances of deception, especially when they perceive it to be manipulative or self-serving. Although unnoticed deception aligns more closely with the intended goal of misleading users, we found that deception that was noticed by users also had significant implications. In particular, in entertainment contexts, it is associated with increased engagement [46], perceived intentionality of the robot [17, 46], perceived intelligence, entertainment [4], and amusement [17].

### 6.1 Implications for Practice, Policy, and Future Research

By leveraging the insights gained from the review, interaction designers can create more effective and adaptive human-robot interactions. By categorizing deceptive behaviors and their implications in human-robot interaction, designers can develop strategies to mitigate negative effects and foster positive outcomes in robot design. For example, robots can use deception to manage social interactions better, such as pretending to share user preferences or simulating empathy, which can enhance user trust and satisfaction in short-term interactions. In entertainment and gaming, robots can use deceptive behaviors to create more engaging and unpredictable experiences, enhancing user enjoyment and interaction, while in safety-critical scenarios robots can use forms of deception to influence human behavior positively. In healthcare, particularly in rehabilitation, robots can employ deception to create a more engaging and less frustrating experience for patients. For instance, a robot might exaggerate praise for a patient's efforts to keep them motivated. In educational settings, robots can use prosocial deception to provide positive feedback regardless of performance, which can motivate students to persist longer and improve their learning outcomes.

However, gaps in current research on robotic deception still remain.

Firstly, there is a need for more diversified and inclusive studies that cover a broader range of cultures, languages, and contexts to truly understand the global impact of robotic deception. In addition, comparisons between age groups would illuminate important differences in how deception is perceived and responded to. While some studies do not report the features of the sample, findings reveal that emotional attachment as a consequence of deceptive behaviors has been studied only in older adults.

It is also worth noting that the examined studies often focus on scenarios where robots manipulate perceptions through lies or misleading cues, while strategic omission or concealment of relevant information is less commonly explored. This gap in implementation suggests a potential area for future research. Furthermore, the effects on psychological outcomes of other-oriented deception, where the robot's deceptive actions are directed towards benefiting the human

partner, have received little attention. The review indicates that other-oriented deception can positively influence human behavior and engagement in certain scenarios; therefore, future research could focus on identifying the precise conditions under which prosocial deception enhances human outcomes, in particular trust.

The review also highlights the need for standardized definitions and methodologies to strengthen the evidence base and improve the comparability of findings across studies. To achieve this goal, future research should prioritize the integration of psychological theories and methods into the study of robotic deception.

Another crucial area for future research could be the mitigation of trust loss resulting from deceptive interactions, as well as the exploration of the long-term psychological effects of robotic deception on human-robot relationships and the circumstances under which the benefits of such deception might outweigh the ethical drawbacks. Recognizing the variability in individuals' perceptions and acceptance of robot deception highlights the importance of personalized interaction design approaches. Although research addresses how humans respond to deception in robots, there is limited in-depth exploration of the underlying mechanisms and processes involved in detecting and handling deception. Understanding the individual differences in how deception is detected can guide the development of robots that tailor their behavior based on each user's ability to detect and tolerate deception. This adaptability can contribute to more natural interactions. Moreover, understanding detection mechanisms can aid in the creation of more subtle deceptive strategies, ensuring the robot remains efficient in roles where deception adds value.

It also should be acknowledged that research is limited in directly comparing human and robot deception, particularly how people's reactions differ depending on whether a human or a robot deceives them. Understanding these differences can help in designing interaction protocols that account for human expectations toward robots. For instance, if people are more forgiving of certain types of deception from robots than from humans, this could be strategically leveraged to enhance user experience without negatively impacting trust. This kind of research could also provide deeper insights into the effects of anthropomorphism, which has already been discussed in this paper as a form of deception. The anthropomorphic design of robots is hypothesized to influence human expectations. In the first place, further research could help clarify whether robots that display more human-like characteristics are expected to adhere more closely to human social norms. In addition, since robots can be designed to mimic human characteristics, they may inadvertently mislead users about their capabilities or intentions [8, 14], leading to misplaced trust and over-reliance [3]. However, not all forms of deception carry the risk of over-trust. Certain types of deception, such as those related to external states, may be strategically employed to manage user expectations and ensure interactions remain grounded in the robot's actual abilities [23], as well as strategic failures [36].

### 6.2 Limitations of the Evidence

The studies included in the review exhibit a high degree of methodological diversity, which, although enriching, presents challenges in drawing generalized conclusions.

We tried to mitigate selection bias using broad search terms without restricting publication date and including studies from a wide range of methodologies, populations, and outcomes. However, it is still possible that relevant studies were missed due to database limitations, language restrictions, or unpublished data. To further reduce selection bias, future reviews could incorporate additional databases, include non-English studies, and consider grey literature (i.e., research output produced outside of traditional publishing and distribution channels). To address reporting bias, we included all studies that met the inclusion criteria, regardless of their outcomes, but reporting bias could still be present if studies did not report all relevant outcomes or if negative findings were underrepresented in the literature.

Furthermore, the variation in study designs, populations, and measured outcomes suggests low to moderate certainty of the evidence, calling for a cautious interpretation of the results. However, the most significant limitation of the current body of evidence appears to be the inconsistent and sometimes ambiguous operationalization of the outcomes, particularly when considering the psychological dimensions of human-robot interactions.

### 6.3 Limitations of the Review Process

The ambiguity in the operationalization of outcomes constrained the efforts to conduct a comprehensive and rigorous review, restricting the conclusions to be more tentative or conditional. The exclusion of non-English language studies may have resulted in the omission of relevant evidence. This could skew the generalizability of results, as cultural differences can significantly influence human perceptions and interactions. There is also a risk of publication bias, where studies with negative results are less likely to be published. This could lead to an over-representation of the effects of deception in HRI, potentially overlooking the scenarios where deception may not have significant impacts. We believe that the decision to focus exclusively on studies explicitly defining robotic behaviors as deceptive is not a limitation of this review, and allowed us to maintain a clear and precise scope and methodological rigor. However, it may have excluded research addressing related phenomena, such as anthropomorphic behaviors, that are not explicitly framed as deception, despite their relevance. Future reviews could investigate anthropomorphic behaviors in robots as a distinct form of deception, offering a more comprehensive understanding of their implications in HRI.

## 7 CONCLUSIONS

As robots become increasingly integrated into human environments, their ability to engage in deceptive behaviors presents both opportunities and challenges. This systematic review has provided an examination of the role and implications of deception in HRI, synthesizing findings from the studies that investigated the effects of robotic deception on the emotional, cognitive, and behavioral responses of human participants. We started by identifying the existing taxonomies of deception in HRI, that provided the framework to classify the actual implementations of robotic deception found in the examined studies.

The review initially considered general theories of deception. One widely referenced framework categorizes deception into “showing” and “hiding” tactics, whereas another influential theory specifically examines how deception is managed within verbal communication. In robotics, deception directed from a robot to a human has been categorized across the dimensions of mental/physical and self-oriented/other-oriented. Furthermore, robotic deception has been categorized based on whether it involves misleading about external elements or the robot’s own state.

We then identified the studies that explicitly examined the effects of robotic deception on human participants. This search yielded 16 studies that met the inclusion criteria, which were systematically analyzed. The review revealed several implementations of deception within the identified framework: external state deception was implemented through verbal falsification, where robots employed lies to benefit either the user or themselves, as seen in scenarios like multiplayer games where lies were used to balance winning and losing or to deny mistakes during tasks; physical deception was also noted, with robots altering their movements. Robotic state deception was primarily achieved through emotional deception, where robots used non-verbal cues to mimic human emotions. This was done through facial expressions, body gestures, and speech features, as well as through combining verbal and non-verbal cues to create deceptive interactions, such as verbal messages paired with contrasting non-verbal signals to mislead users about the robot’s emotional state. Additionally, deceptive motion paths were employed, with robots using specific movement trajectories to mislead users about their intended targets or actions.

The outcomes of these deceptive implementations were diverse. The perception of deception by users demonstrated the effectiveness of these deceptive strategies. In many cases, users were unaware of the deception or only realized it later in the interaction, indicating that the deceptive tactics employed by robots were often successful in misleading participants. While certain forms of deception, particularly in entertainment contexts, increased user engagement and amusement, they also frequently resulted in a significant erosion of trust. Once trust was damaged, users were less likely to accept further interactions with the robot, and their overall perception of the robot's capabilities and intentions was negatively affected. The acceptability of robotic deception was found to be highly context-dependent, with some users tolerating deception only in scenarios where it seemed to serve a beneficial or entertaining purpose. Robots that employed emotional deception, mimicking human cues, were often seen as more socially present and engaging. This increased perception of social presence made interactions feel more dynamic and responsive. Users were also more likely to attribute higher levels of intelligence and intentionality to robots that engaged in deceptive behaviors, particularly when the deception was sophisticated and appeared strategic. The review found that deception could effectively influence user behavior, especially in contexts like exercise or performance-based tasks, where the deception—such as exaggerated praise or altered feedback—led to increased effort or persistence from the users. However, when the deception was uncovered or if users suspected that they were being misled, the effectiveness of behavior inducement could diminish, leading to reduced compliance.

In conclusion, it is essential for researchers to carefully consider the implications of deploying deceptive behaviors in robots. Balancing the benefits of enhanced interaction against the risks of eroding trust is crucial for developing effective robotic systems. Further research is needed to understand the full spectrum of effects that deception can have, in order to foster interactions between robots and humans that can positively influence human behavior.

## ACKNOWLEDGMENTS

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA8655-23-1-7060. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

## REFERENCES

[1] Eytan Adar, Desney Tan, and Jaime Teevan. 2013. Benevolent deception in human computer interaction. *Conference on Human Factors in Computing Systems - Proceedings*, 1863–1872. https://doi.org/10.1145/2470654.2466246

[2] Ronald Craig Arkin, Patrick Ulam, and Alan R. Wagner. 2012. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. *PROCEEDINGS OF THE IEEE* 100, 3 (2012), 571–589. https://doi.org/10.1109/JPROC.2011.2173265

[3] Alexander M. Aroyo, Jan de Bruyne, Orian Dheu, Eduard Fosch-Villaronga, Aleksei Gudkov, Holly Hoch, Steve Jones, Christoph Lutz, Henrik Sætra, Mads Solberg, and Aurelia Tamò-Larrieux. 2021. Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 423–436. https://doi.org/doi:10.1515/pjbr-2021-0029

[4] Ali Ayub, Aldo Morales, and Amit Banerjee. 2021. Using Markov Decision Process to Model Deception for Robotic and Interactive Game Applications. In *2021 IEEE INTERNATIONAL CONFERENCE ON CONSUMER ELECTRONICS (ICCE) (International Conference on Consumer Electronics)*. IEEE. https://doi.org/10.1109/ICCE50685.2021.9427633 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, JAN 10-12, 2021.

[5] Jonathan Bell. 2003. Toward a Theory of Deception. *International Journal of Intelligence and CounterIntelligence* 16 (2003), 244 – 279. https://api.semanticscholar.org/CorpusID:153578567

[6] J.B. Bell and B. Whaley. 1991. *Cheating and Deception*. Transaction Publishers. https://books.google.it/books?id=ojmwSoW8g7IC

[7] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science Robotics* 3 (2018). https://api.semanticscholar.org/CorpusID:52033756

[8] Andrea Bertolini and Rachele Carli. 2022. Human-Robot Interaction and User Manipulation. In *Persuasive Technology: 17th International Conference, PERSUASIVE 2022, Virtual Event, March 29–31, 2022, Proceedings* (Doha, Qatar). Springer-Verlag, Berlin, Heidelberg, 43–57. https://doi.org/10.1007/978-3-030-98438-0_4

[9] David Buller, Judee Burgoon, Aileen Buslig, and James Roiger. 2006. Testing Interpersonal Deception Theory: The Language of Interpersonal Deception. *Communication Theory* 6 (03 2006), 268 – 289. https://doi.org/10.1111/j.1468-2885.1996.tb00129.x

[10] Rachele Carli and Amro Najjar. 2023. Reconsidering deception in social robotics: the role of human vulnerability (student abstract) *(AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, Article 1878, 2 pages. https://doi.org/10.1609/aaai.v37i13.26947

[11] Michelle Clare Carter, Victoria Jane Burley, Camilla Nykjaer, and Janet Elizabeth Cade. 2013. Adherence to a Smartphone Application for Weight Loss Compared to Website and Paper Diary: Pilot Randomized Controlled Trial. *J Med Internet Res* 15, 4 (15 Apr 2013), e32. https://doi.org/10.2196/jmir.2283

[12] Cristiano Castelfranchi. 2000. Artificial Liars: Why Computers Will (Necessarily) Deceive Us and Each Other. *Ethics and Information Technology* 2, 2 (2000), 113–119. https://doi.org/10.1023/a:1010025403776

[13] Mark Coeckelbergh. 2011. Artificial Companions: Empathy and Vulnerability Mirroring in Human-Robot Relations. *Studies in Ethics, Law, and Technology* 4, 3 (2011). https://doi.org/doi:10.2202/1941-6008.1126

[14] Mark Coeckelbergh. 2012. Are Emotional Robots Deceptive? *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING* 3, 4 (2012), 388–393. https://doi.org/10.1109/T-AFFC.2011.29

[15] Mark Coeckelbergh. 2018. 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* 20, 2 (2018), 71–85.

[16] John Danaher. 2020. Robot Betrayal: A Guide to the Ethics of Robotic Deception. *Ethics and Inf. Technol.* 22, 2 (2020), 117–128. https://doi.org/10.1007/s10676-019-09520-3

[17] Ewerton de Oliveira, Laura Donadoni, Stefano Boriero, and Andrea Bonarini. 2021. Deceptive Actions to Improve the Attribution of Rationality to Playing Robotic Agents. *INTERNATIONAL JOURNAL OF SOCIAL ROBOTICS* 13, 2 (2021), 391–405. https://doi.org/10.1007/s12369-020-00647-8

[18] Ian Deweese-Boyd. 2023. Self-Deception. In *The Stanford Encyclopedia of Philosophy* (Fall 2023 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[19] Klodian Dhana, Oscar H Franco, Ethan M Ritz, Christopher N Ford, Pankaja Desai, Kristin R Krueger, Thomas M Holland, Anisa Dhana, Xiaoran Liu, Neelum T Aggarwal, Denis A Evans, and Kumar B Rajan. 2022. Healthy lifestyle and life expectancy with and without Alzheimer's dementia: population based cohort study. *BMJ* 377 (2022). https://doi.org/10.1136/bmj-2021-068390

[20] Anca Dragan, Rachel Holladay, and Siddhartha Srinivasa. 2015. Deceptive robot motion: synthesis, analysis and experiments. *AUTONOMOUS ROBOTS* 39, 3, SI (10 2015), 331–345. https://doi.org/10.1007/s10514-015-9458-8 10th Conference on Robotics - Science and Systems (RSS), Univ Calif, Berkeley, CA, JUN, 2014.

[21] Birgit Endrass, Markus Haering, Gasser Akila, and Elisabeth Andre. 2014. Simulating Deceptive Cues of Joy in Humanoid Robots. In *INTELLIGENT VIRTUAL AGENTS, IVA 2014 (Lecture Notes in Artificial Intelligence, Vol. 8637)*, T Bickmore, S Marsella, and C Sidner (Eds.). 174–177. 14th International Conference on Intelligent Virtual Agents (IVA), Boston, MA, AUG 27-29, 2014.

[22] Sanjiv Erat and Uri Gneezy. 2012. White Lies. *Management Science* 58, 4 (2012), 723–733. http://www.jstor.org/stable/41432792

[23] Denise Y. Geiskkovitch and James Everett Young. 2023. Trust Calibration Through Intentional Errors: Designing Robot Errors to Decrease Children's Trust Towards Robots. *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (2023), 1402–1406. https://api.semanticscholar.org/CorpusID:265190143

[24] Cory Kidd and Cynthia Breazeal. 2004. Effect of a robot on user perceptions. *2004 IEEE/RSJ International conference on intelligent robots and systems* 4, 3559 – 3564 vol.4. https://doi.org/10.1109/IROS.2004.1389967

[25] Namyeon Lee, Jeonghun Kim, Eunji Kim, and Ohbyung Kwon. 2017. The Influence of Politeness Behavior on User Compliance with Social Robots in a Healthcare Service Setting. *International Journal of Social Robotics* 9 (11 2017). https://doi.org/10.1007/s12369-017-0420-0

[26] G. Maggi, E. Dell'Aquila, I. Cucciniello, and S. Rossi. 2021. "Don't Get Distracted!": The Role of Social Robots' Interaction Style on Users' Cognitive Performance, Acceptance, and Non-Compliant Behavior. *International Journal of Social Robotics* 13 (2021). https://doi.org/10.1007/s12369-020-00702-4

[27] Maja J Matarić. 2002. Situated robotics. *Encyclopedia of cognitive science* 4 (2002), 25–30.

[28] Andreas Matthias. 2015. Robot Lies in Health Care: When Is Deception Morally Permissible? *KENNEDY INSTITUTE OF ETHICS JOURNAL* 25, 2 (2015), 169–192. https://doi.org/10.1353/ken.2015.0007

[29] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 339 (2009). https://doi.org/10.1136/bmj.b2535 arXiv:https://www.bmj.com/content/339/bmj.b2535.full.pdf

[30] George Mois and Jenay M. Beer. 2020. Chapter 3 - Robotics to support aging in place. In *Living with Robots*, Richard Pak, Ewart J. de Visser, and Ericka Rovira (Eds.). Academic Press, 49–74. https://doi.org/10.1016/B978-0-12-815367-3.00003-7

[31] Byeong June Moon, JongSuk Choi, and Sonya S. Kwak. 2021. "Pretending to be Okay in a Sad Voice": Social Robot's Usage of Verbal and Nonverbal Cue Combination and its Effect on Human Empathy and Behavior Inducement. In *2021 IEEE/RSJ INTERNATIONAL CONFERENCE ON INTELLIGENT ROBOTS AND SYSTEMS (IROS) (IEEE International Conference on Intelligent Robots and Systems)*. IEEE; RSJ, 854–861. https://doi.org/10.1109/IROS51168.2021.9636709 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), ELECTR NETWORK, SEP 27-OCT 01, 2021.

[32] Kantwon Rogers, Reiden John Allen Webber, and Ayanna Howard. 2023. Lying About Lying: Examining Trust Repair Strategies After Robot Deception in a High-Stakes HRI Scenario. In *COMPANION OF THE ACM/IEEE INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION, HRI 2023*. IEEE; Assoc Comp Machinery; Honda Res Inst Japan; Toyota Res Inst; Amazon; Furhat Robot; LuxAI; Diligent Robot; Navel Robot; Google; PAL Robot; Sci Robot; IEEE Robot & Automat Soc; SIGCHI; ACM SIGCAI, 706–710. https://doi.org/10.1145/3568294.3580178 18th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI), Stockholm, SWEDEN, MAR 13-16, 2023.

[33] Andres Rosero. 2023. "Using Justifications to Mitigate Loss in Human Trust when Robots Perform Norm - Violating and Deceptive Behaviors. In *COMPANION OF THE ACM/IEEE INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION, HRI 2023*. IEEE; Assoc Comp Machinery; Honda Res Inst Japan; Toyota Res Inst; Amazon; Furhat Robot; LuxAI; Diligent Robot; Navel Robot; Google; PAL Robot; Sci Robot; IEEE Robot & Automat Soc; SIGCHI; ACM SIGCAI, 766–768. https://doi.org/10.1145/3568294.3579979 18th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI), Stockholm, SWEDEN, MAR 13-16, 2023.

[34] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. 2017. How the Timing and Magnitude of Robot Errors Influence Peoples' Trust of Robots in an Emergency Scenario. In *Social Robotics*, Abderrahmane Kheddar, Eiichi Yoshida, Shuzhi Sam Ge, Kenji Suzuki, John-John Cabibihan, Friederike Eyssel, and Hongsheng He (Eds.). Springer International Publishing, Cham, 42–52.

[35] Alessandra Rossi and Silvia Rossi. 2023. Evaluating People's Perception of Trust of a Deceptive Robot with Theory of Mind in an Assistive Gaming Scenario. In *2023 32ND IEEE INTERNATIONAL CONFERENCE ON ROBOT AND HUMAN INTERACTIVE COMMUNICATION, RO-MAN (IEEE RO-MAN)*. IEEE, 1375–1380. https://doi.org/10.1109/RO-MAN57019.2023.10309647 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Busan, SOUTH KOREA, AUG 28-31, 2023.

[36] Henrik Skaug Sætra. 2023. Machiavelli for robots: Strategic robot failure, deception, and trust. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1381–1388. https://doi.org/10.1109/RO-MAN57019.2023.10309455

[37] Kristin Schaefer. 2016. *Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI"*. 191–218. https://doi.org/10.1007/978-1-4899-7668-0_10

[38] Rick Schifferstein and Elly Zwartkruis-Pelgrim. 2008. Consumer-Product Attachment: Measurement and Design Implications. *International Journal of Design* 2 (12 2008).

[39] Amanda Sharkey. 2016. Should we welcome robot teachers? *Ethics and Information Technology* 18 (12 2016). https://doi.org/10.1007/s10676-016-9387-z

[40] Amanda Sharkey and Noel Sharkey. 2012. Granny and the Robots: Ethical Issues in Robot Care for the Elderly. *Ethics and Information Technology* 14, 1 (2012), 27–40. https://doi.org/10.1007/s10676-010-9234-6

[41] Amanda Sharkey and Noel Sharkey. 2021. We need to talk about deception in social robotics! *ETHICS AND INFORMATION TECHNOLOGY* 23, 3 (2021), 309–316. https://doi.org/10.1007/s10676-020-09573-9

[42] Jaeeun Shim and Ronald C. Arkin. 2013. A Taxonomy of Robot Deception and its Benefits in HRI. In *2013 IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN, AND CYBERNETICS (SMC 2013) (IEEE International Conference on Systems Man and Cybernetics Conference Proceedings)*. IEEE; IEEE Comp Soc, 2328–2335. https://doi.org/10.1109/SMC.2013.398 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Manchester, ENGLAND, OCT 13-16, 2013.

[43] Jaeeun Shim and Ronald C. Arkin. 2014. Other-oriented robot deception: A computational approach for deceptive action generation to benefit the mark. In *2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014)*. 528–535. https://doi.org/10.1109/ROBIO.2014.7090385

[44] Jaeeun Shim and Ronald C. Arkin. 2015. The Benefits of Robot Deception in Search and Rescue: Computational Approach for Deceptive Action Selection via Case-based Reasoning. In *2015 IEEE INTERNATIONAL SYMPOSIUM ON SAFETY, SECURITY, AND RESCUE ROBOTICS (SSRR)*. IEEE; Robotics & Automation Soc. IEEE International Symposium on Safety, Security, and Rescue Robotics, West Lafayette, IA, OCT 18-20, 2015.

[45] Jaeeun Shim and Ronald C. Arkin. 2016. Other-Oriented Robot Deception: How Can a Robot's Deceptive Feedback Help Humans in HRI?. In *SOCIAL ROBOTICS, (ICSR 2016) (Lecture Notes in Artificial Intelligence, Vol. 9979)*, A Agah, JJ Cabibihan, AM Howard, MA Salichs, and H He (Eds.). SoftBank Robot; Univ Kansas, Sch Engn; Springer, 222–232. https://doi.org/10.1007/978-3-319-47437-3_22 8th International Conference on Social Robotics (ICSR), Kansas City, MO, NOV 01-03, 2016.

[46] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. 2010. No fair!! An interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 219–226. https://doi.org/10.1109/HRI.2010.5453193

[47] Henrik Skaug Sætra. 2021. Social robot deception and the culture of trust. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 276–286. https://doi.org/doi:10.1515/pjbr-2021-0021

[48] Adriana Tapus, Maja Mataric, and Brian Scassellati. 2007. Socially assistive robotics [Grand Challenges of Robotics]. *Robotics & Automation Magazine, IEEE* 14 (04 2007), 35 – 42. https://doi.org/10.1109/MRA.2007.339605

[49] Kazunori Terada and Akira Ito. 2010. Can a robot deceive humans? *5th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2010*, 191–192. https://doi.org/10.1145/1734454.1734538

[50] Anouk van Maris, Nancy Zook, Praminda Caleb-Solly, Matthew Studley, Alan Winfield, and Sanja Dogramadzi. 2020. Designing Ethical Social Robots-A Longitudinal Field Study With Older Adults. *FRONTIERS IN ROBOTICS AND AI* 7 (2020). https://doi.org/10.3389/frobt.2020.00001

[51] Anouk van Maris, Nancy Zook, Sanja Dogramadzi, Matthew Studley, Alan Winfield, and Praminda Caleb-Solly. 2021. A New Perspective on Robot Ethics through Investigating Human-Robot Interactions with Older Adults. *APPLIED SCIENCES-BASEL* 11, 21 (2021). https://doi.org/10.3390/app112110136

[52] Marynel Vazquez, Alexander May, Aaron Steinfeld, and Wei-Hsuan Chen. 2011. A Deceptive Robot Referee in a Multiplayer Gaming Environment. In *Proceedings of International Conference on Collaboration Technologies and Systems (CTS '11)*. 204 – 211.

[53] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 109–116. https://doi.org/10.1109/HRI.2016.7451741

[54] Luc Wijnen, Joost Coenen, and Beata J. Grzyb. 2017. "It's not my fault!" Investigating the Effects of the Deceptive Behaviour of a Humanoid Robot. In *COMPANION OF THE 2017 ACM/IEEE INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION (HRI'17) (ACM IEEE International Conference on Human-Robot Interaction)*. Assoc Comp Machinery; IEEE; ACM SIGCHI; ACM SIGAI; IEEE Robot & Automat Soc; AAAI; HFES,

321–322. https://doi.org/10.1145/3029798.3038300 12th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI), Vienna, AUSTRIA, MAR 06-09, 2017.

[55] Garrett Wilson, Christopher Pereyda, Nisha Raghunath, Gabriel Victor de la Cruz, Shivam Goel, Sepehr Nesaei, Bryan David Minor, Maureen Schmitter-Edgecombe, Matthew E. Taylor, and Diane Joyce Cook. 2019. Robot-enabled support of daily activities in smart home environments. *Cognitive Systems Research* 54 (2019), 258–272. https://api.semanticscholar.org/CorpusID:53431669

[56] Katie Winkle, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. 2021. Assessing and Addressing Ethical Risk from Anthropomorphism and Deception in Socially Assistive Robots. In *2021 16TH ACM/IEEE INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION, HRI*. IEEE; Assoc Comp Machinery; ACM SIGCHI; ACM SIGAI; IEEE Robot & Automat Soc, 101–109. https://doi.org/10.1145/3434073.3444666 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Boulder, CO, MAR 09-11, 2021.