Can a robot lie?

Exploring the folk concept of lying as applied to artificial agents

MARKUS KNEER[1]

Department of Philosophy, University of Zurich

The potential capacity for robots to deceive has received considerable attention recently. Many papers focus on 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 with 399 participants which explores the following three questions: (i) Are ordinary people willing to ascribe intentions to deceive to artificial agents? (ii) Are they as willing to judge a robot lie as a lie as they would be when human agents engage in verbal deception? (iii) Do they 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 it presently attracts.

CCS CONCEPTS • Human-centered computing~Human computer interaction (HCI)~Empirical studies in HCI

Additional Keywords and Phrases: Concept of Lying, Robot Deception, Human-Robot Interaction, Robot Ethics

ACM Reference Format: Markus Kneer. 2020. Can a robot lie? 2020.

1 INTRODUCTION

Innovation in artificial intelligence and machine learning has spurred increasing human-robot interaction in diverse domains, ranging from search and rescue via manufacturing to navigation [1-4]. For teamwork of this sort to succeed when complex tasks are at stake, humans and robots frequently 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 [5], and its role in cooperative human-robot interaction has received considerable attention recently [6-11], for a review see [12].

Once an artificial agent comes equipped with a theory of mind, it is *prima facie* capable of deception. Differently put, an agent of this sort is can 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 [13-15], for a review, see [16]. A considerable chunk of this literature focuses on *beneficial* deception [17], for instance in contexts of search and rescue, healthcare, and education, where "white lies" can, under certain conditions, have positive consequences (e.g. by inciting more effort in learning or rehabilitation activities see [19,20]). These are interesting case studies. As scholars with a bent for ethics have begun to highlight [22-25], 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

[1] markus.kneer@uzh.ch

marketing, politics, privacy and military applications. This is precisely what the present paper does. We'll focus on (i) *verbal* rather than *nonverbal* deception [26], of the (ii) *non-beneficial* rather than the *beneficial* kind, concentrating on (iii) the *human* rather than the *robot* perspective so as to explore the (iv) downstream *normative consequences* that matter most. Differently put, we'll explore whether humans, when interacting with artificial agents, are prone to attribute lies as readily and according to the same criteria to artificial agents as when they are interacting with other human agents.

The paper proceeds as follows: The concept of human lying is briefly examined in section 2.1, followed by a brief 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 (section 2.2). Section 3 presents a preregistered empirical experiment which explores (i) the propensity to judge different agent types (human v. robot) as lying (section 3.3.1), (ii) the willingness to ascribe an intention to deceive and actual deception across agent types (section 3.3.2) and (iii) blame attributions for lying across agent types (section 3.3.3). The implications of the findings are discussed in section 3.4, section 4 concludes.

2 THE CONCEPT OF LYING

2.1 Standard Accounts and Empirical Data

There is a large philosophical literature on the concept of lying ([27-32], for a review, see [33]), and the folk concept of lying has received considerable attention by empirically minded philosophers and linguists (for a review, see [34]). The following three criteria are frequently considered central to the prototype concept of lying [35]:

P1: The proposition uttered by the speakers is false. [Falsity]
P2: The speaker believes the proposition she utters to be false. [Untruthfulness]
P3: In uttering the proposition, the speaker intends to deceive the addressee. [Intention to deceive]

Coleman & 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 [32,33]. This would mean that a speaker cannot lie by uttering a true proposition which 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 [36], the overwhelming majority of findings suggests robust support for the subjective view for English-speaking adults [35, 37-39]. 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. Towards 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-face 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 [28,29, 31,40]. The standard response consists in denying that bald-face lies are lies in the first place [41-43]. Alternatively, one could also argue that they involve an intention to deceive (for an overview, see [44]). Empirical findings support the latter view [45,46]: Most people categorize bald-face lies as lies, though they *also* ascribe an intention to deceive to 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.

2.2 Lying artificial agents

Among the three general prototype criteria of a lie, falsity (P1) proves the least controversial when it comes to robots: Clearly artificial agents can utter propositions, and clearly these can be false. The untruthfulness and the intention to deceive, 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. [47-49]) – regarding which certain authors also caution care [50].

Let’s start with untruthfulness: While it might irk some to ascribe *belief* to artificial agents, it’s relatively unproblematic to say that artificial agents can entertain *informational states* and thus, in some limited sense, be aware of representations (what philosophers call “propositions”). Once this is granted, nothing obstructs positing a capacity for second-order propositions, such as taking a certain proposition *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.

How about the *intention* to deceive? What, precisely, intentions are, is controversial both philosophically (for a review, see [51]), and psychologically (see e.g. the debate surrounding the Knobe effect, [52,53]). However, most scholars agree that doing X intentionally entails (i) a pro-attitude such as a desire to bring about X as well as (ii) *some* epistemic state that one is bringing about X – be it knowledge (as suggested by Anscombe, see [54]) or mere belief (as argued by Davidson, see [55]). While care regarding the use of rich psychological states (“intends”, “wants”, “desires”, “knows”, “believes” etc., see [50].) is once again in order, we have already established the *prima facie* plausibility of (ii), i.e. epistemic states of sorts, for artificial agents. It is presumably also uncontroversial to say that such agents can have *goal states*, *objectives* or *quasi-desires* broadly conceived. The central question, then, is whether quasi-beliefs, quasi-desires, and quasi-intentions suffice to fulfil the capacities we expect an agent – human or not – to be capable of lying.

So far it has been established that, at least prima facie, artificial agents might have the required capacities for lying, or at least some proto-capacities that could suffice. 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 as a *pro tanto* wrong, and – odd cases aside – blame other people for lying, it is not clear that our moral assessment carries over neatly to artificial agents. One worry is that we might simply consider artificial agents as the wrong sort of agent for attributions of blame or moral responsibility (cf. Sparrow’s “responsibility gaps”, [58, 59]) – *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. There are some findings by Malle and colleagues which suggest that this is not the case. These authors, however, raise another problem: There might be things which are morally inacceptable (and/or blameworthy) when done by a human agent, yet morally acceptable (and blameless) when done by an artificial agent. Differently put, the normative landscape, and moral evaluation, might be sensitive to agent type. This is, in fact, exactly what previous studies found [56,57], and we will explore whether this kind of agent-dependent two-tiered morality also applies in the domain of deceptive linguistic communication.

3 EXPERIMENT

3.1 Participants

A total of 399 participants were recruited on Amazon Mechanical Turk to complete a Qualtrics online study for a small compensation. The experiment was rigorously preregistered. [2] The IP address of all participants was restricted to the US. 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 of 333 participants (42.9% female; age M=42 years, SD=13 years).

3.2 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, in another condition what he says is accidentally – and unbeknownst to the agent – true. The experiment thus took a 2 agent-type (human v. robot) x 2 truth value (false v. 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.]

2 https://aspredicted.org/blind.php?x=vn5vr3

3 https://www.reuters.com/article/us-health-coronavirus-japan-robot-hotels-idUSKBN22D4PC

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 anchored at 1 with "no blame at all" and 7 with "a lot of blame)

3.3 Results

3.3.1 Lying

The responses to the main question – whether Ken lied – are graphically represented in Figure 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 cases (binomial tests significantly above chance, all ps<.028, two-tailed). For false propositions, the proportion or participants who judged the human as lying was identical to the one judging the robot lying (92%). For true propositions, the proportion of participants who judged the human (75%) as lying exceeded those for the robot (64%), the difference was just about significant (binomial test, test proportion =.75, p=.038, two-tailed). Broadly speaking, 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.

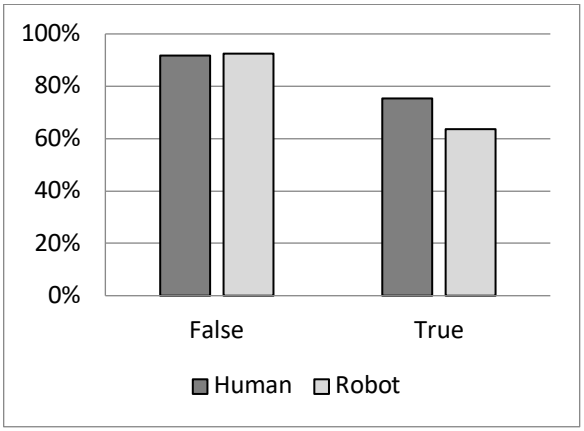

Figure 1: Proportions of participants who judged that Ken lied across *agent type* (human v. robot) and *truth value* (false v. true).

Table 1: Logistic regression predicting lying judgments

	B	SE	Wald	df	p	Odds Ratio
Agent type	0.076	0.539	0.02	1	0.887	1.079
Truth value	1.942	0.461	17.719	1	<.001	6.976
Interaction	-0.64	0.654	0.959	1	0.327	0.527
Constant	-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 type: robot, reference class for truth-value: false.

3.3.2 Deception

Figure 2 and 3 report the proportions of participants who thought the human and the robot had an *intention to deceive* and *actually deceived* their interlocutor respectively. 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), though the effect size was small: As Figure 2 illustrates the proportion of participants ascribing an intention to deceive to the human is marginally higher in the true than in the false condition, whereas the pattern is reversed for robots.[4] Overall the interaction matters but little: around 80% of participants or more ascribe an intention to deceive across all four conditions, which is significantly above chance (binomial tests, all ps<.001, two-tailed).

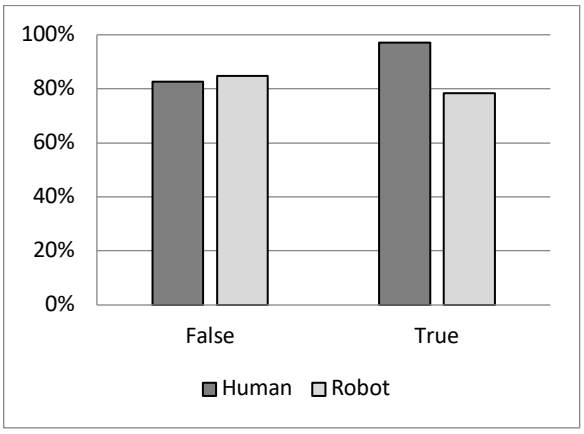

Figure 2: Proportions of participants who judged that Ken had an intention to deceive across *agent type* (human v. robot) and *truth value* (false v. true).

[4] Binomial test human false v. human true, test proportion=.83, p<.001, two-tailed. Binomial test robot false v. human true, test proportion=.78, p=.322.

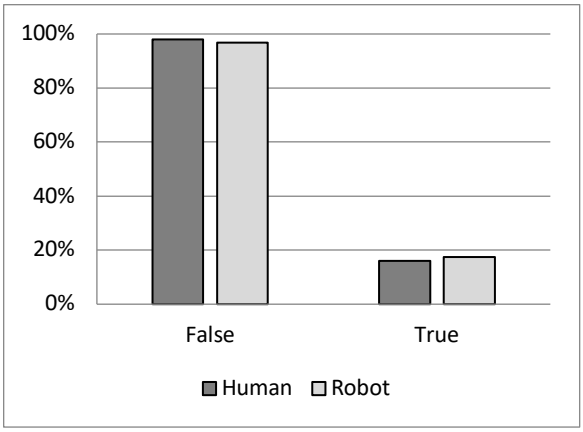

Figure 3: Proportions of participants who judged that Ken actually deceived their interlocutor across *agent type* (human v. robot) and *truth value* (false v. true).

Table 2: Logistic regression predicting intention to deceive

	B	SE	Wald	df	p	Odds Ratio
Agent type	0.156	0.394	0.157	1	0.692	1.169
Truth value	0.43	0.405	1.126	1	0.289	1.537
Interaction	-2.38	0.866	7.552	1	0.006	0.093
Constant	-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, reference class for truth-value: false.

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: Nearly all participants judged the intentional assertion of a proposition that was believed false and in fact false as actual, whereas less than 20% judged it a case of actual deception when the proposition asserted was accidentally true. The interaction was nonsignificant ($p=.561$), see Table 3.

Table 3: Logistic regression predicting actual deception

	B	SE	Wald	df	p	Odds Ratio
Agent type	-0.481	0.925	0.271	1	0.603	0.618
Truth value	4.936	0.662	55.639	1	<.001	139.205
Interaction	0.598	1.028	0.338	1	0.561	1.818
Constant	-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, reference class for truth-value: false.

3.3.3 Blame

A 2 agent type (human v. robot) x 2 truth value (false v. true) ANOVA for blame revealed a nonsignificant main effect of agent type ($F(1,329)=.277$, $p=.599$), an expectedly significant effect of truth-value ($F(1,329)=16.52$, $p<.001$) and a nonsignificant interaction ($F(1,329)=.011$, $p=.916$). As Figure 5 illustrates, and as the absence of a main effect for agent type confirms, people viewed the robot pretty much exactly as blameworthy as the robot for lying across conditions.

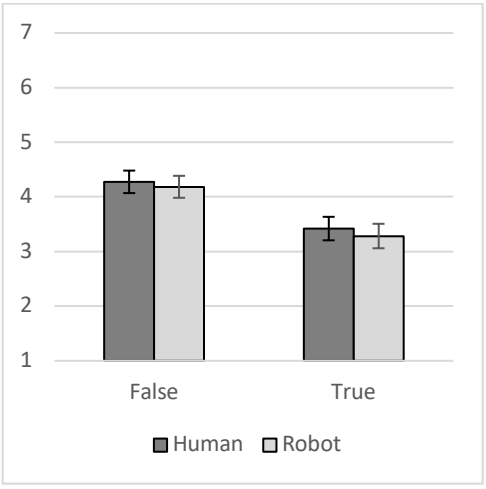

Figure 5: Mean blame rating across *agent type* (human v. robot) and *truth value* (false v. true). Error bars denote standard error of the mean.

3.4 Discussion

The findings of our experiment are loud and clear: The folk concept of lying applies to artificial agents in just the same way as it does for human agents. Consistent with previous research, we found that, *first,* it is possible for humans to tell a lie with a true statement (see [37-39]), and that this finding extends to robots (although the proportion who ascribe a lie in this case is 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 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 [50]. From a pragmatic point of view, this matters but little, since the folk is perfectly willing to *ascribe* intentions to robots. 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 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. [56]), where the moral evaluation of artificial agents differs significantly than 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 manipulate scenario and could, by aid of different illustrations of the robot agents (as done e.g. by Malle and colleagues, see [57]), investigate whether anthropomorphism has an effect on lying attributions and moral evaluation. Moreover, lab-experiments with deceptive embodied robots (see e.g. [16,26]) should be conducted to test the external validity of the findings presented. On the theoretical front, it is core to investigate the normative consequences of the presented findings (see [27]). 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" [58,59]: If robots are judged as capable of lying, and *are* attributed - contrary to what Sparrow and colleagues 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. Consequently, it must be explored whether it might be commendable to create norms, rules, possibly even laws, to restrict the use of actively deceptive robots in certain domains.

4 CONCLUSION

In a large-scale (N=399), preregistered experiment, we explored the folk concept of lying for both human agents and robots. Consistent with previous findings for human agents, lying is independent of the truth value of the proposition uttered, as well as the question whether the attempted deception succeeds. Instead, lying predominantly depends on an agent's intention to deceive. 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 pragmatically necessary to restrict the use of deceptive artificial agents.

REFERENCES

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

[2] Nikolaidis S, Lasota P., Rossano G., Martinez C., Fuhlbrigge T., Shah J. Human-robot collaboration in manufacturing: Quantitative evaluation of predictable, convergent joint action. In: IEEE ISR 2013, pp. 1–6 IEEE; 2013.

[3] Dragan A. D., Srinivasa S. S. Formalizing assistive teleoperation MIT Press. 2012.

[4] Rios-Martinez J., Spalanzani A., Laugier C. From proxemics theory to socially-aware navigation: a survey. Int. J. Soc. Robot. 2015;7(2):137–153.

[5] 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.

[6] Devin S., Alami R. An implemented theory of mind to improve human-robot shared plans execution. 2016.

[7] Zhao Y., Holtzen S., Gao T., Zhu S.-C. Represent and infer human theory of mind for human-robot interaction. In: 2015 AAAI fall symposium series, vol. 2; 2015.

[8] Scassellati B. Theory of mind for a humanoid robot. Auton. Robot. 2002;12(1):13–24.

[9] Leyzberg D, Spaulding S., Scassellati B. Personalizing robot tutors to individuals' learning differences. In: Proceedings of the 2014 ACM/IEEE international conference on Human- robot interaction, pp. 423–430 ACM; 2014.

[10] 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.

[11] Brooks C., Szafir D. Building second-order mental models for human-robot interaction. 2019. arXiv:1909.06508.

[12] Tabrez, A., Luebbers, M. B., & Hayes, B. (2020). A Survey of Mental Modeling Techniques in Human–Robot Teaming. Current Robotics Reports, 1-9.

[13] Shim and R. C. Arkin, "Biologically-inspired deceptive behavior for a robot," 12th International Conference on Simulation of Adaptive Behavior, pp. 401–411, 2012.

[14] A. R. Wagner and R. C. Arkin, "Acting deceptively: Providing robots with the capacity for deception," I. J. Social Robotics, vol. 3, no. 1, pp. 5–26, 2011.

[15] Chakraborti, T., & Kambhampati, S. (2018). Algorithms for the greater good! on mental modeling and acceptable symbiosis in human-ai collaboration. arXiv preprint arXiv:1801.09854.

[16] Dragan, A., Holladay, R., & Srinivasa, S. (2015). Deceptive robot motion: synthesis, analysis and experiments. Autonomous Robots, 39(3), 331-345.

[17] Shim, J., & Arkin, R. C. (2013, October). A taxonomy of robot deception and its benefits in HRI. In 2013 IEEE International Conference on Systems, Man, and Cybernetics (pp. 2328-2335). IEEE.

[18] E. Adar, D. S. Tan, and J. Teevan, "Benevolent deception in human computer interaction," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI '13. New York, NY, USA: ACM, 2013, pp. 1863–1872.

[19] B. Brewer, R. Klatzky, and Y. Matsuoka, "Visual-feedback distortion in a robotic rehabilitation environment," Proceedings of the IEEE, vol. 94, no. 9, pp. 1739–1751, 2006.

[20] S. Matsuzoe and F. Tanaka, "How smartly should robots behave?: Comparative investigation on the learning ability of a care-receiving robot," in RO-MAN, 2012 IEEE, 2012, pp. 339–344.

[21] F. Tanaka and T. Kimura, "Care-receiving robot as a tool of teachers in child education," Interaction Studies, vol. 11, no. 2, pp. 263–268, 2010.

[22] Kaminsky, M., Ruben, M., Smart, W., & Grimm, C. (2017). Averting robot eyes. Maryland Law Review, 76, 983.

[23] Leong, B. and Selinger, E. (2019). Robot eyes wide shut: Understanding dishonest anthropomorphism. FAT* Conference 2019. https://doi.org/10.1145/3287560.3287591

[24] Turkle, S. (2010). In Good Company. In Y. Wilks (Ed.), Close engagements with artificial companions. Amsterdam: John Benjamins Publishing.

[25] Danaher, J. (2020). Robot Betrayal: a guide to the ethics of robotic deception. Ethics and Information Technology, 1-12.

[26] Wagner AR. Lies and deception: Robots that use falsehood as a social strategy. Robots that talk and listen: Technology and social impact. De Grutyer https://doi. org/10.1515/9781614514404. 2016.

[27] Bok, S. (1999). Lying: Moral choice in public and private life. Vintage.

[28] Carson, T. L. (2006). The definition of lying. Noûs, 40(2), 284-306.

[29] Fallis, D. (2009). What is lying?. The Journal of Philosophy, 106(1), 29-56.

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

[31] Stokke, A. (2013). Lying and asserting. The Journal of philosophy, 110(1), 33-60.

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

[33] Mahon, J. E. (2016) The Definition of Lying and Deception, The Stanford Encyclopedia of Philosophy

[34] Wiegmann, A., & Meibauer, J. (2019). The folk concept of lying. Philosophy compass, 14(8), e12620.

[35] Coleman, L., & Kay, P. (1981). Prototype semantics: The English word lie. Language, 57(1), 26-44.

[36] Turri, A., & Turri, J. (2015). The truth about lying. Cognition, 138, 161-168.

[37] Strichartz, A. F., & Burton, R. V. (1990). Lies and truth: A study of the development of the concept. Child development, 61(1), 211-220.

[38] Wiegmann, A., Samland, J., & Waldmann, M. R. (2016). Lying despite telling the truth. Cognition, 150, 37-42.

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

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

[41] Lackey, J. (2013). Lies and deception: an unhappy divorce. Analysis, 73(2), 236-248.

[42] Meibauer, J. (2014). Bald-faced lies as acts of verbal aggression. Journal of language Aggression and Conflict, 2(1), 127-150.

[43] 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.

[44] Krstić, V. (2019). Can you lie without intending to deceive?. Pacific Philosophical Quarterly, 100(2), 642-660.

[45] Meibauer, J. (2016). Understanding bald-faced lies: an experimental approach. International Review of Pragmatics, 8(2), 247-270.

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

[47] Görür, O. C., Benjamin S. Rosman, G. Hoffman, and S. Albayrak. 2017. Toward Integrating Theory of Mind into Adaptive Decision-Making of Social Robots to Understand Human Intention. https://researchspace.csir.co.za/dspace/handle/10204/9653.

[48] Winfield, Alan F. T. 2018. “Experiments in Artificial Theory of Mind: From Safety to Story-Telling.” Frontiers in Robotics and AI 5. https://doi.org/10.3389/frobt.2018.00075.

[49] Rabinowitz, Neil C., Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. “Machine Theory of Mind”. https://arxiv.org/abs/1802.07740.

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

[51] Setiya, K. (2009). Intention. Stanford Encyclopedia of Philosophy.

[52] Knobe, J. (2003). Intentional action and side effects in ordinary language. Analysis, 63(3), 190-194.

[53] Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. Philosophical studies, 130(2), 203-231.

[54] Anscombe, G. E. M. (2000). Intention. Harvard University Press.

[55] Davidson, D. (1971). Agency. In R. Binkley, R. Bronaugh, & A. Marras (Eds.), Agent, Action, and Reason. Toronto: University of Toronto Press.

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

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

[58] Sparrow, R. (2007). Killer robots. Journal of applied philosophy, 24(1), 62-77.

[59] Leveringhaus, A. (2018). What's So Bad About Killer Robots?. Journal of Applied philosophy, 35(2), 341-358.