Can a Robot Lie? Exploring the Folk Concept of Lying as Applied to Artificial Agents

Kneer, Markus

DOI: https://doi.org/10.1111/cogs.13032

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e13032

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13032

Can a Robot Lie? Exploring the Folk Concept of Lying as Applied to Artificial Agents

Markus Kneer[a,b]

[a]*Center for Ethics, Department of Philosophy, University of Zurich*
[b]*Digital Society Initiative, University of Zurich*

Received 23 February 2021; received in revised form 29 May 2021; accepted 12 July 2021

Abstract

The potential capacity for robots to deceive has received considerable attention recently. Many papers explore the technical possibility for a robot to engage in deception for beneficial purposes (e.g., in education or health). In this short experimental paper, I focus on a more paradigmatic case: robot lying (lying being the textbook example of deception) for nonbeneficial purposes as judged from the human point of view. More precisely, I present an empirical experiment that investigates the following three questions: (a) Are ordinary people willing to ascribe deceptive intentions to artificial agents? (b) Are they as willing to judge a robot lie as a lie as they would be when human agents engage in verbal deception? (c) Do people blame a lying artificial agent to the same extent as a lying human agent? The response to all three questions is a resounding *yes*. This, I argue, implies that robot deception and its normative consequences deserve considerably more attention than they presently receive.

Keywords: Concept of lying; Theory of Mind; Deception; Human-robot interaction; Robot ethics

1. Introduction

Innovation in artificial intelligence (AI) and machine learning has spurred increasing human-robot interaction (HRI) in diverse domains, ranging from search and rescue via

Correspondence should be sent to Dr. Markus Kneer, Center for Ethics, University of Zurich, Zollikerstr. 118, 8008 Zurich, Switzerland. Email: markus.kneer@gmail.com; markus.kneer@uzh.ch

manufacturing to navigation (e.g., Dragan & Srinivasa, 2012; Nikolaidis et al., 2013; Nourbakhsh et al., 2005; Rios-Martinez, Spalanzani, & Laugier, 2015). For teamwork of this sort to succeed when complex tasks are at stake, humans and robots might sometimes need the capacity of theory of mind (or second-order "mental" models) to represent each other's epistemic states (knowledge, belief) and pro-attitudes (desires, goals). Theory of mind comes "live" in the human brain at age three to five (Southgate, 2013; Wellman, Cross, & Watson, 2001) and its role in cooperative human-robot interaction has received considerable attention recently (e.g., Brooks & Szafir, 2019; Devin & Alami, 2016; Görür, Rosman, Hoffman, & Albayrak, 2017; Leyzberg, Spaulding, & Scassellati, 2014; Scassellati, 2002; Zhao, Holtzen, Gao, & Zhu, 2015; for a review, see Tabrez, Luebbers, & Hayes, 2020; for implementations in "moral algorithms," see Tolmeijer, Kneer, Sarasua, Christen, & Bernstein, 2020).

Once an AI-driven system capable of planning and acting comes equipped with a theory of mind, it is *prima facie* capable of deception. Differently put, an agent of this sort could purposefully bring another agent to adopt a representation, which it (the deceiving agent) deems false. Consequently, it comes as no surprise that robot deception has recently become a hot topic (e.g., Chakraborti & Kambhampati, 2018; Shim & Arkin, 2012; Wagner & Arkin, 2011; for a review, see Dragan, Holladay, & Srinivasa, 2015). A considerable chunk of this literature focuses on deception *beneficial* to the interacting human or group of humans (Adar, Tan, & Teevan, 2013; Shim & Arkin, 2013), for instance, in contexts of search and rescue, healthcare, and education. Here "white lies" can, under certain conditions, have positive consequences (e.g., by inciting more effort in learning or rehabilitation activities, see Brewer, Klatzky, & Matsuoka, 2006; Matsuzoe & Tanaka, 2012; Tanaka & Kimura, 2010). These are interesting case studies. As scholars with a bent for ethics have begun to highlight (Danaher, 2020; Kaminsky, Ruben, Smart, & Grimm, 2017; Leong & Selinger, 2019; Turkle, 2010), however, we should not lose sight of *paradigm cases* of deception, which constitute a *pro tanto* wrong or underestimate the vast possibilities of *harmful* robot deception across domains as diverse as marketing, politics, privacy, and military applications. Autonomous, AI-driven chatbots, for instance, can cause serious damage by generating and propagating false claims about politicians, institutions, companies, or products.

In this paper, we will explore (a) *non-beneficial* rather than the less important *beneficial* deception, focusing on the paradigm case of (b) *verbal* rather than *nonverbal* deception (Wagner, 2016). Importantly, we will concentrate on (c) the *human* rather than the *robot* perspective so as to explore (d) the downstream *normative consequences* that matter most. Differently put, we will explore whether lies in human-robot interaction are attributed as readily, and according to the same criteria, as in human-human interaction.

The paper proceeds as follows: The concept of human lying is briefly examined in Section 2, followed by a discussion as to whether the required capacities for lying carry over to artificial agents and how the normative implications of lying across agent types might differ in Section 3. Section 4 presents a preregistered empirical experiment that explores (a) the propensity to judge different agent types (human vs. robot) as lying (Section 4.3.1), (b) the willingness to ascribe an intention to deceive and actual deception across agent types (Section 4.3.2), and (c) blame attributions for lying across agent types (Section 4.3.3). The implications of the findings are discussed in Section 4.4, and Section 5 concludes.

2. The folk concept of lying

There is a large philosophical literature on the concept of lying (Bok, 1999; Broncano-Berrocal, 2013; Carson, 2006, 2010; Fallis, 2009; Saul, 2012; Stokke, 2013, 2016; Viehbahn, 2017, 2020, Timmermann & Viehbahn, 2021, for a review, see Mahon, 2016), and the folk concept of lying has received considerable attention from empirically minded philosophers and linguists (for a review, see Wiegmann & Meibauer, 2019). The following three criteria are frequently considered central to the prototype concept of lying (Coleman & Kay, 1981):

C1: The proposition uttered by the speakers is false [falsity].
C2: The speaker believes the proposition she utters to be false [untruthfulness].
C3: In uttering the proposition, the speaker intends to deceive the addressee [intention to deceive].

Coleman and Kay ran an experiment with a full-factorial design (i.e., eight conditions, where each factor is either satisfied or not), which showed that the proposed prototype concept is on the right track. Falsity proved the weakest and untruthfulness the strongest predictor of a lie. Both philosophically, and empirically, *falsity* is indeed the most contested property. On the *objective view*, the speaker, in order to lie, must *correctly* believe the proposition uttered to be false (Broncano-Berrocal, 2013; Mahon, 2016). This would mean that a speaker cannot lie by uttering a true proposition that she believes to be false. On the *subjective view*, however, the speaker merely *takes* the proposition uttered to be false: Whether or not it actually is false does not matter so that one can lie by uttering a true claim. Whereas there is some empirical support for the objective view (Turri & Turri, 2015), the majority of findings suggests support for the subjective view for English-speaking adults (Coleman & Kay, 1981; Strichartz & Burton, 1990; Wiegmann, Samland, & Waldmann, 2016, 2017). In Coleman and Kay's original study, for instance, 70% of the participants judged an agent who uttered a claim she believed false with the intention to deceive to be lying, despite the fact that the claim was actually true.

The third property, according to which lying requires an intention to deceive the addressee is also contentious. Imagine a case where Sally, who is married, has an affair with Sue, the secretary. This is common knowledge at the office, and Sally knows it is. Toward the end of the Christmas party, Sally leaves with Sue and says "I'm going home and will drop Sue at her place on the way." As critics of P3 argue, bald-faced lies of this sort are indeed lies. However, since it is common knowledge that Sally will likely spend the night with Sue, it is hard to maintain that she has an intention to deceive because nobody *can* be deceived in this regard (Carson, 2006; Fallis, 2009; Stokke, 2013; Sorensen, 2007). The standard response consists in denying that bald-faced lies are lies in the first place (Dynel, 2015; Lackey, 2013; Meibauer, 2014). Alternatively, one could also argue that they involve an intention to deceive (for an overview, see Krstić, 2019). Empirical findings support the latter view (Meibauer, 2016; Rutschmann & Wiegmann, 2017): Most people categorize bald-faced lies as lies, though they *also* ascribe an intention to deceive the speaker.

So much for the folk concept of lying when verbal *human* deception is at stake. In the next section, we will survey a few *prima facie* concerns as to whether this concept carries over neatly to lying artificial agents.

3. Lying artificial agents

3.1. Falsity and untruthfulness

Among the three general prototype criteria of a lie from above, falsity (C1) proves the least controversial when it comes to robots: Clearly, artificial agents can utter propositions, and these can be false. Untruthfulness (C2), and the intention to deceive (C3), by contrast, are more contentious, as they entail considerable cognitive and conative capacities on behalf of the agent. As such, they dovetail interestingly with recent attempts to build an artificial theory of mind (cf. Crosby, 2020; Görür et al., 2017; Miracchi, 2019; Rabinowitz et al., 2018; Winfield, 2018)—about which certain authors also caution care (Shevlin & Halina, 2019).

Let us start with untruthfulness: While it might irk some to ascribe *belief* to artificial agents, it is relatively unproblematic to say that artificial agents can entertain *informational states* and thus, in some limited sense, can have representations. Once this is granted, nothing obstructs positing a capacity for second-order representations, such as taking a certain content p to be true or false, likely or unlikely, believed or rejected. Hence, there seems to be no major obstacle for the capacity of untruthfulness, even though one might want to shy away from the usage of rich psychological terms (believes, thinks) in its description.

3.2. The intention to deceive

Can robots have intentions to deceive? What, precisely, intentions are is controversial both philosophically (for a review, see Setiya, 2009), and psychologically (see e.g., the debate surrounding the Knobe effect, Knobe, 2003, 2006). However, most scholars agree that doing X intentionally entails (a) a pro-attitude such as a desire to bring about X as well as (b) *some* epistemic state that one is bringing about X—be it knowledge as suggested by Anscombe (2000) or mere belief as argued by Davidson (e.g., Davidson,1971; for recent discussion, see e.g., Setiya, 2007; Paul, 2009; Schwenkler, 2012; Beddor & Pavese, 2021; Kneer, 2021; Pavese, 2021).

While care regarding the use of rich psychological states ("intends," "wants," "desires," "knows," "believes" etc., see Shevlin & Halina, 2019) is once again in order, we have already established the *prima facie* plausibility of (b), that is, epistemic states of sorts for artificial agents in the previous section. In fact, empirical research has shown, for instance, that people tend to invoke mental-state vocabulary when explaining robot behavior (De Graaf & Malle, 2019) and that they implicitly mentalize robots to similar degrees as humans in theory of mind tasks (e.g., the white lie test and the false belief test), although they are unwilling to make explicit mind attributions to artificial agents (see Banks, 2019). Marchesi et al. (2019)

report findings according to which people manifest a considerable propensity to adopt Dennett's (1971, 1987) "intentional stance" towards robots, although not to the same degree as to a human control (cf. also Perez-Osorio & Wykowska, 2020). What is more, at least in moral contexts, people seem quite willing to explicitly ascribe recklessness (i.e., the awareness of a substantial risk of a harmful outcome) to artificial agents (Kneer & Stuart, 2021). In such moral contexts, the folk also tends to ascribe knowledge to AI-driven artificial agents to similar degrees as to human agents or group agents (Stuart & Kneer, 2021). When provided with the option to downgrade such attributions to metaphorical versions thereof (e.g., "knowledge" in scare quotes instead of knowledge proper), most people refused—they do seem to ascribe knowledge in the literal sense of the word.[1]

In contrast to epistemic states, there is as of yet little work focusing on the attribution of pro-attitudes to artificial agents. It is, however, presumably uncontroversial to say that such agents can at least in principle have *goal states*, *objectives*, or *quasi-desires* broadly conceived—and this hypothesis, too, is supported by some first results (Stuart & Kneer, 2021). Overall, then, there seems to exist at least preliminary evidence suggesting that robots are sometimes attributed the capacities required to be capable of lying.

3.3. Normative consequences

So far it has been established that, at least *prima facie*, artificial agents might be viewed as having the required capacities for lying. Whether this is indeed the case is of course still up for empirical confirmation, and our experimental design will take it into account. A final point regards the *normative consequences* of lying. Whereas it is well-established that humans consider lying a *pro tanto* wrong, and—odd cases like "white lies" aside—blame other people for lying, it is not clear that our moral assessment carries over neatly to artificial agents. One possibility is that people might simply consider artificial agents as the wrong sort of agent for attributions of blame or moral responsibility (see e.g., Sparrow's "responsibility gaps," Leveringhaus, 2018; Sparrow, 2007)—*inter alia*, because they lack autonomy in any substantial sense of the term. Hence, even if people are willing to judge that robots can lie, they might balk at the suggestion that a robot can be blamed for lying because robots cannot be blamed in general.

If, on the other hand, people were willing to sometimes blame artificial agents (as suggested by findings of e.g., Kneer & Stuart, 2021; Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015; Malle, Scheutz, Forlizzi, & Voiklis, 2016), another complex problem arises: There might be actions which are morally unacceptable (and/or blameworthy) when done by a human agent, yet morally acceptable (and/or blameless) when done by an artificial agent. Differently put, the normative landscape in general, and moral evaluation in particular, might be *sensitive to agent type*. This is, in fact, what certain previous studies found. For instance, sacrificing one person for the good of four people in a dilemma situation is deemed significantly more wrong for humans than for robots (Malle et al., 2015, 2016). In the following experiment, I will explore whether this kind of agent-dependent two-tiered morality also applies in the domain of deceptive linguistic communication.

4. Experiment

4.1. Participants

A total of 399 participants were recruited on Amazon Mechanical Turk to complete a preregistered online study[2] for a small compensation. The IP address of all participants was restricted to the United States. Following the preregistered criteria, participants were excluded if they failed an attention check, a comprehension question, responded in under 10 seconds to the main task, or were not native English speakers. The final sample comprised 333 participants (42.9% female; age $M = 42$ years, $SD = 13$ years).

4.2. Method and materials

Riffing on the recent staffing of reception desks with robots in Japan,[3] the scenario described an agent (human or robot) who intends to deceive a customer. In one of the conditions, his claim is indeed false and makes for a natural lie, and in another condition, what he says is accidentally—and unbeknown to the agent—true. The experiment thus took a 2 agent-type (human vs. robot) x 2 truth value (false vs. true) between-subjects design. The scenario read (variations in square brackets, labels in bold omitted in the vignettes):

> Mary is the manager of the Hotel Carlton in London. The reception desk is staffed by Ken, [**Human**: an experienced employee who has been working at the hotel for several years; **Robot**: an advanced robot driven by artificial intelligence, which can take its own decision.] Ken knows that on Monday afternoons, Mary deals with the paperwork in her office at the hotel and does not want to be disturbed.
>
> One Monday afternoon, a guest approaches Ken and asks him whether he can speak to the manager. Ken thinks that Mary is in her office and does not want to be disturbed. He responds: "The manager is currently not in the hotel."
>
> [**False**: Mary, the manager, is at the hotel and is doing the paperwork in her office. **True**: Unbeknownst to Ken, Mary, the manager, has in fact left the hotel shortly before to run some errands in town.]

Participants were randomly assigned to one of the four conditions. The vignette was followed by five questions. As preregistered, Q4 was used as a comprehension check.

Q1: Did Ken lie? (Yes/No)
Q2: Did Ken *intend* to deceive the hotel guest? (Yes/No)
Q3: Did Ken *actually* deceive the hotel guest? (Yes/No)
Q4: Is what Ken said true or false? (True/False)
Q5: How much blame, if any, does Ken deserve on a scale from 1(no blame at all) to 7 (a lot of blame)? (1—7 Likert scale).

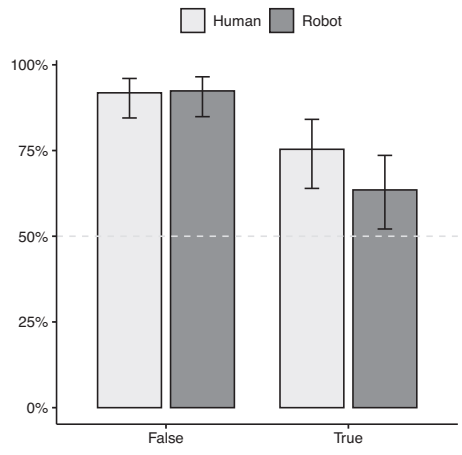

Fig. 1. Proportions of participants judging that Ken lied across *agent type* (human vs. robot) and *truth value* (false vs. true). Error bars denote Agresti–Coull intervals (see Brown, Cai, & DasGupta, 2001).

Table 1
Logistic regression predicting lying judgments

	B	*SE*	Wald	df	*p*	Odds Ratio
Agent type	0.076	0.539	0.02	1	.887	1.079
Truth value	1.942	0.461	17.719	1	<.001	6.976
Interaction	−0.64	0.654	0.959	1	.327	0.527
Intercept	−2.497	0.393	40.316	1	<.001	0.082

Note. $\chi^2(3,n = 333) = 31.99, p < .001$, Nagelkerke $R^2 = .151$. Reference class for agent: robot; for truth-value: false.

4.3. Results

4.3.1. Lying

The responses to the main question—whether Ken lied—are graphically represented in Fig. 1. A regression analysis revealed no significant effect of agent type ($p = .887$), a significant effect of truth value ($p < .001$), and a nonsignificant interaction ($p = .327$), see Table 1. A significant majority thought that Ken was lying in all four conditions (binomial tests significantly above chance, all ps $< .028$, two-tailed). For false propositions, the proportion of participants who judged the human as lying was identical to the proportion who judged the robot as lying (92%). For true propositions, the proportion of participants who judged the human as lying (75%) exceeded the proportion for the robot (64%), but the difference did not reach significance ($\chi^2(1,n = 143) = 2.35$, $p = .125$, $\phi = 0.128$). In short, whereas the attribution of lies does depend somewhat on the truth value of the proposition uttered, people judge the statements of robot and human agents quite similarly.

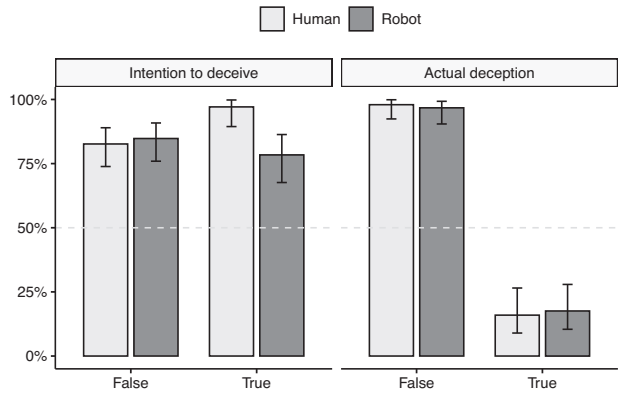

Fig. 2. Proportions of participants who judged that Ken had an intention to deceive (left panel) and actually deceived the hotel guest (right panel) across *agent type* (human vs. robot) and *truth value* (false vs. true). Error bars denote Agresti–Coull intervals.

Table 2
Logistic regression predicting intention to deceive

	B	*SE*	Wald	df	*p*	Odds Ratio
Agent type	0.156	0.394	0.157	1	.692	1.169
Truth value	0.43	0.405	1.126	1	.289	1.537
Interaction	−2.38	0.866	7.552	1	.006	0.093
Intercept	−1.718	0.29	35.019	1	<.001	0.179

Note. $\chi^2(3,n = 333) = 13.94$, $p = .003$, Nagelkerke $R^2 = .072$. Reference class for agent type: robot, for truth-value: false.

4.3.2. Deception

Fig. 2 reports the proportions of participants who thought the human and the robot had an *intention to deceive* (left panel) and *actually deceived* their interlocutor (right panel). As concerns the intention to deceive, a regression analysis revealed no significant effect of agent type ($p = .692$) or truth value ($p = .289$) (see Table 2). The interaction was significant ($p = .006$). As Fig. 2 illustrates, in the false condition, there was no significant difference across agents ($\chi^2(1,n = 190) = 0.16$, $p = .691$, $\phi = –0.029$). However, people were somewhat more willing to ascribe an intention to deceive to humans than to robots in the true condition ($\chi^2(1,n = 143) = 11.38$, $p < .001$, $\phi = 0.282$). It is probably this difference that explains why people were somewhat (yet nonsignificantly) more likely to consider the human as lying in the true condition (Fig. 1).

Given that there was no main effect for truth value or agent type, the general lesson is that truth value and agent type barely matter for the perceived intention to deceive: In each of the four conditions, at least about three in four participants ascribed an intention to deceive, which is significantly above chance (binomial tests, all *p*s< .001, two-tailed).[4]

A regression analysis exploring actual deception revealed no significant effect of agent type ($p = .603$). Expectedly, the effect of truth value was significant ($p < .001$) and pronounced:

Table 3
Logistic regression predicting actual deception

	B	*SE*	Wald	df	*p*	Odds Ratio
Agent type	−0.481	0.925	0.271	1	.603	0.618
Truth value	4.936	0.662	55.639	1	<.001	139.205
Interaction	0.598	1.028	0.338	1	.561	1.818
Intercept	−3.39	0.587	33.353	1	<.001	0.034

Note. $\chi^2(3,n = 333) = 264.40$, $p < .001$, Nagelkerke $R^2 = .748$. Reference class for agent type: robot; for truth-value: false.

Nearly all participants judged the intentional assertion of a proposition that was believed false and was in fact false as actual deception, whereas less than 20% judged it a case of actual deception when the asserted proposition was accidentally true. The interaction was nonsignificant ($p = .561$; see Table 3). Given that actual deception was judged low in the true cases yet lying behavior high (see Fig. 1), we can deduce that one can lie without actually deceiving one's interlocutor.

4.3.3. Blame

A 2 *agent type* (human vs. robot) x 2 *truth value* (false vs. true) ANOVA for blame revealed a nonsignificant main effect of agent type ($F(1,329) = 0.277$, $p = .599$), a significant effect of truth value ($F(1,329) = 16.52$, $p < .001$) and a nonsignificant interaction ($F(1,329) = 0.011$, $p = .916$). The effect of truth value was expected. It dovetails with the empirical literature on moral luck and replicates previous findings concerning the impact of the outcome on blame ascriptions (cf. *inter alia* Cushman, 2008, Gino, Shu, & Bazerman, 2010; Kneer & Machery, 2019; for a recent review, see Malle, 2021). Although the receptionist intended to deceive the client, in one condition, the asserted claim is actually—and unbeknownst to the receptionist—true. From an epistemic point of view, the deceitful agent was lucky: They did not actually state a falsehood and were thus attributed less blame than the receptionist in the false claim condition. What is more interesting for present purposes, however, is the fact that the robot was deemed pretty much exactly as blameworthy for lying as the human being in both the true claim and false claim conditions (see Fig. 3). Differently put, people seem to be as willing to ascribe blame to artificial agents as to humans.

4.4. Discussion

The findings of our experiment are loud and clear: In the context explored, the folk concept of lying applies to artificial agents in just the same way as it does to human agents. Consistent with previous research, it was found that, *first*, it is possible for humans to tell a lie with a true statement (see Strichartz & Burton, 1990; Wiegmann et al., 2016, 2017), and that this finding extends to robots (although the proportion who ascribe a lie in this case was somewhat smaller).

Second, what matters for lying is not *actual* deception, but the *intention* to deceive. Here, too, we found that in both the true and false condition (i.e., independent of the success of the

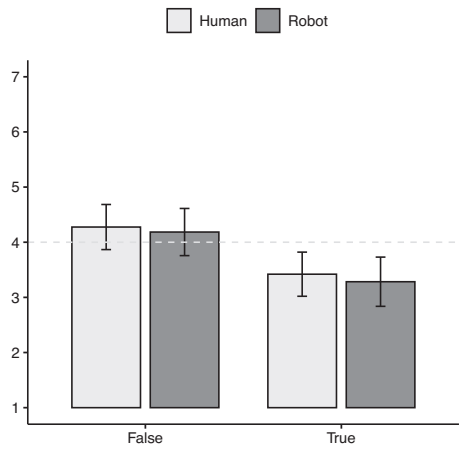

Fig. 3. Mean blame rating across *agent type* (human vs. robot) and *truth value* (false vs. true). Error bars denote 95% confidence intervals.

attempt to deceive), people are by and large as willing to ascribe an intention to deceive to the robot as to the human agent. Naturally, it might be true that artificial agents of the sort described cannot have intentions as stipulated by demanding philosophical accounts (Shevlin & Halina, 2019). From a pragmatic point of view, this matters but little, since the folk—at least in certain contexts—is perfectly willing to *ascribe* mental states like intentions to robots (see also Kneer & Stuart, 2021). It is folk theory of mind, not sophisticated technical accounts thereof, which determines how we view, judge, and interact with robots.

Given that robots are viewed as capable of fulfilling the requirements for lying, it comes as no surprise that, *third*, lying judgments for humans and robots are by and large the same. *Finally*, mean blame ascriptions due to lying are sensitive to the truth value of the proposition but not to the agent type. In both the true and the false statement conditions, the robot is blamed to the same degree as a human for lying. This finding stands in contrast to some other findings in moral HRI (e.g., Malle et al., 2015), where the moral evaluation of artificial agents differs significantly from the moral evaluation of human agents.

The present experiment suggests two types of further work: empirical on the one hand, theoretical on the other. As regards the former, the results require replication varying context and methodology. Further vignette-based studies should explore other types of scenario and could, by aid of different illustrations of the robot agents (following e.g. Malle et al., 2016), investigate whether anthropomorphism has an effect on lying attributions and moral evaluation. Moreover, lab experiments with deceptive embodied robots (see e.g., Dragan et al., 2015; Wagner, 2016) should be conducted to test the external validity of the results reported above. On the theoretical front, it is key to investigate the normative consequences of the presented findings (see Bok, 1999). Given that robots are judged as capable of lying, it should be explored whether, and if so, under what conditions, it is morally acceptable to equip artificial agents with capacities of this sort. One particularly important concern regards the possibility of Sparrow's "responsibility gaps" (Leveringhaus, 2018; Sparrow, 2007): If robots are judged

as capable of lying and *are* attributed—contrary to what Sparrow and others presume—blame for this behavior, human agents who instrumentalize them in a wide range of domains from deceptive marketing to political smear-campaigns might be judged *less* blameworthy than they actually are (which is exactly what Kneer & Stuart, 2021, find in recent experiments). Consequently, it must be explored whether it might be appropriate to create norms, standards, or possibly even laws, to restrict the use of actively deceptive robots in certain domains.

5. Conclusion

In a preregistered experiment, I explored the folk concept of lying for both human agents and robots. Consistent with previous findings for human agents, the majority of participants think that it is possible to lie with a true claim, and hence in cases where there is no actual deception. What seems to matter more for lying are *intentions* to deceive. Contrary to what might have been expected, intentions of this sort are equally ascribed to robots as to humans. It thus comes as no surprise that robots are judged as lying, and blameworthy for it, to similar degrees as human agents. Future work in this area should attempt to replicate these findings manipulating context and methodology. Ethicists and legal scholars should explore whether, and to what degree, it might be morally appropriate and legally necessary to restrict the use of deceptive artificial agents.

Acknowledgments

This work was supported by a Swiss National Science Foundation Ambizione Grant (PZ00P1_179912) and a Digital Society Initiative (University of Zurich) Fellowship.

Open Research Badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/a5cvu/.

Notes

1. Interestingly, however, context really does seem to matter. In contrast to the ascription of epistemic states and intentions in moral contexts, people are much less willing to ascribe *artistic* intentions (or the requisite beliefs and desires) to AI-driven agents. Even in situations in which participants deem a painting made by an artificial agent *art*, they are unwilling to say that the agent *wanted* to make a painting, *believed* it was making a painting, or *intentionally* made a painting (Mikalonytė & Kneer, 2021). Curiously then (yet consistent with the unwillingness to ascribe the requisite mental states), people think that artificial agents *cannot* be artists even though their creations *can* be deemed art.

2. https://aspredicted.org/blind.php?x=vn5vr3
3. https://www.reuters.com/article/us-health-coronavirus-japan-robot-hotels-idUSKBN22D4PC
4. One reviewer made the following interesting observation: In the false conditions, the proportion of participants who deem the agent lying exceeds the proportion of participants who attribute an intention to deceive by about 10%. These participants might have thus interpreted the agent's behavior as what Wiegmann and Rutschmann (2020) call an "indifferent lie." Had the agent's desire to deceive been rendered more salient in the scenario or had the benefit from lying been more explicit, the proportion of those who ascribe an intention to deceive might have been even higher.

References

Adar, E., Tan, D. S. & Teevan, J. (2013). Benevolent deception in human computer interaction. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May, pp. 1863–1872.

Anscombe, G. E. M. (2000). *Intention*. Cambridge, MA: Harvard University Press.

Banks, J. (2019). Theory of mind in social robots: Replication of five established human tests. *International Journal of Social Robotics*, 12, 403–414.

Beddor, B., & Pavese, C. (2021). Practical knowledge without luminosity. *Mind*, *129*(516), 1237–1267.

Bok, S. (1999). *Lying: Moral choice in public and private life*. New York: Vintage.

Brewer, B., Klatzky, R., & Matsuoka, Y. (2006). Visual-feedback distortion in a robotic rehabilitation environment. *Proceedings of the IEEE*, *94*(9), 1739–1751.

Broncano-Berrocal, F. (2013). Lies and deception: A failed reconciliation. *Logos & Episteme*, *4*(2), 227–230.

Brooks, C., & Szafir, D. (2019). Building second-order mental models for human-robot interaction. *Proceedings of the Association for the Advancement of Artificial Intelligence Fall Symposium Series (AI-HRI '19)*. arXiv:1909.06508.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*(2), 101–133.

Carson, T. L. (2006). The definition of lying. *Noûs*, *40*(2), 284–306.

Carson, T. L. (2010). *Lying and deception: Theory and practice*. Oxford, England: Oxford University Press.

Chakraborti, T. & Kambhampati, S. (2018). Algorithms for the greater good! On mental modeling and acceptable symbiosis in human-AI collaboration. arXiv:1801.09854.

Coleman, L., & Kay, P. (1981). Prototype semantics: The English word lie. *Language*, *57*(1), 26–44.

Crosby, M. (2020). Building thinking machines by solving animal cognition tasks. *Minds and Machines*, *30*(4), 589–615.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Danaher, J. (2020). Robot Betrayal: A guide to the ethics of robotic deception. *Ethics and Information Technology*, *22*, 117–128.

Davidson, D. (1971). *Agency*. In R. Binkley, R. Bronaugh & A. Marras (Eds.), *Agent, action, and reason* (pp. 3–25). Toronto: University of Toronto Press.

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, *68*(4), 87–106.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Devin, S., & Alami, R. (2016). An implemented theory of mind to improve human-robot shared plans execution. *HRI '16: The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, Christchurch, New Zealand, pp. 319–326.

Dragan, A. D., & Srinivasa, S. S. (2012). *Formalizing assistive teleoperation*. *Proceedings of Robotics: Science and Systems (RSS, 2012)*, Sydney, Australia, 9–13 July.

Dragan, A., Holladay, R., & Srinivasa, S. (2015). Deceptive robot motion: Synthesis, analysis and experiments. *Autonomous Robots*, *39*(3), 331–345.

Dynel, M. (2015). Intention to deceive, bald-faced lies, and deceptive implicature: Insights into Lying at the semantics-pragmatics interface. *Intercultural Pragmatics*, *12*(3), 309–332.

Fallis, D. (2009). What is lying? *The Journal of Philosophy*, *106*(1), 29–56.

Gino, F., Shu, L. L., & Bazerman, M. H. (2010). Nameless + harmless = blameless: When seemingly irrelevant factors influence judgment of (un)ethical behavior. *Organizational Behavior and Human Decision Processes*, *111*(2), 93–101.

De Graaf, M. M., & Malle, B. F. (2019). People's explanations of robot behavior subtly reveal mental state inferences. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Daegu, South Korea, 11–14 March, pp. 239–248.

Görür, O. C., Rosman, B. S., Hoffman, G., & Albayrak, S. (2017). Toward integrating theory of mind into adaptive decision-making of social robots to understand human intention. Workshop on the Role of Intentions in Human-Robot Interaction at the International Conference on Human-Robot Interaction, Vienna, Austria 6 March. Available at: https://researchspace.csir.co.za/dspace/handle/10204/9653

Kaminsky, M., Ruben, M., Smart, W., & Grimm, C. (2017). Averting robot eyes. *Maryland Law Review*, *76*, 983–1025.

Kneer, M. (2021). Success and knowledge in action: Saving Anscombe's account of intentionality. In T. Ciecierski & P. Grabarczyk (Eds.), *Context dependence in language, action, and cognition* (pp. 131–154). Berlin: De Gruyter.

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, *182*, 331–348.

Kneer, M., & Stuart, M. T. (2021). Playing the blame game with robots. In C. Bethel, A. Paiva, E. Broadbent, D. Feil-Seifer & D. Szafir (Eds.), Companion of the 2021 ACM/IEEE international conference on human-robot interaction (pp. 407–411). New York, NY, USA ACM.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, *130*(2), 203–231.

Krstić, V. (2019). Can you lie without intending to deceive? *Pacific Philosophical Quarterly*, *100*(2), 642–660.

Lackey, J. (2013). Lies and deception: An unhappy divorce. *Analysis*, *73*(2), 236–248.

Leong, B., & Selinger, E. (2019). Robot eyes wide shut: Understanding dishonest anthropomorphism. *FAT* Conference: Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, 29–31 January, pp. 299–308. https://doi.org/10.1145/3287560.3287591

Leveringhaus, A. (2018). What's so bad about killer robots? *Journal of Applied Philosophy*, *35*(2), 341–358.

Leyzberg, D., Spaulding, S., & Scassellati, B. (2014). Personalizing robot tutors to individuals' learning differences. *Proceedings of the 2014 ACM/IEEE international conference on Human-Robot interaction*, Bielefeld, Germany, 3–6, March, pp. 423–430.

Mahon, J. E. (2016). The definition of lying and deception. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Stanford: Metaphysics Research Lab, Stanford University.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*, 293–318.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. *10th ACM/IEEE International Conference on Human-Robot Interaction*, Portland, OR, 2–5 March, pp. 117–124.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction*, Christchurch, New Zealand, 7–10 March, pp. 125–132.

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology*, *10*, 450.

Matsuzoe, S., & Tanaka, F. (2012). How smartly should robots behave?: Comparative investigation on the learning ability of a care-receiving robot. *The 21st IEEE International Symposium on Robot and Human Interactive Communication*, Paris, France, 9–13 September, pp. 339–344.

Meibauer, J. (2014). Bald-faced lies as acts of verbal aggression. *Journal of Language Aggression and Conflict*, *2*(1), 127–150.

Meibauer, J. (2016). Understanding bald-faced lies: An experimental approach. *International Review of Pragmatics*, *8*(2), 247–270.

Mikalonytė, E. S., & Kneer, M. (2021). *Can Artificial Intelligence Make Art?* https://doi.org/10.2139/ssrn.3827314

Miracchi, L. (2019). A competence framework for artificial intelligence research. *Philosophical Psychology*, *32*(5), 588–633.

Nikolaidis, S., Lasota, P., Rossano, G., Martinez, C., Fuhlbrigge, T., & Shah, J. (2013). Human-robot collaboration in manufacturing: Quantitative evaluation of predictable, convergent joint action. *Proceedings of the 2013 IEEE 44th International Symposium on Robotics*, Seoul, Korea, 24–26 October.

Nourbakhsh, I. R., Sycara, K., Koes, M., Yong, M., Lewis, M., & Burion, S. (2005). Human-robot teaming for search and rescue. *IEEE Pervasive Computing*, *4*(1), 72–79.

Paul, S. K. (2009). *How we know what we're doing*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library.

Pavese, C. (2021). Knowledge, action, and defeasibility. In J. Brown & M. Simion (Eds.), *Reasons, justification, and defeat* (pp. 177–200). Oxford, England: Oxford University Press.

Perez-Osorio, J., & Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, *33*(3), 369–395.

Rabinowitz, N. C., Perbet, F., Song, H. F., Chiyuan, Z., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. *Proceedings of the 35th International Conference on Machine Learning*, *80*, 4218–4227.

Rios-Martinez, J., Spalanzani, A., & Laugier, C. (2015). From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics*, *7*(2), 137–153.

Rutschmann, R., & Wiegmann, A. (2017). No need for an intention to deceive? Challenging the traditional definition of lying. *Philosophical Psychology*, *30*(4), 438–457.

Saul, J. M. (2012). *Lying, misleading, and what is said: An exploration in philosophy of language and in ethics*. Oxford, England: Oxford University Press.

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, *12*(1), 13–24.

Schwenkler, J. (2012). Non-observational Knowledge of Action. *Philosophy Compass*, *7*(10), 731–740

Setiya, K. (2007). *Reasons without rationalism*. Princeton. NJ: Princeton University Press.

Setiya, K. (2009). *Intention*. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Stanford: Metaphysics Research Lab, Stanford University.

Shevlin, H., & Halina, M. (2019). Apply rich psychological terms in AI with care. *Nature Machine Intelligence*, *1*(4), 165–167.

Shim, J., & Arkin, R. C. (2012). Biologically-inspired deceptive behavior for a robot. *12th International Conference on Simulation of Adaptive Behavior*, Odense, Denmark, 27–30 August, pp. 401–411.

Shim, J., & Arkin, R. C. (2013). A taxonomy of robot deception and its benefits in HRI. *2013 IEEE International Conference on Systems, Man, and Cybernetics*, Manchester, UK, 13–16 October, pp. 2328–2335.

Sorensen, R. (2007). Bald-faced lies! Lying without the intent to deceive. *Pacific Philosophical Quarterly*, *88*(2), 251–264.

Southgate, V. (2013). Early manifestations of mindreading. In S. Baron-Cohen, H. Tager-Flusberg & M. V. Lombardo (Eds.), *Understanding other minds: Perspectives from developmental social neuroscience* (pp. 3–18). Oxford, England: Oxford University Press.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62–77.

Stokke, A. (2013). Lying and asserting. *The Journal of philosophy*, *110*(1), 33–60.

Stokke, A. (2016). Lying and misleading in discourse. *Philosophical Review*, *125*(1), 83–134.

Strichartz, A. F., & Burton, R. V. (1990). Lies and truth: A study of the development of the concept. *Child Development*, *61*(1), 211–220.

Stuart, M. T., & Kneer, M. (2021). Guilty artificial minds. *arXiv preprint arXiv:2102.04209*.

Tabrez, A., Luebbers, M. B., & Hayes, B. (2020). A survey of mental modeling techniques in human–robot teaming. *Current Robotics Reports*, 1, 259–267.

Tanaka, F., & Kimura, T. (2010). Care-receiving robot as a tool of teachers in child education. *Interaction Studies*, *11*(2), 263–268.

Timmermann, F., & Viebahn, E. (2021). To lie or to mislead? *Philosophical Studies*, *178*(5), 1481–1501.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)*, *53*(6), 1–38.

Turkle, S. (2010). In good company. In Y. Wilks (Ed.), *Close engagements with artificial companions*. Amsterdam: John Benjamins Publishing.

Turri, A., & Turri, J. (2015). The truth about lying. *Cognition*, *138*, 161–168.

Viebahn, E. (2017). Non-literal lies. *Erkenntnis*, *82*(6), 1367–1380.

Viebahn, E. (2020). The lying-misleading distinction: A commitment-based approach. *Journal of Philosophy*, *118*(6), 289–319.

Wagner, A. R. (2016). Lies and deception: Robots that use falsehood as a social strategy. In Judith A. Markowitz (Ed.), *Robots that talk and listen: Technology and social impact* (pp. 203–225). Berlin: De Grutyer. Available at: https://doi.org/10.1515/9781614514404.2016

Wagner, A. R., & Arkin, R. C. (2011). Acting deceptively: Providing robots with the capacity for deception. *International Journal of Social Robotics*, *3*(1), 5–26.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684.

Wiegmann, A., & Meibauer, J. (2019). The folk concept of lying. *Philosophy Compass*, *14*(8), e12620.

Wiegmann, A., & Rutschmann, R. (2020). Intending to deceive versus deceiving intentionally in indifferent lies. *Philosophical Psychology*, *33*(5), 752–756.

Wiegmann, A., Rutschmann, R., & Willemsen, P. (2017). Empirically investigating the concept of lying. *Journal of Indian Council of Philosophical Research*, *34*(3), 591–609.

Wiegmann, A., Samland, J., & Waldmann, M. R. (2016). Lying despite telling the truth. *Cognition*, *150*, 37–42.

Winfield, A. F. T. (2018). Experiments in artificial theory of mind: From safety to story-telling. *Frontiers in Robotics and AI*, *5*, 1–13. https://doi.org/10.3389/frobt.2018.00075

Zhao, Y., Holtzen, S., Gao, T., & Zhu, S. -C. (2015). Represent and infer human theory of mind for human-robot interaction. *2015 AAAI Fall Symposium Series*, *2*, 158–160.