Research Article

Henrik Skaug Sætra*

Social robot deception and the culture of trust

https://doi.org/10.1515/pjbr-2021-0021
received November 25, 2020; accepted April 11, 2021

Abstract: Human beings are deeply social, and both evolutionary traits and cultural constructs encourage cooperation based on trust. Social robots interject themselves in human social settings, and they can be used for deceptive purposes. Robot deception is best understood by examining the effects of deception on the recipient of deceptive actions, and I argue that the long-term consequences of robot deception should receive more attention, as it has the potential to challenge human cultures of trust and degrade the foundations of human cooperation. In conclusion: regulation, ethical conduct by producers, and raised general awareness of the issues described in this article are all required to avoid the unfavourable consequences of a general degradation of trust.

Keywords: deception, trust, culture, social robots, cooperation

1 Introduction

When social robots are introduced into human environments, they are embedded in a highly complex web of social structures and mechanisms. Social robots can be programmed to mimic basic human social norms and behaviors [1,2]. However, their social interactions are marred by a deceptive approach to the social aspect of human relations [3,4]. In this article, I examine how deception by way of social robots can disrupt and change human social structures and degrade the foundations of human cooperation. While most attention has been paid to the short-term and direct ethical implications of robot deception, I here broaden the scope of the analysis to include medium- and long-term consequences and effects beyond the particular individuals involved in situations where deception occurs.

Human beings are intensely social. Our *radically social* nature has led to human cultures based on widespread cooperation [5]. In modern western societies, the default behavior is based on a culture of *trust*.[1] Williams [7] further notes that trust is a "necessary condition of cooperative activity." Trust is a dangerous thing, however, when it is misplaced. It leaves us vulnerable to deception, as trust can be *breached*.

Human behavior is the result of the interplay between biology, evolutionary adaptions, and *culture*. Despite the differences between these foundations of behavior, they tend to pull in the same direction – toward cooperation [8]. Evolutionary traits can be analyzed both at the individual level and at the group level. At the group level, survival of strong groups has led to the evolution of traits that promote group efficiency. Strong individuals are easily conquered by groups of individuals who are individually less fit for survival but strong due to social alliances [9]. While human beings evolve, cultures evolve more rapidly [6]. Evolutionary game theory shows how traits such as trust, and other foundations of cooperative behavior, can evolve in various social settings [10]. Similarly, unfavorable settings, and environments with hostile social settings, may lead to the *degradation* of trust and cooperation [6]. How deception can lead to such degradation is the topic of this article, and the mechanisms involved are discussed in more detail after we have established an understanding of what deception is and how robots can be used for deception.

Culture can be understood as a mechanism that promotes group success and survival by promoting cooperation [8].

* **Corresponding author: Henrik Skaug Sætra,** Østfold University College, Faculty of Business, Languages and Social Sciences, NO-1757 Halden, Norway, e-mail: henrik.satra@hiof.no

1 Societal and cultural differences are, however, great, and it might be objected that the article is either limited to a very reduced set of societies, or that it is too general to be applicable to any society. I speak broadly of modern western societies and examine general human social tendencies and proclivities as described by, for example, evolutionary biologists, neuroscientists, and philosophers [6]. If there is such a thing as broad and general human proclivities, the analysis will be applicable to all, or most, human societies. However, the article is situated in a traditional western context, and most of the literature cited is also either about, or derived from, subjects in what can broadly be referred to as modern western societies.

Culture is used to describe shared patterns of, among other things, *norms*, *values*, and *attitudes* [11]. Valsiner [12] notes that culture is also a loaded term, and that it can also be seen as a way to separate *us* from *them* – a way to divide people into groups. In this article, I focus on culture as humanity's way to adapt more rapidly to new challenges than evolution would allow for [6]. Culture lets us counteract our biological and instinctual behaviors as these evolve too slow for our rapidly changing environments.

The environmental change I focus on is the introduction of new entities capable of activating human social mechanisms and eliciting social responses from us. Social robots are examples of such entities [3,13]. If social robots disrupt human social mechanisms when embedded in social settings, this is an example of how evolutionary traits and social responses can make us vulnerable, while culture and social norms may aid us in coping with the introduction of new entities in our environments. The result, however, is also that our culture and social norms need to change in order to account for these new entities and that these changes may also have effects for human social affairs. How robots can change the human culture of trust is one aspect of robot ethics that has not received much attention in previous literature.

Humans have, however, also evolved the capacity to deceive, along with many other primates [6,14–16]. What is labeled *tactical* deception requires advanced social cognition, and it requires organisms to be able to attribute minds to others and to partly simulate these minds of others [14]. Gorelik and Shackelford [17] also argue that human technologies might in turn "display cultural manifestations of deceptive" adaptions. Evolution may have rewarded deceptive behavior and thus promoted the deception and manipulation of our rivals [17,18]. However, people *prefer* not to lie and deceive, and the pleasure deriving from our own gains from lying is partially offset by the pain experienced by inflicting losses on others [19].

An intuitive and common assumption is that deception leads to a degradation of trust. However, some argue that deception may, in certain cases, *increase* trust. They show how *prosocial deception*, which we return to in Section 2.2, can lead to increased trust and increased cooperation in certain games [20]. Robot deception is similarly argued to be conducive to efficient and beneficial human–robot interactions (HRIs), and the literature on robot deception details many reasons to deceive as well as reflections on the ethical challenges associated with such deception.

I argue from an approach to agency that considers humans to be responsible for the actions of social robots [21]. Technically, a robot cannot deceive, but it can be the tool of deception, and thus the humans involved in the production and deployment of social robots are ultimately responsible for the consequences social robots have on the culture of trust. In addition to this, understanding robot deception requires us to understand the effects of deception in the victims of deceptive agents rather than the action of the deceivers [4]. Robots' effects on human culture and trust result from the effects on targets of deception, and not factoring in effects may partly stem from an unfortunate tendency to exclusively focus on the deceivers.

Responsible social robotics requires that we do not create robots that exploit human sociability to the degree that human trust, and thus sociability, will be reduced as a consequence. This could occur through both individual and group mechanisms, as (a) individuals *learn* and become less cooperative once they are deceived and (b) robot deception may degrade trust through changing human culture and evolutionary pressures.

I begin by establishing the concept of deception, before robot deception is examined in more detail. Lastly, the potential consequences of robot deception for human cooperation and culture are discussed.

2 Understanding deception

A working definition of deception is a necessary preparation for examining the consequences of *robot* deception. I first examine some of the key definitions and terms associated with deception. I then briefly establish why deception is not all bad, and that understanding *prosocial* deception is required for understanding the nature of deception.

2.1 The concept of deception

Deception is a concept that can be applied to a wide range of human activities, and a categorical denouncement of deception seems impossible in a world with a plethora of social relations, play, and fun. Deception is, nevertheless, potentially problematic for a number of reasons. Williams [7] provides two important reasons *not* to deceive: (a) it involves a breach of *trust* by the deceiver and (b) the deceiver *manipulates* the deceived. Trust is a relational concept, and a working definition entails that trust is present when one entity expects another one to do as expected. It is also often used in situations in which

the one that trust *relies on* is the trustee [7]. Levine and Schweitzer [20] distinguish between benevolence- and integrity-based trust. The first entails the reputation for goodness on the part of the trustee, while the latter entails "the belief that the trustee adheres to a set of acceptable ethical principles, such as honesty and truthfulness" [20].

As already noted, humans generally expect truth and dislike deception [22]. *Truthfulness* most naturally relates to the verbal action and communication of what the communicator knows to be factually correct [22]. As people expect to be told the truth, we become susceptible to deception. Levine [23] proposes the *truth default theory*, which emphasizes that people are inclined to believe others unless good reasons not to are provided. They *trust*, or expect, them to tell the truth. Communication is also, in general, "based on a presumption of truth" [24].

Deception is by some said to involve misleading others into believing something we do not ourselves believe in [25]. It can be achieved through "gesture, through disguise, by means of action or inaction, even through silence" [26]. It is often used to describe *behavior*, particularly when the behavior is "fraudulent, trickly, and/or misleading" [27]. Knapp et al. [27] emphasize the importance of *context* in determining what is deceptive and what is, for example, good-natured play. Deception then is defined by "the way people perceive certain features of communicative acts in context" [27].

Lying is the use of verbal or written statements in order to achieve deception, and, as such it is a subset of deception [26]. As lying, unlike deception, requires language, it is a uniquely human phenomenon, and this is why Hobbes [28] describes language as a key reason why humans are plagued by a range of conflicts not experienced by other social animals.

Some, like Vrij [29], defines the attempt to deceive as deception regardless of the success of the attempt. I will not consider deception to have occurred unless the attempt was successful. We might still speak of deceptive behavior as the attempts to achieve deception, even if we do not know, or care about, the actual outcomes. As such, a plan to deceive is not deception, but it can be deceptive.

Of equal importance is the potential requirement of *intent*, present in all the definitions of deception referred to above. According to Frank and Svetieva [30], lying is always intentional but *deception* is not. Deception is here taken as any action that misleads someone, and this, as we shall see when I turn to social robots, may easily occur without the *intention* to mislead.

What Byrne and Whiten [14] refer to as *tactical deception* involves "acts from the normal repertoire of the agent, deployed such that another individual is likely to misinterpret what the acts signify, to the advantage of the agent." While *tactical* can be taken to imply a rational conscious approach to deception, this definition does not necessitate an intent to deceive.

I emphasize the distinction between the approach of focusing only on the deceiving agent and including the effects it has on the *target of deception* [4]. This is achieved by relying on evaluations of the result of the deceptive action. By focusing on the target, the criteria of intentionality become less important. Rather than focusing on the intentions of the deceiver, the expectations and experiences of the target are what matters. As established in the introduction, human cooperation is based on initial expectations of truthfulness in most non-combative situations. Unless the context makes it reasonable to expect deceit (such as when we watch a theater play or play games based on deception), the target's expectations of nondeceit are taken as a standard basis of interaction. As already noted, this focus on the deceived is crucial for understanding some of the medium- and long-term effects of deception.

Deception only involves betrayal when there is an expectancy of truth, with the accompanying *trust* that nondeception is the norm in a given relationship. This is where deception connects with culture and social norms. For example, on a market in parts of the world, it would not be considered *betrayal* for a seller to claim that the price of a product is much higher than it actually is. Such behavior could be the normal form of interacting between seller and buyer, and the buyer is expected to know this.[2]

2.2 Prosocial deception

While deception usually comes with negative connotations, it is not all bad, as deception has a range of useful functions in human social settings. When we go to see a movie, actors pretend to be something they are not. When children play, they do the same. Many games are built upon the very premise of telling and exposing lies

2 It has been established that levels of trust and the occurrence of deception vary considerably between different cultures [6,27,31]. Triandis et al. [11] provide an account of how deception occurs more frequently in collectivist than in individualist cultures, but that there are also interesting variations at different levels of society. I will not pursue specific cultural constellations of trust and sociality in this article.

and deception. Clearly, not all deception is wrong, as Coeckelbergh [32] emphasizes as he argues that we should see deception as a neutral – or even positive – phenomenon. Furthermore, social norms often entail the suppression of true feelings and emotions or exhibiting dishonest ones. The little white lies that we all tell can be considered necessary in order to avoid conflict. Lies and deception can be *beneficial* and *prosocial*.

"White lies" can be told for the benefit of the liar alone, or for the benefit of both liar and the lied to. One example is how a manager can be overly positive when giving employee feedback, which can result in *improved* performance and better relations in the workplace [33]. Furthermore, if my partner asks me if the outfit they try on for tonight's dinner looks good, I might think that it is best for both of us if I simply answer *yes*, without endeavoring to be entirely truthful and comprehensive when relaying my thoughts.

Such lies may seem innocuous, but I argue that even these lies come at a cost. The lies will often be accepted, but both parties will most likely be aware of the game that is being played. Consequently, such lies have the consequence of eliminating one's own partner as a *real* judge of outfits, should there ever be a need for that. White lies may not reduce trust in a person *in general*, but they *will* reduce trust in particular areas.

3 Social robots and robot deception

Trust is essential for human–human cooperation and is also essential for human–robot (and human-automation) cooperation [34,35]. Trust is also intimately connected to *deception*, as it "plays an important, but often overlooked, role in the development and maintenance of trust" in machines [36].

Social robots are used for various purposes, such as entertainment and companionship, therapeutic purposes, and even love and intimacy [37–39]. Some parts of the deception that I discuss in the following might also be applicable to robots that are *not* particularly advanced, due to our tendency to anthropomorphize technology [40–42]. I focus on *social robots* in particular, as *social* robots are most problematic due to their explicit aim of interacting with human beings and thus taking advantage of the various human instincts and norms of social behavior [3,32]. Anthropomorphism is discussed in more detail in Section 3.4.

3.1 Attribution of deception

First, a key question: can robots *own actions*? The question of robot agency is widely debated, and Bryson [43] and Bryson et al. [44] argue that giving robots agency – even in the form of limited *legal* personhood – is problematic. In the following, I assume that robots *cannot* own their own actions and that some human being is always responsible for the acts of a robot [21].

In this framework, a robot that makes a deceptive statement is not considered the source of this statement – some human is held responsible. A robot *may* be an artificial person, and an actor, but it is not considered a *natural* person. A natural person is always considered the *author* of the actions of an artificial person [21,28], the producer, the distributor, or the owner who deploys it. As such, robot deception is more properly understood as *human deception by way of robots*. The sophistication of modern machines – a *veil of complexity* [21] – may obscure the source of responsibility, but it does not fundamentally change how responsibility should be attributed. Even if the future actions of a machine are fully unknown, and unknowable, to the human responsible, they are responsible, as they have designed the machine in a way that makes it unpredictable [21].

In some respects, social robots can be said to be *sycophantic* [3]. A sycophant is someone who acts in a certain way in order to curry favor and not because of some inherent desire to act the way they do. Social robots are often designed and programmed to please and promote certain behavior in the humans they encounter. Not on the basis of some internal desire in the social robot but in order to achieve some predetermined goal created by its designer.

Robot deception thus refers to deception *by* robot, and robots are merely the vessel of deception. They are akin to advanced costumes worn by a designer in a masquerade. This implies that when someone refers to someone being deceived by a robot, the one deceiving is actually a human being. Proper attribution of responsibility is required both for purposes of legislation and regulation and for making the general public aware of what robots are, and are not, capable of. The latter point relates to a potential avenue for facing the threats described here and will be further discussed in the conclusion.

3.2 Why designers deceive

Facilitating effective interaction between humans and robots is one of the primary objectives of the field of

HRI [35,45–47]. We have already seen how prosocial deception can function as lubrication in human–human interactions (HHIs), and deception is perhaps even more important as a lubricant in HRI, if the goal is to make this interaction more humanlike. Many potential benefits are to be reaped from robot deception, and these depend on how we define benefits and whose benefits we focus on [48–50]. The argument of the article is not that all robot deceptions must be eliminated, but that in order to understand how to regulate and face this phenomenon, we need to understand the medium- to long-term societal consequences of such deception. For example, Arkin et al. [46] suggest that a deceptive robot may be better at inducing a patient with Alzheimer's to accept what is deemed a proper treatment. They go on to use *interdependence theory* to decide when a robot should deceive and state that "deception is most warranted when the situation is one of greatest interdependence and greatest conflict" [46].

One prime example of a cooperative mechanism in human and animal social settings is that of *joint attention*, which is dependent on using gaze as a method of communication [8]. While a robot does not need eyes that mimic human functionality in terms of visibly pointing to objects of attention, much research is devoted to how robots can perform social eye gaze [45]. A robot need not use its "eyes" for perceiving its visual environment, but it can still simulate social eye gaze functions, such as *eye contact*, *referential gaze*, *joint attention*, and *gaze aversion* [45]. If these functions are merely superficial attributes implemented to elicit a certain response from humans, they may be deceptive. Eye contact, or mutual gaze, also increases the levels of trust between humans; and studies show that imitating human features and encouraging anthropomorphism in general are beneficial for human–machine trust [13,51]. This makes the intentional design of such features highly relevant to issues of trust and deception. Anthropomorphism is discussed in more detail in Section 3.4.

Another example is how robots that do not disclose their true nature to human opponents are able to achieve higher levels of cooperation in prisoner dilemma games than more honest machines [52]. Short et al. [53] also show that making a robot *cheat* in a game actually increases social engagement and encourages anthropomorphism. Furthermore, Almeshekah [48] shows how deception can be used to enhance the security of computer systems, including robots. While human trust can be increased by deceptive measures, *too much* trust has also proven to be potentially problematic. Exaggerated trust in robots – often referred to as the problem of overtrust – may, for example, lead to dangerous situations as people trust a robot giving poor advice rather than their own judgments [54,55].

Coeckelbergh [56] notes that it might be beneficial for robots to adhere to certain *ideals* of nondeception, but that, in general, classical human ideals may need to be negotiated to facilitate effective HRI. While fully acknowledging the potential benefits of robot deception in order to create better user experiences and facilitate cooperation, my emphasis is on the potential long-term and broader consequences of robot deception for human trust and cooperation.

3.3 Typologies of robot deception

A fundamental premise in this article is that robots can be used for deception. This is also widely accepted in the field of HRI, where robot deception is acknowledged, and the focus is more on when and how a robot can be used for deception than *if* it can be deceptive [36,57]. While Arkin et al. [46] consider all "false communication that tends to benefit the communicator" as deception, there is a need to distinguish between various forms of deception.

3.3.1 External, superficial, and hidden state deception

Danaher [58] examines robot deception and divides it into three forms: *external state deception*, *superficial state deception*, and *hidden state deception*. When a robot tells a *lie*, it deceives you about something external to the robot itself. It may, for example, tell you that it will most likely rain tomorrow, even if it knows that the weather will most likely be fine. This is *external state deception*, and an example would be a lying robot [59,60], regardless of whether the lies are white or the deception prosocial.

When a robot emits signals that imply that it has capacities or characteristics it does *not* have, we have *superficial state deception*. This could be the case if the robot was, for example, programmed to appear sad when it delivered bad news to a human. This might be perceived as the presence of some form of empathy, even if there is no trace of empathy or sadness to be found in the robot. This kind of deception might be crucial for facilitating efficient HRI, but it is nevertheless deceptive. Another example would be simulated social eye gaze. Eckel and Wilson [61] show how facial expressions have clear effects on perceived trustworthiness. This is an example of how social robots may easily be designed to take advantage of subconscious social mechanisms in

human beings. Such signals have evolved to promote social efficiency. Human nonverbal actions are sources of "leakage" or "deception cues" [62]. Robots can be designed to (a) not display such cues or (b) exploit such mechanisms in order to deceive even more effectively [2]. The later could involve, for example, a poker playing robot giving false tells, implying that it is bluffing when it is not. When social robots use and exploit the potential of such signals, the signals' value for human beings are reduced, and responding to them in traditional ways becomes potentially harmful.

The final form of deception involves acting in order to *conceal* some characteristics that the robot actually has. Kaminski et al. [63] use the example of a robot that turns its head away from you – leading you to think that it cannot "see" you – while it has sensors and eyes that can record at any angle. When this occurs, the robot acts a certain way to hide its true capabilities, and this is *hidden state deception*.

Danaher [58] argues that superficial and hidden state deception are different in important ways, and that hidden state deception is the most worrisome and constitutes what he calls *betrayal*. He also argues that the two forms are often conflated and that this conflation makes us miss important nuances of robot deception. It is important to separate them, he argues, because superficial state deception is usually not problematic, and actually not deception at all. He bases such a view on the theory of *ethical behaviorism*, which implies that simulated feelings are in a certain respect equal to genuine feelings, and that we should focus on the outward states of actors since we do not have access to the inner minds of humans or machines. This is contrary to the approach of, for example, Turkle [64], who portrays human feelings and relationships as authentic, and the robot equivalents as somehow inauthentic or fake. As noted by Coeckelbergh [56], however, it may be more fruitful to examine what constitutes appropriate human–machine relationships than to focus on vague notions of authenticity.

If a robot fakes empathy, it can be perceived as *actually* empathizing, according to Danaher's [58] ethical theory. If we disregard the inner states of entities, robot deception is similar to acts of courtesy between human beings, for example, and he argues that there is no reason we should treat such behavior from robots as different to that from humans.

Some have argued that focusing on the relational aspects of HRI, rather than what we and robots *are*, allows us to better understand the mechanisms at play and ethical implications of these interactions [32,65,66]. This approach is well suited for analyzing how robots trigger and enjoy human social mechanisms and can thus help show the breadth and depth of the consequences of robot deception.

3.3.1.1 The problem of "If not A, then B"

However, there is a conceptual difficulty involved in Danaher's separation between the two forms. My objection is that actively signaling the presence of characteristics one does *not* have involves concealing the characteristics one *actually* has. When a robot fakes empathy, it conceals the lack of it. If a robot is programmed not to respond to questions immediately but appears to ponder such questions – perhaps holding its chin with fake puzzlement – it hides its true capacity for instantaneous reasoning. If it fakes physical pain, it conceals its physical nature and immunity to the same.

Superficial states and hidden states are connected, and people necessarily draw inferences about hidden states from the superficial ones. The conflation of the two is bad for certain analytical purposes, but when it comes to deception, separating them may not make things as clear as Danaher hopes. I argue that both forms are *deception*, and I also argue that both forms may be conceived of as *betrayal*.

Danaher approaches the problem from the perspective of ethical behaviorism and a focus on the deceiving agent. Arkin et al. [46] similarly note that they "focus on the actions, beliefs, and communication of the deceiver, not the deceived," without justifying this choice. I have argued that focusing on the one *deceived* is of more interest when analyzing the ethical implications of robot deception. What a robot *actually* is, and does, is of less interest than how robot deception influences human behavior and culture.

3.3.2 Full and partial deception

Focusing on the recipient of deceptive acts, I introduce two forms of deception connected to Danaher's superficial and hidden state deception. I label these *full* deception and *partial* deception. These two categories aim at describing the results that occur in human beings who encounter social robots.

When a person believes that a social robot is not a robot, *full deception* has occurred. The person fully believes, at both a conscious and a subconscious level, that the machine is not a machine, but a human being,

an animal, or something else distinctively different from its true nature. The machine passes the Turing test of the human involved [67]. The difficulty in achieving full deception depends as much, or more, on the individual being deceived as on the deceptive actor. As such, context matters, and the result in the individual at the receiving end of deception is of great importance.

The second form of deception is the one I am most interested in, and this occurs when a human being has a rational appreciation of the nature of the device it interacts with but at a subconscious level cannot help reacting to it as if it is real. *Partial deception* occurs when social robots elicit emotional responses as if they were alive, even if we know that they are not. Turkle [64] and Darling [68] discuss this phenomenon in detail. This is similar to the effects on behavior we see from various social cues from nonlive sources, such as the decrease in antisocial behavior in the presence of a poster with a pair of eyes [69].

Partial deception is related to the idea of the "the willing suspension of disbelief" [70] and *self-deception*. However, our subconscious responses are, per definition, not voluntary, and thus deception at this level cannot be categorized as fully encompassed by the term *self-deception*.

3.4 The problem of anthropomorphism

Human beings are "compulsive meaning makers" [12]. We enter the world, and we fill out the blanks, so to speak, as we only perceive what our senses provide us with, but the blanks are filled with *our* creations. As gestalt psychology has demonstrated in a number of fascinating ways, we humans have a way of finding meaning and completeness in the incomplete. The relevance of such considerations is that the designers of social robots create somewhat incomplete entities, with suggestions of similarities to human beings, animals, etc. When people encounter these entities, they happily fill out the blanks, but the blanks may be filled with deeply human characteristics, even if the entities in question are little more than a plastic shell with an engine inside [71].

Darling et al. [72], for example, show how human attitudes toward robots are influenced by encouraging anthropomorphism through providing robots with *stories* and lifelike movement. On the other hand, anthropomorphism may lead to an exaggerated belief in the capabilities of artificial intelligence on the part of humans, and it may also hinder the proper evaluation of the state of robot development, as human beings do not objectively evaluate robots for what they really are – instead, they "find intelligence almost everywhere" [71].

Encouraging anthropomorphism can be both superficial and hidden state deception, and the benefits of making humans see robots as more – or less – than they really are include, but are not limited to, facilitating effective HRI, using robots to test human psychological and social mechanisms [71,73]. This leads some to argue that designers "should consider the incorporation of human-like features as a deliberate design choice" [13], while others argue that designers must take steps to account for both intended and unintended deception, or "nonintended illusions" [32].

One objection to the argument presented in this article is that deception by robot is unproblematic and the result of anthropomorphism. As such, robot deception is akin to the people deceived by the shapes of Heider and Simmel [74] or pretty much *any* other device, as shown in the classical book by Reeves and Nass [40].

The fact that *other* devices can also be deceptive is no excuse for intentionally, or neglectfully, causing such deception by robot. A key ethical aspect related to robot deception is thus how to signal *actual* capabilities and to *counteract* human tendencies to anthropomorphize technology.

4 Discussion

We have seen that human beings are deeply social and cooperative and that trust and deception are central to understanding our social nature. We have also seen how robots are highly effective vessels of deception, and the main argument proposed in this article is that the medium- and long-term consequences of such deception on a cultural and social scale must be considered.

Much attention has previously been devoted to the benefits of robot deception for HRI, and the ethics community has examined many of the problematic aspects of deception. However, most of the latter has involved narrower limited short-term, direct, or individual consequences of robot deception, while I have here argued that there are medium- to long-term consequences on a cultural level that might also be important.

The argument is that in societies built on trust and the expectancy of truthful signals of both superficial and hidden states, repeated deception will erode this trust and change the culture and social norms. This, I argue, is one reason why robot deception is problematic in terms of cultural sustainability. The degree to which trust is affected by robot

deception must be subjected to more thorough philosophical *and* empirical research, and this must involve examinations beyond those focused on individuals or dyads. Two mechanisms that support the notion that robot deception may be detrimental to human culture of cooperation and trust has emerged in this article. The first involves partly well-known short-term effects related to individual learning while the other details long-term group-level dynamics related to culture and evolution.

4.1 Fool me once…

First, we can assume that deception will have consequences on the individuals being deceived, if we simply assume that individuals *learn*. When people trust, and are betrayed, they will be less likely to trust again. This is a rational individualist approach to trust, and culture is affected by the level of trust exhibited by its individual members. Such changes can occur rapidly, even within a generation.

As noted by Williams [7], long-term consideration about the favorable consequences of being trustworthy inclines people against deception. If a person deceives, he may not be trusted again. We are inclined to *assume* truthfulness in our initial meetings with others in neutral or familiar setting [23]. Axelrod and Hamilton [75] famously showed how applying the simple strategy of *tit for tat* could be responsible for the evolution of cooperation. This involves *starting* with trust but then subsequently punishing any violations of that trust with noncooperation. This reciprocity-oriented version of trust, however, often requires a high degree of transparency, and this has important implications for the design of social robots intended to foster, rather than degrade, trust.

The general idea is that people *learn* and that there are strong mechanisms of *reinforcement* involved in human social settings. Trust and cooperation are fundamental for human society, and breaches of trust are not only rationally remembered but accompanied by strong biological mechanisms which make sure that such transgressions are not easily forgotten [6].

4.2 Cultural devolution

> When established institutions become unreliable or corrupt, trust is withdrawn, with suspicion of strangers, familiars, and even family members becoming the standard. [6]

Trust and cooperation are, as we have seen, dependent on certain conditions of transparency and punishment, and it is built gradually over time. The *culture of trust* describes cultures in which the just mentioned institutions exist and are relatively stable. This culture can, however, quickly be broken. If social robots make institutionalized cooperation *unreliable* or *corrupt*, the consequence can be cultural change. It is important to note that the withdrawal of trust described by Churchland [6] is *generalized*, and not necessarily restricted to the causes of the corruption of institutions of cooperation. This means that HRI will have spillover effects to HHI. Exploiting trust in human–robot relationships might thus lead to less trust between *humans*, and examining the relationship between trust in HRI and HHI is thus an avenue for future research.

Second, when deception is based on exploiting social proclivities, individuals less sensitive to social cues will be less vulnerable to deception, and thus evolutionary dynamics may lead to a general degradation of sociality based on trust. As evolutionary pressures have selected *for* cooperation and trust, an increased tendency to deceive, for example by robot, may change such pressures and reward *distrust* and noncooperative behavior [6].

This mechanism may have effects even if individuals are *not* aware of the deception that takes place, and even partial deception may thus lead to a changing culture of trust. Such change, however, is much slower than the first mechanism described.

5 Conclusion

Since I argue from an approach to agency that considers *human beings* as responsible for the actions of the social robots we know today, the *producers* of social robots become the target of my examination. A robot cannot deceive, but it can be the tool of deception, and thus the humans involved in the production and deployment of social robots are the ones responsible for the consequences of social robots on the level of trust in our cultures. The individual and group mechanisms described in the previous section deserve more attention in the literature on robot deception, and the consequences of robot deception on human trust and human culture must be studied both theoretically and empirically. Much attention has been paid to important theoretical ethical considerations and the direct results of robot deception, but this article has highlighted another important facet of such deception. Combining these is required for attaining a complete understanding of the societal implications of robot deception.

Two approaches are often proposed for solving the issues discussed here: (a) regulation and (b) ethical producers. Regulating the social nature of robots is a difficult task, yet the potential consequences of robot deception necessitate increased and sustained efforts both to understand and regulate such deception. As shown in this article, designers of social robots have many incentives to produce robots that are deceptive, which implies that self-regulation by the industry alone will tend to be insufficient.

There is, however, one other avenue to be pursued for responsible robotics – one in which both the academic community and media is crucial. This is a bottom-up approach to responsible robotics, which implies that we must first seek an understanding of the both short- and long-term consequences of deceptive robots and then educate people on these effects. Human beings must be armed against the potential danger, and increased awareness has two effects: (a) it decreases the chances of being deceived somewhat, as one may actively counter it (even the partial deception) and (b) it will affect the demand for such robots. As self-regulation seems utopian, and regulation of anthropomorphic features is exceedingly difficult, and might never become perfect, social pedagogy aimed at disarming social robots of their deceptive arsenal seems to be both a necessary and valuable pursuit that has received too little attention thus far.

Funding information: The author states no funding involved.

Author contributions: The author has accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The author states no conflict of interest.

Data availability statement: Data sharing is not applicable to this article as no data sets were generated or analyzed during the current study.

References

[1] J. Mumm and B. Mutlu, "Human-robot proxemics: physical and psychological distancing in human-robot interaction," in *Proceedings of the 6th International Conference on Human-Robot Interaction*, 2011, pp. 331–338.

[2] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, "Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, 2009, pp. 69–76.

[3] H. S. Sætra, "The parasitic nature of social AI: sharing minds with the mindless," *Integr. Psychol. Behav. Sci.*, vol. 54, pp. 308–326, 2020, DOI: https://doi.org/10.1007/s12124-020-09523-6.

[4] A. Sharkey and N. Sharkey, "We need to talk about deception in social robotics!," *Ethics Inf. Technol.*, 2020, DOI: https://doi.org/10.1007/s10676-020-09573-9.

[5] T. Yıldız, "Human-computer interaction problem in learning: could the key be hidden somewhere between social interaction and development of tools?" *Integr. Psychol. Behav. Sci.*, vol. 53, no. 3, pp. 541–557, 2019.

[6] P. S. Churchland, *Braintrust: What Neuroscience Tells Us About Morality*, Princeton: Princeton University Press, 2011.

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

[8] M. Tomasello, *Why We Cooperate*, MIT Press, Cambridge, 2009.

[9] R. Kurzban, "Biological foundations of reciprocity," in *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research*, E. Ostrom, and J. Walker, Eds., Russel Sage Foundation, New York, 2003, pp. 105–127.

[10] M. A. Nowak, *Evolutionary Dynamics: Exploring the Equations of Life*, Harvard University Press, Cambridge, 2006.

[11] H. C. Triandis, P. Carnevale, M. Gelfand, C. Robert, S. A. Wasti, and T. Probst, et al., "Culture and deception in business negotiations: A multilevel analysis," *Int. J. Cross Cult. Manag.*, vol. 1, no. 1, pp. 73–90, 2001.

[12] J. Valsiner, *An Invitation to Cultural Psychology*, SAGE Publications Ltd., Los Angeles, 2014.

[13] E. J. de Visser, S. S. Monfort, R. McKendrick, M. A. B. Smith, P. E. McKnight, F. Krueger, and R. Parasuraman, "Almost human: Anthropomorphism increases trust resilience in cognitive agents," *J. Exp. Psychol. Appl.*, vol. 22, no. 3, pp. 331–349, 2016.

[14] R. W. Byrne and A. Whiten, "Cognitive evolution in primates: evidence from tactical deception," *Man*, vol. 27 pp. 609–627, 1992.

[15] K. Hall and S. F. Brosnan, "Cooperation and deception in primates," *Infant Behav. Dev.*, vol. 48, pp. 38–44, 2017.

[16] R. W. Mitchell, "A framework for discussing deception," in *Deception: Perspectives on Human and Nonhuman Deceit*, R. W. Mitchell and N. S. Thompson, Eds., Suny Press, New York: State University of New York Press, 1986, pp. 3–40.

[17] G. Gorelik and T. K. Shackelford, "Culture of deception," *Behav. Brain Sci.*, vol. 34, no. 1, pp. 24–25, 2011.

[18] L. McNally and A. L. Jackson, "Cooperation creates selection for tactical deception," *Proc. R. Soc. B: Biol. Sci.*, vol. 280, no. 1762, p. 20130699, 2013.

[19] U. Gneezy, "Deception: The role of consequences," *Am. Econ. Rev.*, vol. 95, no. 1, pp. 384–394, 2005.

[20] E. E. Levine and M. E. Schweitzer, "Prosocial lies: When deception breeds trust," *Organ. Behav. Hum. Decis. Process.*, vol. 126, pp. 88–106, 2015.

[21] H. S. Sætra, "Confounding complexity of machine action: a hobbesian account of machine responsibility," *Int. J. Technoethics*, vol. 12, no. 1, pp. 87–100, art. 6, 2021, DOI: https://doi.org/10.4018/IJT.20210101.oa1.

[22] P. J. Kalbfleisch and T. Docan-Morgan, "Defining truthfulness, deception, and related concepts," in *The Palgrave*

Handbook of Deceptive Communication. T. Docan-Morgan, Ed., Springer, Cham, 2019, pp. 29–39, DOI: https://doi.org/10.1007/978-3-319-96334-1_2.
[23] T. R. Levine, "Truth-default theory (TDT) a theory of human deception and deception detection," *J. Lang. Soc. Psychol.*, vol. 33, no. 4, pp. 378–392, 2014.
[24] D. B. Buller and J. K. Burgoon, "Interpersonal deception theory," *Commun. Theory*, vol. 6, no. 3, pp. 203–242, 1996.
[25] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," *Psychol. Bull.*, vol. 129, no. 1, p. 74, 2003.
[26] S. Bok, *Lying: Moral Choice in Public and Private Life*, Vintage Books, New York, 1979.
[27] M. L. Knapp, M. S. McGlone, D. L. Griffin, and B. Earnest, *Lying and Deception in Human Interaction*, Kendall Hunt Publishing, Dubuque, 2015.
[28] T. Hobbes, *Leviathan*, London: Basil Blackwell, 1651.
[29] A. Vrij, *Detecting Lies and Deceit: The Psychology of Lying and Implications for Professional Practice*, Wiley, Chichester, 2000.
[30] M. G. Frank and E. Svetieva, "Deception," in *Nonverbal Communication: Science and Applications*, D. Matsumoto, M. G. Frank, and H. S. Hwang, Eds., Sage Publications, Los Angeles, 2013, pp. 121–144.
[31] J. P. Henrich, R. Boyd, S. Bowles, E. Fehr, C. Camerer, and H. Gintis, *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, Oxford University Press, Oxford, 2004.
[32] M. Coeckelbergh, "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 Inf. Technol.*, vol. 20, no. 2, pp. 71–85, 2018.
[33] S. Erat and U. Gneezy, "White lies," *Manag. Sci.*, vol. 58, no. 4, pp. 723–733, 2012.
[34] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Hum. Factors*, vol. 57, no. 3, pp. 407–434, 2015.
[35] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Hum. Factors*, vol. 53, no. 5, pp. 517–527, 2011.
[36] P. A. Hancock, D. R. Billings, and K. E. Schaefer, "Can you trust your robot?" *Ergon. Des.*, vol. 19, no. 3, pp. 24–29, 2011.
[37] H. S. Sætra, *First, they came for the old and demented: Care and relations in the age of artificial intelligence and social robots*. Human Arenas, 2019, DOI: 10.2139/ssrn.3494304.
[38] M. Scheutz and T. Arnold, "Are we ready for sex robots?," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*), IEEE, 2016, pp. 351–358.
[39] D. Levy, *Love and Sex with Robots: The Evolution of Human-Robot Relationships*, Harper Perennial, New York, 2009.
[40] B. Reeves and C. I. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, Cambridge University Press, Cambridge, 1996.
[41] A. Sharkey and N. Sharkey, "Children, the elderly, and interactive robots," *IEEE Robot. Autom. Mag.*, vol. 18, no. 1, pp. 32–38, 2011.
[42] N. Sharkey and A. Sharkey, "The eldercare factory," *Gerontology*, vol. 58, no. 3, pp. 282–288, 2012.
[43] J. J. Bryson, "Patiency is not a virtue: the design of intelligent systems and systems of ethics," *Ethics Inf. Technol.*, vol. 20, no. 1, pp. 15–26, 2018.
[44] J. J. Bryson, M. E. Diamantis, and T. D. Grant, "Of, for, and by the people: the legal lacuna of synthetic persons," *Artif. Intell. Law*, vol. 25, no. 3, pp. 273–291, 2017.
[45] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: a review," *J. Human-Robot Interact.*, vol. 6, no. 1, pp. 25–63, 2017.
[46] R. C. Arkin, P. Ulam, and A. R. Wagner, "Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception," *Proc. IEEE*, vol. 100, no. 3, pp. 571–589, 2011.
[47] K. E. Oleson, D. R. Billings, V. Kocsis, J. Y. Chen, and P. A. Hancock, "Antecedents of trust in human-robot collaborations," in *2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support* (*CogSIMA*), IEEE, 2011, pp. 175–178.
[48] M. H. Almeshekah, *"Using deception to enhance security: A taxonomy, model, and novel uses,"* PhD thesis, Purdue University, 2015.
[49] J. Shim and R. C. Arkin, "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*), IEEE, 2014, pp. 528–535.
[50] J. Shim and R. C. Arkin, "A taxonomy of robot deception and its benefits in HRI," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 2013: IEEE, pp. 2328–2335.
[51] A. Normoyle, J. B. Badler, T. Fan, N. I. Badler, V. J. Cassol, and S. R. Musse, "Evaluating perceived trust from procedurally animated gaze," in *Proceedings of Motion on Games*, 2013, pp. 141–148.
[52] F. Ishowo-Oloko, J. -F. Bonnefon, Z. Soroye, J. Crandall, I. Rahwan, and T. Rahwan, "Behavioural evidence for a transparency-efficiency tradeoff in human-machine cooperation," *Nat. Mach. Intell.*, vol. 1, no. 11, pp. 517–521, 2019.
[53] E. Short, J. Hart, M. Vu, and B. Scassellati, "No fair!! An interaction with a cheating robot," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*), IEEE, 2010, pp. 219–226.
[54] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*), IEEE, 2016, pp. 101–108.
[55] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*), IEEE, 2015, pp. 1–8.
[56] M. Coeckelbergh, "Are emotional robots deceptive?" *IEEE Trans. Affective Comput.*, vol. 3, no. 4, pp. 388–393, 2011.
[57] A. R. Wagner and R. C. Arkin, "Acting deceptively: Providing robots with the capacity for deception," *Int. J. Soc. Robot.*, vol. 3, no. 1, pp. 5–26, 2011.
[58] J. Danaher, "Robot betrayal: a guide to the ethics of robotic deception," *Ethics Inf. Technol.*, vol. 22, pp. 1–12, 2020, DOI: https://doi.org/10.1007/s10676-019-09520-3.

[59] O. Bendel, K. Schwegler, and B. Richards, "The LIEBOT Project," in *Machine Ethics and Machine Law*, Jagiellonian University, Cracow, 2016, pp. 8–10.
[60] O. Bendel, "Chatbots as moral and immoral machines: Implementing artefacts in machine ethics" *CHI 2019 Workshop on Conversational Agents*, Glasgow, UK, 2019.
[61] C. C. Eckel and R. K. Wilson, "The human face of game theory: Trust and reciprocity in sequential games," in *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research*, E. Ostrom, and J. Walker, Eds., Russel Sage Foundation, New York, 2003, pp. 245–274.
[62] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.
[63] M. E. Kaminski, M. Rueben, W. D. Smart, and C. M. Grimm, "Averting robot eyes," *Md. L. Rev.*, vol. 76, no. 4, pp. 983–1025, 2017.
[64] S. Turkle, *Alone Together: Why We Expect More from Technology and Less from Each Other*, Basic Books, New York, 2017.
[65] R. A. Jones, "Relationalism through social robotics," *J. Theory Soc. Behav.*, vol. 43, no. 4, pp. 405–424, 2013.
[66] D. J. Gunkel, *Robot Rights*, London: MIT Press, 2018.
[67] A. M. Turing, "Computing machinery and intelligence," in *Parsing the Turing Test*, R. Epstein, G. Roberts, and G. Beber, Eds., Springer, Netherlands, 2009, pp. 23–65.
[68] K. Darling, "'Who's Johnny?' Anthropomorphic framing in human-robot interaction, integration, and policy," in *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, P. Lin, K. Abney, and R. Jenkins, Eds., Oxford University Press, New York, 2016.
[69] K. Dear, K. Dutton, and E. Fox, "Do 'watching eyes' influence antisocial behavior? A systematic review & meta-analysis," *Evol. Hum. Behav.*, vol. 40, no. 3, pp. 269–280, 2019.
[70] M. Jacobsen, "Looking for literary space: The willing suspension of disbelief re-visited," *Res. Teach. English*, vol. 16 pp. 21–38, 1982.
[71] J. Złotowski, D. Proudfoot, K. Yogeeswaran, and C. Bartneck, "Anthropomorphism: opportunities and challenges in human-robot interaction," *Int. J. Soc. Robot.*, vol. 7, no. 3, pp. 347–360, 2015.
[72] K. Darling, P. Nandy, and C. Breazeal, "Empathic concern and the effect of stories in human-robot interaction," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication* (*RO-MAN*), IEEE, 2015, pp. 770–775.
[73] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *J. Exp. Soc. Psychol.*, vol. 52, pp. 113–117, 2014.
[74] F. Heider and M. Simmel, "An experimental study of apparent behavior," *Am. J. Psychol.*, vol. 57, no. 2, pp. 243–259, 1944.
[75] R. Axelrod and W. D. Hamilton, "The evolution of cooperation," *Science*, vol. 211, no. 4489, pp. 1390–1396, 1981.