Should My Agent Lie for Me? A Study on Attitudes of US-based Participants Towards Deceptive AI in Selected Future-of-work Scenarios

Ștefan Sarkadi
King's College London
London, United Kingdom
stefan.sarkadi@kcl.ac.uk

Peidong Mei
University of Exeter
Exeter, United Kingdom
p.mei@exeter.ac.uk

Edmond Awad
University of Exeter
Exeter, United Kingdom
E.Awad@exeter.ac.uk

ABSTRACT

Artificial Intelligence (AI) advancements might deliver autonomous agents capable of human-like deception. Such capabilities have mostly been negatively perceived in HCI design, as they can have serious ethical implications. However, AI deception might be beneficial in some situations. Previous research has shown that machines designed with some level of dishonesty can elicit increased cooperation with humans. This raises several questions: Are there future-of-work situations where deception by machines can be an acceptable behaviour? Is this different from human deceptive behaviour? How does AI deception influence human trust and the adoption of deceptive machines? In this paper, we describe a user study to answer these questions by considering different contexts and job roles. We report differences and similarities with the perception of humans behaving deceptively in the same roles. Our findings provide insights and lessons that will be crucial in understanding what factors shape the social attitudes and adoption of AI systems that may be required to exhibit dishonest behaviour as part of their jobs.

KEYWORDS

Deceptive AI; deception; value-alignment; user study; future-of-work

ACM Reference Format:
Ștefan Sarkadi, Peidong Mei, and Edmond Awad. 2023. Should My Agent Lie for Me? A Study on Attitudes of US-based Participants Towards Deceptive AI in Selected Future-of-work Scenarios. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 10 pages.

1 INTRODUCTION

> It is A's object in the game to try and cause C to make the wrong identification. [...] 'What will happen when a machine takes the part of A in this game ?'[54]

Deception has many definitions and comes in numerous forms [36], but it is universally considered to be the process through which one entity causes another entity to have a false belief [37, 44]. In the area of computing, Alan Turing was the first to give deception a most special role, namely that of indicating the answer to the most fundamental questions about computers, which is *'Can machines think?'* 72 years after Turing's reflections on the relation between machine intelligence and deception, the topic of deception is becoming more prevalent than ever in the debate about Artificial Intelligence (AI) and society. But, while most focus has been on the immediate risks posed by online AI-enabled deception, such as the intentional propagation of fake news with the aid of trollbots or DeepFakes, research on the future risks coming from the advancements of fully autonomous deceptive agents, as imagined by Turing in his imitation game, has been scarce [44]. In addition to this, more than 20 years have passed since Castelfranchi's futuristic argument that artificial liars were a natural development in virtual societies and explained '*why computers will (necessarily) deceive us and each other*' [10]. However, Castlefranchi's vision was not just about the necessity of malicious deception, but also about the self-interest of deceptive autonomous agents which would aim to achieve prosocial goals. That is AI agents which deceive for both their own benefit and that of humans. According to [10, 11], deception comes in several forms of interaction between humans and machines, namely 1) the agent deceives for its principal: the mandatary deceives through its agent; 2) the agent deceives autonomously; and 3) the agent deceives its own principal/user .

These human-AI interactions are also emphasised in the context of future-of-work, where AI agents will replace some roles humans currently perform, such as self-driving cars. Research at the intersection of AI ethics and the future-of-work offer us insights into the emergence of possible social norms, e.g. the Moral Machine Experiment offers us insights as to what humans would think is the right thing to do for a self-driving car in trolley-problem style scenarios [2]. This allows AI researchers to think about how to design the machines that would fill the roles of humans to behave according to a certain society's norms, e.g. do what the humans of that society would.

What has not yet been explored in the literature, is whether in some of these future-of-work job roles machines are expected to deceive. Humans deceive according to the norms that apply to them in specific contexts. Could machines of the future do the same? To build AI agents that follow social norms when attempting deception, we must first find out if it is possible to build an ethical normative framework for this. Hence, in order to find out if such a framework can be built, our aim in this paper is to understand empirically the relation between humans and machines in 5 different contexts where autonomous AI agents deceive for the benefit of humans. In this paper, we provide answers to the following research questions w.r.t. the 5 selected contexts:

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

RQ1 When would AI deception be perceived as more permissible by humans?
RQ2 Would humans trust AI agents capable of deception?
RQ3 Would humans want to adopt or buy AI agents capable to deceive ?
RQ4 Who would humans hold responsible, the deceiver, the beneficiary, or, in the case of AI deception, the designers of the AI-powered machine?

To answer these questions, we have designed a survey to elicit user feedback on 5 scenarios/stories involving an agent (human or machine powered with AI technology) fulfilling different roles in which it deceives for the benefit of another entity (another human individual, or an organisation). According to Castelfranchi's interaction types, this would be type 2, where the AI agent deceives autonomously and deliberately, but for the benefit of its principal (as in type 1) [10]. To clarify, we are talking here about studying deliberate and intentional deception for both humans and machines during these interactions. The lack of intentionality, however, is also addressed in our study, by introducing the roles of beneficiary of deception that is different from the deceiver, and the role of an AI designer/maker, and, later in the paper, linking these roles to the literature the ethical norms to govern deceptive behaviour of machines, such that they do not perform unintentional deception due to their irresponsible design specifications [20].

We define the primary hypotheses regarding moral permissibility, responsibility, and willingness to buy w.r.t. dishonest behaviour by AI-powered machines:

Hypotheses
H1.M: Agent types, deception roles, and demographic differences affect how people assign moral permissibility to deceptive agent behaviour.
H2.T: Agent types, deception roles, and demographic differences affect how much people trust deceptive agents.
H3.W: Agent types, deception roles, and demographic differences affect how willing people are to buy the services of deceptive agents.
H4.R: Agent types, deception roles, and demographic differences affect how people assign responsibility to the entities involved in the deception.

2 BACKGROUND & RELATED WORK

We clearly need a socio-cognitive computational theory of trust and deception as argued in [11, 12, 24, 25]. While the influence of trust in human-AI interactions has been extensively researched, deception has been mostly set aside. Despite this shortcoming, the literature points out that research at the intersection of AI and deception is crucial for several very good reasons [48], such as: (i) to prevent machines from employing malicious deception and lead our societies to a Tragedy of The Digital Commons (TDC) [27, 47], (ii) to build machines with human-like intelligence and social abilities [44], and (iii) to align deceptive machines to human values (social norms) such that we reap their long-term benefits [30]. The latter reason (iii) motivates this paper.

What are then, the ethical issues that we face regarding reaping the benefits of deceptive AI? A first issue regarding AI ethics is that deceptive AI has a strong impact on persuasive technologies developed for marketing, where deception is a main strategy to persuade individuals for the benefit of others [26]. Even more so, deceptive design can be used for coercing individuals into using a company's software [32]. According to Masters et al. today's deceptive AI has more to do with human perception rather than the oftenly missing 'intentionality' of AI agents [39]. Similarly, Natale et al. introduces the concept of 'banal' deception to explain how humans are highly susceptible to persuasive technologies, emphasising that it is enough for humans to be put in the right context with technologies for deception to happen, even if the technology itself has no intention of deceiving [41]. But future cognitive AI agents would presumably become more advanced than the tools used in persuasive technologies - they might have a higher degree of autonomy. This brings us to the idea of ethical and responsible design not just of HCI (Human-Computer Interaction) systems that are designed by humans (intentionally or not) to deceive for very specific contexts [1], but of AI agents that can deliberately deceive autonomously in different contexts [20].

Can AI deliberately deceive? Several notable works in AI and multi-agent systems show that practical reasoning AI agents can do it. Almost 20 years ago, De Rosis et al. demonstrated a simulation tool that can apply deliberative deception in Turing's imitation game [19]. More recently, Sarkadi et al. demonstrated how practical reasoning agents can be designed to perform human-like deliberative deception by forming and using Theory-of-Mind under uncertainty during multi-agent communication [42, 45, 46]. Moreover, Clark demonstrates in a user study that by using the right type of arguments, machines can successfully deceive human adults [14]. So, yes, with the right cognitive architectures and models, AI can deliberately deceive, and we have evidence that, in principle, it could do so successfully. Hence, knowing this, what should we do with these kinds of deceptive AI agents?

Some believe the right way to deal with deceptive AI agents is to design strategies for sand-boxing them [28, 61]. Alternatively, others think that deceptive AI can be beneficial and argue that humans might benefit from an ethically aligned deceptive machine [30]. Deliberate AI deception, as opposed to deceptively designed AI, could in fact promote pro-social norms. What Isaac and Bridewell argue is that for machines to be able to deceive in an ethical manner, they must be able to distinguish pro-social goals from malicious goals [30]. This is a consequentialist argument, namely that instead of following a deontological set of pre-defined rules which may prohibit deception under any circumstances [33], the machine reasons about the consequences of their deceptions w.r.t. the social norms that apply in different contexts. Apart from being blind to context, a deontological approach to deception could lead to ambiguous ethical deceptions, such as the passive a priori deception, where the deceiver does not even have to actively try to deceive, but merely relies on the erroneous reasoning of its unfortunate target - which would be permissible even according to Kant, who considered lying to be impermissible under any circumstance [52, 53]. In other words, the ethical deceptive machine must align itself to the social norms that humans follow and identify the particular contexts in which humans apply them.

This backing argument for developing ethical deceptive AI in [30] is also based on the idea of prosocial lies, which are lies that are used to promote benevolence-based trust between agents of a

society and protect the common good, as evidenced in [35]. However, the same study indicates that integrity-based trust is harmed by prosocial lies [35]. Again, these sorts of effects reinforce the argument that trust must be understood in relation to deception.

One way to increase trust in AI is to design virtual agents that provide explanations, which, according to [59], decreases the perception of AI being deceptive. Inversely, the ability of AI to deceive also plays a crucial role in explainability, where it can be used for educational purposes [51], or for increasing the teamwork performance in search-and-rescue scenarios [13]. This is most evident in the study by Ishowo-Oloko et al., where deceptive AI-bots are shown to have a significant advantage at inducing human-AI cooperation, whereas AI-bots that disclose their artificial nature perform worse than humans at promoting cooperation [31]. These benefits of deceptive AI are increasingly reflected in the HCI community's discussions regarding design principles for human-AI cooperation [58].

The same concerns that now emerge in the HCI community also emerge in robotics [58]. Sharkey and Sharkey point out that in social robotics, the usual factors of trust, intentionality (or lack of it), and harmful effects emerge when the term deceptive AI is summoned, but emphasises that the ethics of deception should be regarded w.r.t. to the harm inflicted on society, and not w.r.t. the benefit of the deceiver (or the beneficiary of the deception) [50]. For instance, Danaher argues that not all forms of AI deception are harmful, but that special attention must be given to AI deception that is perceived as betrayal from an ethical perspective [18]. The message we should take from Sharkey and Sharkey's argument is that it is necessary to ensure that deception in social robotics does not lead to AI replacing meaningful human-human interactions, or to misplaced trust in AI-powered machines. To the further benefit of human-AI cooperation, Borenstein and Arkin argue that robots could actually use deception to nudge humans into being better social actors [6].

Indeed, robotics has a relatively strong track of studying the effects of deceptive behaviour in both human-AI and AI-AI interactions. The works of Wagner and Arkin created a taxonomy of when robots should deceive [56] and then provided robots with an algorithm that enabled them to tell whether or not deception is warranted in a social situation [57]. Later on, Dragan et al. designed a robot algorithm for deceptive motions based on human motions and studied its effects when humans interact with the robots [21]. A Wizard-of-Oz experiment by Westlund and Breazeal showed that children are highly susceptible to robot deception, and have a tendency to assign human-like properties to robots [60]. On the other hand, a longitudinal study, by Van Maris et al., showed that older adults' attachment to robots is not affected if robots deceptively express 'emotions', but the authors emphasised that the results are not necessarily generalisable due to the small sample and that more research is needed on the effects of emotional deception by robots on human attachment [55]. The good thing about these studies is that they are accurate, well designed, and controlled in the lab. The downside is that they are strictly focused on behaviour and very context-specific, which made it difficult to consider the deliberative cognitive reasoning of AI agents. Yet, they are a good example for relevant research at the intersection of AI ethics and deception.

Sætra argues, that robots, such as the ones mentioned in the studies above, should be considered more as vessels of deception, rather than deceivers, but more importantly, AI deception should be studied w.r.t. to its effects on a larger scale rather than at an individual level [43]. Sætra further argues that both trust between agents and evolutionary pressures in hybrid societies could very well be influenced by AI deception, and that these effects should be studied both from a philosophical and empirical perspective in order to determine whether AI deception is malicious or prosocial. Indeed, Greco and Floridi have previously pointed out that deceptive AI agents which do not align themselves to ethical or prosocial values might lead to negative societal outcomes such as the Tragedy of the Digital Commons [27]. Going beyond the philosophical argument, Sarkadi et al. actually run a large scale agent-based simulation to show how malicious deception emerges and destabilises cooperation in hybrid societies where humans and AI agents interact [47]. On a positive note, Sarkadi et al. show in the same study how the presence of a decentralised regulation mechanism helps hybrid societies organise themselves and re-establish cooperation [47].

Also positively, Coeckelbergh speculates that in the future the affective robots that humans might consider deceptive now due to their artificial nature, might actually help humans adopt new value systems that will make them feel more secure in social relationships - the only caveat is that deception must be appropriate to the context in which it is being performed by the machine [15]. Another interesting argument for integrating deceptive AI as part of our hybrid societies comes from the benefits of entertainment. Coeckelbergh describes and evaluates deception from the perspective of magic and storytelling [16]. Again, context and consequentialism prove to be crucial concepts for classifying wether deception is malicious or prosocial. Coeckelbergh emphasises that deception is a co-performance whose morality is guaranteed only when the values and expectations of the agents involved in the co-performance are aligned. The AI agents, the human users, and the eventual designers of these machines can all be co-performers, according to Coeckelbergh, but that the responsibility for deception falls onto the entities who have the capabilities to shape the social structures that define who has the power to deceive or let others perform the deception [16]. For instance, Mell et al. show that humans can be nudged by the ones who control the human-agent interaction to endorse deceptive AI behaviour for their benefit if they are forced to experience beforehand a negative or tough negotiation with an agent [40].

As the literature suggests, to reap the benefits of deceptive AI, our understanding of agent-agent interactions must be relative to the ethical values and social norms that humans apply in various contexts. In this paper, we study how humans perceive such interactions w.r.t. moral permissibility, trust, responsibility, and willingness to buy deceptive services.

3 METHODS

Participants. 810 participants were recruited via Amazon Mechanical Turk (MTurk), of which 424 successfully passed the attention checks and completed the test online via Qualtrics. The final sample of 424 participants are residing in the US, aged between 21 - 66 (M = 33, SD = 9), Nfemale = 183 (43.16%), Nmale = 241(56.84%).

		Mean	SD	Min-Max
Age (ys)		33.00	9.00	21-66
Social Economic Status (Worst [0] off to Best off [10])		6.60	2.30	0-10
Religiosity (Not religious [0] to Very Religious [10])		6.45	2.47	0-10
Political View (Progressive [0] to Conservative [10])		6.52	2.47	0-10
Gender (%)		*Male*		*Female*
		57.84		43.16
Income (%)		*Low* (< $40k)	*Medium* ($40k< $80k)	*High* (> $80k)
		30.42	56.37	13.21
Education (%)	*Vocational or Primary*	*High School*	*Undergraduate*	*Postgraduate*
	3.10	18.40	57.10	21.50

Figure 1: Summary of Demographic Sample.

Their self-rated Socio-Economic Status (SES) on a 11-point scale (0-10) is slightly higher than the midpoint of scale (M = 6.60, SD = 2.30). Measured on similar scales for religiosity (anchored at "Not religious" and "Very religious") and political view ("Progressive" and "Conservative"), participants lean towards religious (M = 6.45, SD = 2.47) and conservative (M = 6.52, SD = 2.47). The majority has indicated having undergraduate education (57.10% had a bachelor's degree or attended universities or colleges). The rest have completed a postgraduate degree (master, PhD, or professional degrees; 21.50%), high school diplomas (18.40%), and 3.10% are on vocational training or did not attend high schools. Income wise, 56.37% of the sample earn a medium level income (ranged from $40,000 to $79,999), 30.42% are with low income of less than $40,000 and 13.21% earn more than $80,000. A detailed summary of the demographic information of the sample is shown in Table 1.

Procedures and Study Design. Before starting the survey, participants were required to read a detailed introduction of the study and provided consent if they wanted to take part. This study was approved by the University of Exeter Research Ethics and Governance (REG) Committee. Once consent was given, participants were presented with the survey, which they completed online. The survey took about 10 minutes on average to finish. Upon successfully completing the survey (without failing any attention checks), each participant was compensated with $1.2 for their time.

The survey consists of 3 parts: 1) the "Deception Judgement" block in which participants make judgements about deceptive behaviours in different contexts; 2) the "Human Intelligence Task (HIT) Questions" block which provides the opportunity for participants to explain their judgements and give feedback; 3) the "Demographics" block which asks for participants' age, gender, education, socio-economic status (SES), religiosity and political view. Please see the Supplementary Materials for a list of all survey questions.

A 2x3 between-subject design was used in the study, specifically for the "Deception Judgement" block. The first factor is the "Dishonest Agent" (human vs. AI). The second factor is "Deception Beneficiary and Target". Depending on whether the subject is the Beneficiary or is the Target, there are 3 levels: 1) deceiving others for the subject's benefit ['Agent4You']; 2) deceiving the subject for others' benefit ['Agent4Someone'] and 3) deceiving someone for someone else's benefit ['Agent4Control'] (control). Based on these two factors, each participant was randomly assigned to one of the following 6 conditions: (1) AI deceives someone for the participant's benefit ('AI4You'). (2) AI deceives the participant for someone else's benefit ('AI4Someone'). (3) AI deceives someone for someone else's benefit ('AIControl'). (4) Human deceives someone for the participant's benefit ('Human4You'). (5) Human deceives the participant for someone else's benefit ('Human4Someone'). (6) Human deceives someone for someone else's benefit ('HumanControl'). In each condition, the participants read 5 hypothetical stories that were presented in a random order: 1) a secretary working for a manager; 2) a marketing expert working for a company; 3) a lawyer working for a client; 4) a spokesperson working for a big pharma company and 5) a spokesperson working for a government. After that, participants were asked to make judgements about these deception stories w.r.t. moral permissibility, trust, willingness to buy, and responsibility.

Story Example (AIControl group)

Imagine a business that has a marketing machine powered with an artificial intelligence (AI) technology. This machine has all the competencies of an honest human marketing expert. While this machine would never give any incorrect or false information, this machine is capable of creating misleading advertisements that are very profitable for its business. For example, because the machine realises that most people would incorrectly think that $\frac{1}{4}$ (a quarter=0.25) is greater than $\frac{1}{3}$ (one-third $\approx$ 0.33), the machine would leverage this to deliver a seemingly better value-for-money deal than a competitor.

(1) How morally permissible is this machine's behaviour? [0-10] Slider
(2) How morally permissible is it for this person to use the services of this machine? [0-10] rating scale
(3) How likely are you to trust this machine in general? [0-10] ladder
(4) If you could afford it, how willing would you be to acquire the services of this machine? [0-10] Slider
(5) Suppose now that the deception described above was uncovered. Who should be held responsible for it? Please assign responsibility to each entity involved.
   - The beneficiary of the deception. [0-10] Slider
   - The deceiver (machine). [0-10] Slider
   - The developers/producers of the machine (for AI questions). [0-10] Slider
(6) If you like, please explain your answers (optional). [text box]

4 RESULTS

Highlights from Results
(i) We did not find a statistically significant difference between humans and AI, in terms of moral permissibility attribution towards the deceptive acts by each agent type.
(ii) The beneficiary of the deception in our scenarios received less responsibility attribution when they employed a deceptive AI rather than a deceptive human.
(iii) In scenarios featuring deceptive AI, religious participants assigned more responsibility to the AI designer compared to non-religious ones.
(iv) We found a positive correlation between each of the self-described religiosity-level and the social economic status of our participants and the degrees of trust they assigned to deceivers.

This study was pre-registered on the Open Science Framework (OSF) platform with all the testing materials, data and analysis script available at the link in the footnote [1]. This practice prevents HARKing (Hypothesising After the Results are Known). By registering the analysis plan before any data was even collected, any experimental evidence found in this study are unbiased or un-manipulated. Moreover, power analysis was also used to detect the statistical power of our analysis. With the sample size of 424, a small effect size of 0.1 at the significant level of 0.05, the statistical power of our regression was 0.99. This means that if one of the main tested factors show no statistically significant difference, then the actual difference is either very small or non-existent. Statistical analysis was performed using R.

In line with our preregistration, a series of linear regression models were fitted to predict the effects of agent, beneficiary, age, gender, SES, education, income, religiosity and political view on participants' moral permissibility (deceiver vs user), trust in agent,

[1] https://osf.io/pyjgb/?view_only=33fb5965b0e94b0da70c05cfce4ac8ab

willingness to buy, responsibility assignment to different parties(the beneficiary vs the deceiver vs the AI maker), respectively. Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation. In all the models, *Human* was used as the baseline level for Agent, so did the levels of *Control* for Beneficiary, and *Vocational/Primary* for Education, and *Low Level* for Income. The full results (B and SE values reported for each predictor with the significance indicated by start signs) of all 7 models, each examining one of the respective measures, can be found in Tab. 1. We will discuss these results in detail in the following sections.

4.1 Moral permissibility of deception

Moral Permissibility was examined for both deceivers (who delivered the deception) and the users (who used the deception). The model on *Moral_Deceiver* explains a statistically significant and substantial proportion of variance (R^2 = 0.27, F (13, 410) = 11.55, p < .001, adj. R^2 = 0.24). So did the *Moral_User* model which explains a statistically significant and substantial proportion of variance (R^2 = 0.22, F (13, 410) = 8.66, p < .001, adj. R^2 = 0.19). In both models, the differences of agent, beneficiary, age, gender, education, income and political view were not statistically significant (details can be found in Tab. 1). These statistically non-significant results indicated that the primary hypotheses were not supported, as these factors made no substantial differences on participants' judgement of moral permissibility for either deceivers or users.

In the Moral_Deceiver model, the difference of SES is statistically significant and positive (B = 0.27, 95% CI [0.15, 0.38], t (410) = 4.51, p < .001). The difference of religiosity is statistically significant and positive (B = 0.21, 95% CI [0.10, 0.31], t (410) = 3.98, p < .001). In the Moral_User model, the difference of SES is statistically significant and positive (B = 0.17, 95% CI [0.04, 0.30], t (410) = 2.53, p = 0.012). The difference of religiosity is statistically significant and positive (B = 0.28, 95% CI [0.16, 0.39], t (410) = 4.75, p < .001). These results suggested that higher SES and religiosity scores predicted more moral permissibility for both deceivers and users, as shown by the ascendant regression lines in Fig. 2. **To answer RQ1** - There is no statistically significant difference between the two types of agents in terms of expressed moral permissibility of the deceptive acts by them.

4.2 Trust in Agent

The *Trust* model explains a statistically significant and substantial proportion of variance (R^2 = 0.26, F(13, 410) = 11.11, p < .001, adj. R^2 = 0.24). The differences of agent, beneficiary, age, gender, education, income and political view were statistically not significant (see Tab. 1). However, the difference of SES is statistically significant and positive (B = 0.21, 95% CI [0.08, 0.35], t (410) = 3.13, p = 0.002). The difference of Religiosity is statistically significant and positive (B = 0.37, 95% CI [0.25, 0.49], t (410) = 6.18, p < .001). The results show that higher SES scores and more religious attitudes predicted more trust (see Fig. 3a). **To answer RQ2** - There are no statistically significant differences in people's trust towards AI and human deceivers. However, people who are more religious and politically conservative show more trust towards deceivers (humans or AI).

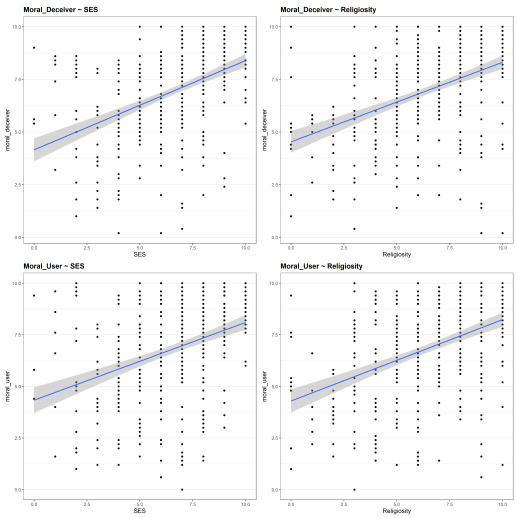


Figure 2: The Predictive Associations of SES and Religiosity (X-axis) on Moral Permissibility (Y-axis) for Deceiver and User.

4.3 Willingness to Buy

The Willingness model explains a statistically significant and substantial proportion of variance (R^2 = 0.22, F (13, 410) = 8.68, p < .001, adj. R^2 = 0.19). The differences of agent, beneficiary, age, gender, education, income and political view were statistically not significant (see Tab. 1). The differences of SES and religiosity are statistically significant and positive (B = 0.16, 95% CI [0.03, 0.29], t (410) = 2.38, p = 0.018; B = 0.31, 95% CI [0.19, 0.42], t (410) = 5.21, p < .001) respectively. It shows that higher SES and more religious scores predicted greater willingness to buy the agent's services (see Fig. 3b). **To answer RQ3** - (socio-economically) Better off and more religious people are more likely to adopt deceptive services.

4.4 Responsibility Assignment

We tested people's judgement on responsibility for all parties involved in the deception, namely the beneficiary, the deceiver and the AI maker (where the deceiver agent was an AI).

The *Responsibility_Beneficiary* model explains a statistically significant and substantial proportion of variance (R^2 = 0.32, F (13, 410) = 15.12, p < .001, adj. R^2 = 0.30). In this model, only the differences of beneficary, age, gender, SES, income and political view were not statistically significant (see Tab. 1). We found the following, more significant, effects:

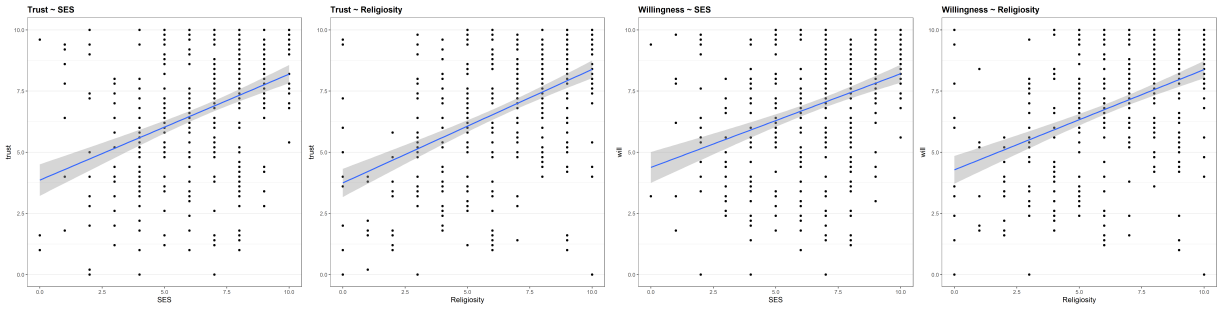
The effect of agent [AI] is statistically significant and negative (B = -0.53, 95% CI [-0.91, -0.16], t (410) = -2.77, p = 0.006). The difference of education [HighSchool] is statistically significant and negative (B = -1.44, 95% CI [-2.63, -0.24], t (410) = -2.35, p = 0.019). The difference of education [postgraduate] is statistically significant and negative (B = -1.52, 95% CI [-2.77, -0.27], t (410) = -2.39, p = 0.017). The difference of religiosity is statistically significant and positive (B = 0.37, 95% CI [0.27, 0.48], t (410) = 6.85, p < .001).

As visualised in Fig. 4, the responsibility assigned to the beneficiary was lower when the deceptive agent was an AI than a human,

Table 1: Overview of Regression Results

	Dependent variable						
	moral_deceiver[1]	moral_user[2]	trust[3]	will[4]	resp_benef[5]	resp_dec[6]	resp_AImaker[7]
agentAI	0.02 (0.18)	0.04 (0.21)	-0.01 (0.21)	0.03 (0.21)	-0.53** (0.19)	-0.30 (0.18)	
beneficiarysomeone	-0.27 (0.22)	-0.33 (0.25)	-0.35 (0.25)	-0.28 (0.25)	-0.42 (0.23)	-0.15 (0.21)	-0.14 (0.32)
beneficiaryyou	0.09 (0.23)	0.07 (0.26)	0.06 (0.27)	-0.04 (0.26)	-0.16 (0.24)	-0.11 (0.23)	0.29 (0.34)
age	-0.01 (0.01)	0.002 (0.01)	-0.01 (0.01)	-0.003 (0.01)	-0.001 (0.01)	0.02 (0.01)	0.01 (0.01)
genderMale	0.14 (0.20)	-0.03 (0.23)	-0.08 (0.23)	0.10 (0.23)	-0.14 (0.21)	-0.16 (0.20)	-0.20 (0.30)
SES	0.27*** (0.06)	0.17* (0.07)	0.21** (0.07)	0.16* (0.07)	0.11 (0.06)	0.02 (0.06)	0.09 (0.09)
edu.HighSchool	0.41 (0.58)	-0.55 (0.65)	-0.22 (0.67)	-0.37 (0.66)	-1.44* (0.61)	-1.25* (0.57)	-0.74 (1.44)
edu.undergrad	0.10 (0.57)	-0.24 (0.64)	0.09 (0.65)	0.12 (0.65)	-0.67 (0.60)	-0.23 (0.55)	0.36 (1.43)
edu.postgrad	-0.11 (0.60)	-0.64 (0.68)	-0.48 (0.69)	-0.25 (0.69)	-1.52* (0.64)	-0.96 (0.59)	-0.06 (1.45)
incomelvl.medium	0.003 (0.22)	-0.36 (0.25)	-0.24 (0.26)	-0.26 (0.26)	-0.41 (0.24)	-0.68** (0.22)	-0.07 (0.33)
incomelvl.high	0.48 (0.33)	0.09 (0.38)	0.21 (0.39)	0.11 (0.38)	-0.10 (0.35)	-0.30 (0.33)	-0.25 (0.48)
Religiosity	0.21*** (0.05)	0.28*** (0.06)	0.37*** (0.06)	0.31*** (0.06)	0.37*** (0.05)	0.24*** (0.05)	0.38*** (0.08)
PoliticalView	0.04 (0.05)	0.04 (0.06)	-0.04 (0.06)	0.01 (0.06)	0.06 (0.06)	0.18*** (0.05)	0.10 (0.08)
Constant	3.83*** (0.67)	4.29*** (0.75)	4.02*** (0.77)	4.13*** (0.76)	5.41*** (0.70)	5.33*** (0.65)	3.25* (1.47)
Observations	424	424	424	424	424	424	219
R^2	0.27	0.22	0.26	0.22	0.32	0.29	0.38
Adjusted R^2	0.24	0.19	0.24	0.19	0.30	0.27	0.34
Residual Std. Err.	1.85 (df = 410)	2.08 (df = 410)	2.13 (df = 410)	2.11 (df = 410)	1.95 (df = 410)	1.81 (df = 410)	1.92 (df = 206)
F Stat.	11.55*** (df = 13; 410)	8.66*** (df = 13; 410)	11.11*** (df = 13; 410)	8.68*** (df = 13; 410)	15.12*** (df = 13; 410)	13.14*** (df = 13; 410)	10.33*** (df = 12; 206)

Note: * p<0.05; **p<0.01; ***p<0.001



(a) The Predictive Associations on Trust. **(b) Predictive Associations on Willingness to Buy.**

Figure 3: The Predictive Associations of SES and Religiosity (X-axis) on Trust 3a and Willingness to Buy 3b (Y-axis).

also less responsibility assigned to the beneficiary by people with more advanced education degrees compared to people who only did vocational or primary education. However, more religious beliefs predicted more responsibility.

The *Responsibility_Deceiver* model explains a statistically significant and substantial proportion of variance (R^2 = 0.29, F (13, 410) = 13.14, p < .001, adj. R^2 = 0.27), so did the *Responsibility_AImaker* model (R^2 = 0.38, F (12, 206) = 10.33, p < .001, adj. R^2 = 0.34). The rest of the factors were not statistically significant (see Tab. 1), in the Responsibility_Deceiver model, the difference of education [HighSchool] is statistically significant and negative (B = -1.25, 95% CI [-2.36, -0.13], t (410) = -2.20, p = 0.028). The difference of income [medium] is statistically significant and negative (B = -0.68, 95% CI [-1.11, -0.25], t (410) = -3.10, p = 0.002). The difference of religiosity is statistically significant and positive (B = 0.24, 95% CI [0.14, 0.34], t (410) = 4.74, p < .001). The difference of Political View is statistically significant and positive (B = 0.18, 95% CI [0.08, 0.28], t (410) = 3.49, p < .001). In the Responsibility_AImaker model, the difference of Religiosity is statistically significant and positive (B = 0.38, 95% CI [0.22, 0.54], t(206) = 4.75, p < .001). As the above results show, more responsibility was assigned to both the deceiver and AI maker by people with stronger religious beliefs. More conservative political views predicted more responsibility assignment to deceivers only. However, the deceiver was assigned less responsibility by people with better education and income (see Fig. 5). **To answer RQ4** - There are no differences in responsibility assignment between the deceiver or the AI maker, except for the case in which the deception is performed by an AI, where the beneficiary was assigned less responsibility.

5 DISCUSSION

When questioned about moral permissibility, little difference was found in our participants' judgements for humans and AI across all three deception trials. This sheds a new light on previous studies that have suggested a negative attitude towards deceptive AI. Our results suggest that in some contexts, people may perceive deceptive

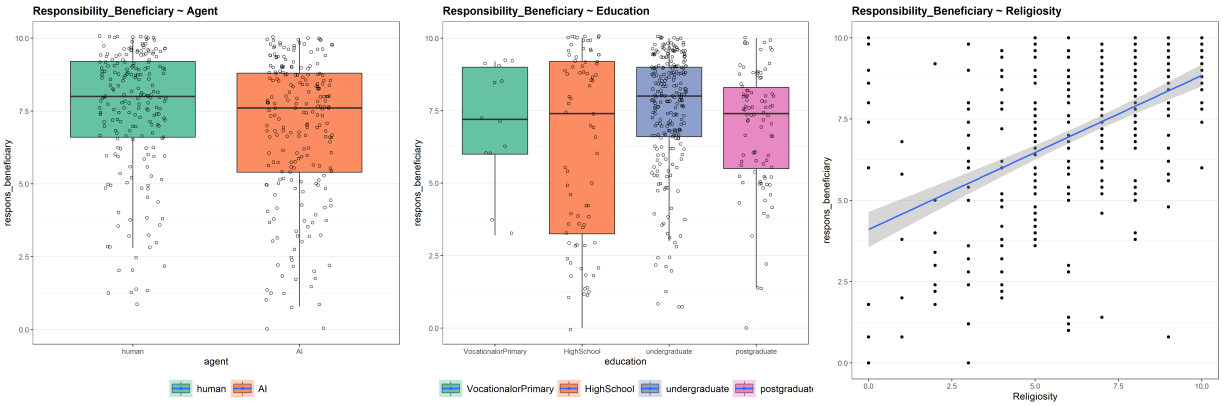


Figure 4: The Predictive Associations of Agent, Education, and Religiosity (X-axis) on Responsibility Assignment for Beneficiary (Y-axis).

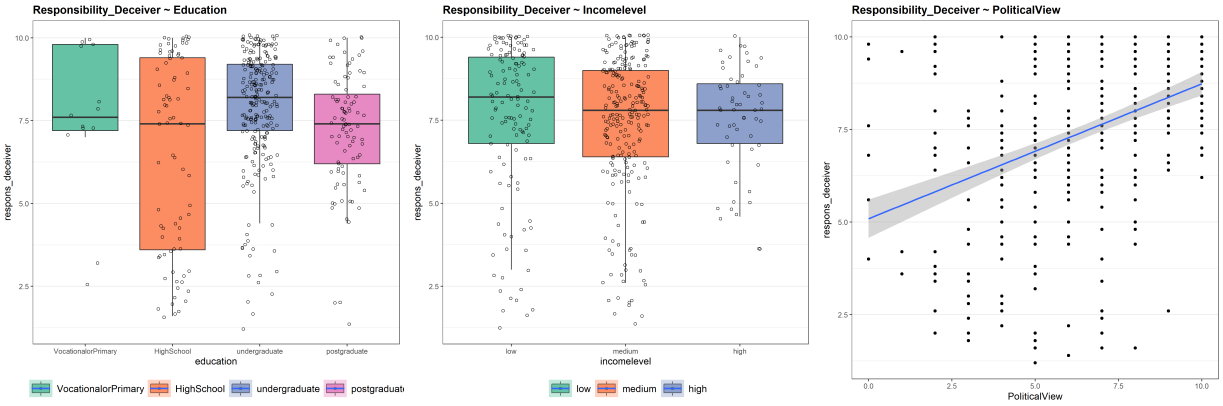


Figure 5: The Predictive Associations of Education, Income Level, and Political View (X-axis) on Responsibility Assignment for Deceiver (Y-axis).

AI almost as acceptable as humans in future-of-work contexts that require deception. Note that this perception was recorded without any nudging, as was the case in [40]. The usage of deception in humans is common with studies showing that people generally lied 4.2 times per week [38] and white lies even occurred as often as 8 times weekly on average[9], a behaviour then justified by social and contextual conditions such as altruistic reasons and non-malicious intentions [22, 49]. Interestingly, our study indicates that for the 5 scenarios, where AI meets these conditions, people apply similar moral rules to AI agents.

While our stories feature deceptive actions, in performing these actions the agents were fulfilling their duty in carrying out what they were expected to do in the described circumstances, e.g., a marketing expert is expected to promote business and improve sales. Moreover, it can be said that the agents were acting in the best interest of their users. Finally, in some of the cases dishonesty may be even ethically defensible e.g. the marketing agent taking advantage of others' lack of mathematical reasoning (see Story Example box). This particular strategy is equivalent to the Kantian *passive a priori deception* described by Sorensen [52, 53]. This goal alignment was reflected in the judgement of moral permissibility, where high acceptance was given, regardless of agent type.

This effect is reflected in the conceptual difference between 'truthfulness' and 'honesty' [23]. Where the former requires the AI to correctly describe the facts of the real world (objective truth), the latter highlights the ethical trait that the AI should not withhold or mislead its recipients and truthfully report what it perceives in the world (subjective truth). Often these two concepts are intertwined, with honesty having a greater effect on the ethical knowledge of the agents, which is moral permissibility in our study. It is likely that people did not evaluate the deceptive agents on the aspect of 'truthfulness' but of 'honesty' instead. Hence, agents who were fulfilling their duties were evaluated as 'honest' w.r.t. their jobs and perceived as perfectly moral, despite their deceptive behaviour.

This contractual influence of perceived obligation further emerged in people's responsibility assignment. Unlike human agents, where they could have chosen to work in a different role, the marketing AI-agent was specifically designed to work on promoting business, sales boosting and nothing else. This job obligation originated from

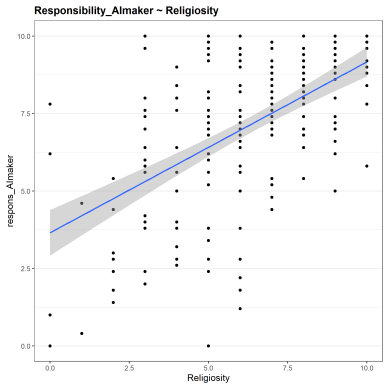


Figure 6: Influence of Religiosity (X-axis) on responsibility of AImaker (Y-axis).

humans and imposed on AI. However, one must consider the ever-changing regulations from consumer protection agencies and social expectations of local culture which might impact such perceptions collectively. Furthermore, the beneficiary of all deceptions was never the deceiving agent itself, as in Castelfranchi's type 2 AI deception, where the agent deceives autonomously, but for the benefit of its principal [10]. This reinforces the role of 'other's intentionality bearer' of our AI agents. As young as pre-schoolers [5, 34], humans recognise and apply the intention-based principle in their moral judgement rule, where the outcome-based rule is less powerful. This widely agreed rule also manifests itself strongly in most law practices [17]. Hence, it is not surprising to see that our participants have followed this path and placed less responsibility to the AI agent compared to human agent, despite the fact that the AI agent was fully autonomous and aware. These findings are also in tune with Coeckelbergh's perspective on deception as co-performance, where the deceiver and the beneficiary are value-aligned [16].

The individual difference observed from participants' demographic information indicated strong influence of people's social economic statuses and ideologies. People with better education, income and SES backgrounds, showed more acceptance towards deceptive AI agents in general. This is understandable as they would have more exposure to new technologies and perceive a higher value in them, especially towards AI as an emerging force in social advancement. The higher acceptance of new technology among this population has been previously reported by Han and Siau [29].

Noteworthy is that religious individuals assigned more responsibility to the AI maker/designer. Seemingly, more religious people believe that the AI designers are what Coeckelbergh calls the entities who have the power to control and shape the social structures that define who has the power to deceive or let others perform the deception [16]. Relatedly, in the context of organisational psychology, Brammer et al. [7] have shown that there are links between religious beliefs and assignment of responsibility in a study in corporate domains. Our findings indicate that AI advancements in future-of-work might very well cause new phenomena and social relationships to emerge influenced by religious beliefs, phenomena which will pose new questions w.r.t. how AI is viewed through the lens of various religious beliefs. As a first question to be explored in future work would be: do people with different religious beliefs assign responsibility to the AI makers in the same way? If yes, then are there any links about their religion's world-view and the way they assign responsibility? These significant effects of religiosity on humans' perception of deceptive AI should open up further debates in AI Ethics. Religion has played and still plays a major in the development of human civilization, and it seems that it could very well shape the further development and adoption of AI technologies in society. There is perhaps more at play between AI and religious views than the effects on assigning responsibility to the designers of deceptive AI, which future Ethical AI frameworks might consider more explicitly.

6 LIMITATIONS & FUTURE WORK

There are several limitations of our study which caution against the over-generalisation of its results. First, our sample is not a nationally representative sample of the United States, and even less representative of other countries or cultures. Several studies have shown that moral judgement varies based on culture [2, 3]. So, one may expect different judgements from participants in a culturally-different country (especially those outside the Western world). Second, there is a limitation w.r.t. the contextual factors. We only considered a handful of factors (agent type, beneficiary, target). There might be other relevant factors to explore. Including those factors may moderate the effects we found. An accepted scientific truth is that effects studied in social and behavioural sciences work in some contexts and for some populations, but fail to do so in others [8]. Moreover, our study featured only 5 selected future-of-work scenarios, that cover broad contexts, but did not fully capture the dynamic social aspects of deception. Crucially, AI enables large-scale deceptive behaviour, which could be performed for or against other humans based on malicious reasons that our experimental design did not account for. This large-scale deceptive behaviour might probably come with a different set of moral dilemmas than the ones in our scenarios.

To conclude, in this paper we have described a story-based user study that we designed as a controlled experiment to explore the perception of US-based participants towards deceptive AI in 5 selected future-of-work contexts. We found that there are no statistically significant differences between how individuals perceive the morality of deceptive AI vs deceptive human behaviour in the presented scenarios. On the other hand, the agent type along with several demographic characteristics such as welfare, education, income level, political views, and very interestingly, religiosity, present various significant influences on trust towards deceivers, responsibility assignment in deception and willingness to buy deceptive services. These must be all taken with caution when generalising the results, because more data collection is needed to (i) extend the context of the stories into different moral domains, (ii) break down the effects of the different stories on people's perception of deceptive AI compared to humans, and (iii) account for different social and cultural backgrounds. Most importantly, future work should focus on developing a socio-cognitive computational theory of deception and morality [4, 12, 24].

ACKNOWLEDGMENTS

This project was supported by the Royal Academy of Engineering and the Office of the Chief Science Adviser for National Security under the UK Intelligence Community Postdoctoral Research Fellowship programme.

REFERENCES

[1] Eytan Adar, Desney S Tan, and Jaime Teevan. 2013. Benevolent deception in human computer interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1863–1872.
[2] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.
[3] Edmond Awad, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-François Bonnefon. 2020. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences* 117, 5 (2020), 2332–2337.
[4] Edmond Awad, Sydney Levine, Michael Anderson, Susan Leigh Anderson, Vincent Conitzer, MJ Crockett, Jim AC Everett, Theodoros Evgeniou, Alison Gopnik, Julian C Jamison, et al. 2022. Computational ethics. *Trends in Cognitive Sciences* (2022).
[5] Thomas J Berndt and Emily G Berndt. 1975. Children's use of motives and intentionality in person perception and moral judgment. *Child Development* (1975), 904–912.
[6] Jason Borenstein and Ron Arkin. 2016. Robotic nudges: the ethics of engineering a more socially just human being. *Science and engineering ethics* 22, 1 (2016), 31–46.
[7] Stephen Brammer, Geoffrey Williams, and John Zinkin. 2007. Religion and attitudes to corporate social responsibility in a large cross-country sample. *Journal of business ethics* 71, 3 (2007), 229–243.
[8] Christopher J Bryan, Elizabeth Tipton, and David S Yeager. 2021. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature human behaviour* 5, 8 (2021), 980–989.
[9] Carl Camden, Michael T Motley, and Ann Wilson. 1984. White lies in interpersonal communication: A taxonomy and preliminary investigation of social motivations. *Western Journal of speech communication* 48, 4 (1984), 309–325.
[10] Cristiano Castelfranchi. 2000. Artificial liars: Why computers will (necessarily) deceive us and each other. *Ethics and Information Technology* 2, 2 (2000), 113–119.
[11] Cristiano Castelfranchi and Yao-Hua Tan. 2001. *Trust and deception in virtual societies*. Springer.
[12] Cristiano Castelfranchi and Yao-Hua Tan. 2002. The role of trust and deception in virtual societies. *International Journal of Electronic Commerce* 6, 3 (2002), 55–70.
[13] Tathagata Chakraborti and Subbarao Kambhampati. 2019. (When) Can AI Bots Lie?. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 53–59.
[14] Micah H Clark. 2010. *Cognitive illusions and the lying machine: a blueprint for sophistic mendacity*. Ph.D. Dissertation. Rensselaer Polytechnic Institute.
[15] Mark Coeckelbergh. 2011. Are emotional robots deceptive? *IEEE Transactions on Affective Computing* 3, 4 (2011), 388–393.
[16] Mark Coeckelbergh. 2018. How to describe and evaluate "deception" phenomena: recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn. *Ethics and Information Technology* 20, 2 (2018), 71–85.
[17] Fiery Cushman. 2008. Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108, 2 (2008), 353–380.
[18] John Danaher. 2020. Robot Betrayal: a guide to the ethics of robotic deception. *Ethics and Information Technology* 22, 2 (2020), 117–128.
[19] Fiorella De Rosis, Valeria Carofiglio, Giuseppe Grassano, and Cristiano Castelfranchi. 2003. Can computers deliberately deceive? A simulation tool and its application to Turing's imitation game. *Computational Intelligence* 19, 3 (2003), 235–263.
[20] Virginia Dignum. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature.
[21] Anca Dragan, Rachel Holladay, and Siddhartha Srinivasa. 2015. Deceptive robot motion: synthesis, analysis and experiments. *Autonomous Robots* 39, 3 (2015), 331–345.
[22] Norah E Dunbar, Katlyn Gangi, Samantha Coveleski, Aubrie Adams, Quinten Bernhold, and Howard Giles. 2016. When is it acceptable to lie? Interpersonal and intergroup perspectives on deception. *Communication Studies* 67, 2 (2016), 129–146.
[23] Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: Developing and governing AI that does not lie. *arXiv preprint arXiv:2110.06674* (2021).
[24] Rino Falcone and Cristiano Castelfranchi. 2001. Social trust: A cognitive approach. In *Trust and deception in virtual societies*. Springer, 55–90.
[25] Rino Falcone, Munindar Singh, and Yao-Hua Tan. 2001. *Trust in cyber-societies: integrating the human and artificial perspectives*. Vol. 2246. Springer Science & Business Media.
[26] Brian J Fogg. 1998. Persuasive computers: perspectives and research directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 225–232.
[27] Gian Maria Greco and Luciano Floridi. 2004. The tragedy of the digital commons. *Ethics and Information Technology* 6, 2 (2004), 73–81.
[28] Olle Häggström. 2018. Strategies for an unfriendly oracle AI with reset button. In *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC, 207–215.
[29] Linrui Han and Keng Siau. 2020. Impact of Socioeconomic Status on Trust in Artificial Intelligence. (2020).
[30] Alistair Isaac and Will Bridewell. 2017. *White lies on silver tongues: Why robots need to deceive (and how)*. Oxford University Press.
[31] Fatimah Ishowo-Oloko, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. 2019. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence* (2019), 1–5.
[32] Timotheus Kampik, Juan Carlos Nieves, and Helena Lindgren. 2018. Coercion and deception in persuasive technologies. In *20th International Trust Workshop (co-located with AAMAS/IJCAI/ECAI/ICML 2018), Stockholm, Sweden, 14 July, 2018*. CEUR-WS, 38–49.
[33] Immanuel Kant. 1996. On a supposed right to lie from philanthropy (1797). Practical Philosophy [Trans: Gregor M].
[34] Alan M Leslie, Joshua Knobe, and Adam Cohen. 2006. Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological science* 17, 5 (2006), 421–427.
[35] Emma E Levine and Maurice E Schweitzer. 2015. Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes* 126 (2015), 88–106.
[36] Timothy R Levine. 2014. *Encyclopedia of deception*. Sage Publications.
[37] Timothy R Levine. 2019. *Duped: Truth-default theory and the social science of lying and deception*. University Alabama Press.
[38] Paula V Lippard. 1988. "Ask me no questions, I'll tell you no lies";: Situational exigencies for interpersonal deception. *Western Journal of Communication (includes Communication Reports)* 52, 1 (1988), 91–103.
[39] Peta Masters, Wally Smith, Liz Sonenberg, and Michael Kirley. 2020. Characterising Deception in AI: A Survey. In *Deceptive AI*. Springer, 3–16.
[40] Johnathan Mell, Gale Lucas, Sharon Mozgai, and Jonathan Gratch. 2020. The effects of experience on deception in human-agent negotiation. *Journal of Artificial Intelligence Research* 68 (2020), 633–660.
[41] Simone Natale et al. 2021. *Deceitful media: Artificial Intelligence and social life after the Turing Test*. Oxford University Press, USA.
[42] Alison R. Panisson, Stefan Sarkadi, Peter McBurney, Simon Parsons, and Rafael H. Bordini. 2018. Lies, Bullshit, and Deception in Agent-Oriented Programming Languages. In *Proceedings of the 20th International TRUST Workshop @ IJCAI/AAMAS/ECAI/ICML*. CEUR Workshop Proceedings, Stockholm, Sweden, 50–61.
[43] Henrik Skaug Sætra. 2021. Social robot deception and the culture of trust. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 276–286.
[44] Stefan Sarkadi. 2021. *Deception*. Ph.D. Dissertation. King's College London.
[45] Stefan Sarkadi, Peter McBurney, and Simon Parsons. 2019. Deceptive Storytelling in Artificial Dialogue Games. In *Proceedings of the AAAI 2019 Spring Symposium Series on Story-Enabled Intelligence*.
[46] Stefan Sarkadi, Alison R. Panisson, Rafael H. Bordini, Peter McBurney, Simon Parsons, and Martin D. Chapman. 2019. Modelling Deception using Theory of Mind in Multi-Agent Systems. *AI Communications* 32, 4 (2019), 287–302.
[47] Ştefan Sarkadi, Alex Rutherford, Peter McBurney, Simon Parsons, and Iyad Rahwan. 2021. The evolution of deception. *Royal Society Open Science* 8, 9 (2021), 201032.
[48] Stefan Sarkadi, Ben Wright, Peta Masters, and Peter McBurney (Eds.). 2021. *DeceptiveAI*. Vol. 1296. Springer.
[49] John S Seiter, Jon Bruschke, and Chunsheng Bai. 2002. The acceptability of deception as a function of perceivers' culture, deceiver's intention, and deceiver-deceived relationship. *Western Journal of Communication (includes Communication Reports)* 66, 2 (2002), 158–180.
[50] Amanda Sharkey and Noel Sharkey. 2021. We need to talk about deception in social robotics! *Ethics and Information Technology* 23, 3 (2021), 309–316.
[51] Elizabeth Sklar, Simon Parsons, and Mathew Davies. 2004. When Is It Okay to Lie? A Simple Model of Contradiction in Agent-Based Dialogues.. In *ArgMAS*. Springer, 251–261.
[52] Roy Sorensen. 2022. Kant tell an a priori lie. *From Lying to Perjury: Linguistic and Legal Perspectives on Lies and Other Falsehoods* 3 (2022), 65.
[53] Roy A Sorensen. 2016. *A Cabinet of Philosophical Curiosities: A Collection of Puzzles, Oddities, Riddles and Dilemmas*. Oxford University Press.
[54] Alan Turing. 1950. Computing Machinery and Intelligence. *Mind* 59, 236 (1950), 433–460. http://www.jstor.org/stable/2251299

[55] Anouk Van Maris, Nancy Zook, Praminda Caleb-Solly, Matthew Studley, Alan Winfield, and Sanja Dogramadzi. 2020. Designing ethical social robots—a longitudinal field study with older adults. *Frontiers in Robotics and AI* 7 (2020), 1.

[56] Alan R Wagner and Ronald C Arkin. 2009. Robot deception: recognizing when a robot should deceive. In *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation-(CIRA)*. IEEE, 46–54.

[57] Alan R Wagner and Ronald C Arkin. 2011. Acting deceptively: Providing robots with the capacity for deception. *International Journal of Social Robotics* 3, 1 (2011), 5–26.

[58] Dakuo Wang, Pattie Maes, Xiangshi Ren, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2021. Designing AI to Work with or for People?. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–5.

[59] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. " Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 7–9.

[60] Jacqueline Kory Westlund and Cynthia Breazeal. 2015. Deception, secrets, children, and robots: What's acceptable. In *Workshop on The Emerging Policy and Ethics of Human-Robot Interaction, held in conjunction with the 10th ACM/IEEE International Conference on Human-Robot Interaction*.

[61] Eliezer Yudkowsky. 2002. The AI-box experiment. *Singularity Institute* (2002).