INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Using social robots to encourage honest behaviours

Sofia Miguel Martins Nunes Petisca

PhD in Psychology

Supervisors:

Doctor Francisco Esteves, Full Professor,
Mid Sweden University

Doctor Ana Paiva, Full Professor,
Instituto Superior Técnico

December, 2020

Social Psychology and Organizations Department

Using social robots to encourage honest behaviours

Sofia Miguel Martins Nunes Petisca

PhD in Psychology

Jury:
Doctor Abílio Oliveira, Assistant Professor, ISCTE – Instituto Universitário de Lisboa
Doctor Arvid Kappas, Full Professor, Jacobs University
Doctor Pedro Albuquerque, Associate Professor, Universidade do Minho
Doctor Elizabeth Collins, Invited Assistant Professor, ISCTE – Instituto Universitário de Lisboa
Doctor Francisco Esteves, Full Professor, Mid Sweden University

December, 2020

Acknowledgments

I want to thank Professor Francisco Esteves and Professor Ana Paiva for guiding me through the PhD, and Professor Elizabeth Collins for positive and valuable feedback during the development of my thesis. I also want to thank Professor Iolanda Leite for receiving me in a collaboration for my thesis, which greatly contributed to improving my PhD.

I would also like to thank the anonymous reviewer that gave a positive review of my PhD project that allowed me to receive a PhD grant, due to it, I was able to develop this thesis.

A special thanks goes to Filipa Correia, which help, positivity, and support allowed me to run my studies. I extend this thanks to Joana Campos, Sarah Gillet, Sandra Oristrell and Francesco Nuzzo, which helped me creating tasks for my studies (and fixing bugs). Lastly, I thank all the participants who were kind enough to give some of their time to contribute to this thesis.

This work was supported by the Social European Fund (FSE) and the Foundation for Science and Technology (FCT) with a Doctoral FCT Grant (Ref. SFRH/BD/118013/2016).

Resumo

Esta tese apresenta uma série de estudos para perceber se os robôs podem promover comportamentos honestos nas pessoas. No Estudo 1 observa-se que um robô que apenas olha para o utilizador, inibe batota, mas um robô que apresenta algum comportamento verbal não tem o mesmo efeito. No estudo 2, vemos que os participantes fazem batota tanto sozinhos, nas suas casas, como na presença de um vídeo de um robô que simplesmente olha. No Estudo 3 incluindo no robô a capacidade de perceber as jogadas dos participantes e reagir a elas, diminui a batota ao longo do jogo. No Estudo 4 a inclusão de um priming para o auto-conceito relacional não aumenta o efeito encontrado no Estudo 3. Finalmente, no Estudo 5 e 6 exploram-se as perceções das pessoas, e verifica-se que consideram errado ser-se desonesto com um robô, mas reportando baixos níveis de culpa. Justificam a desonestidade por: falta de capacidades no robô, falta de presença e a existência de uma tendência humana para a desonestidade. Quando avaliadas as atitudes que os outros teriam ou eles próprios em ser-se desonesto, manipulando o carácter afetivo do robô, não existem efeitos e as pessoas no geral reportam que os outros serão desonestos mantendo-se a si mesmas numa posição neutra. Curiosamente, os que demonstram atitudes mais negativas face a interagirem com robôs, reportam mais desonestidade. Estas são considerações importantes para o desenvolvimento de robôs para colaborarem com humanos no futuro.

Palavras-Chave: desonestidade; comportamento não-ético; batota; interações entre humanos e robôs.

PsycINFO Codes:

3000 Social Psychology

4140 Robotics

Abstract

This thesis presents a series of studies to understand if robots can promote more honest behaviours from people, when they are tempted to behave dishonestly. In Study 1 we see that a robot just presenting gaze behaviour inhibits cheating, but a robot doing small talk, does not. In Study 2 we see that participants cheated to an equal extent when doing the task in their homes alone or with a video of a robot looking at them. In Study 3 we find that including situation awareness in a robot (showing awareness of the participant behaviour), decreased cheating across the game. In Study 4 we see that priming participants for their relational self-concept does not enhance the situation awareness effect on cheating. In study 5 and 6 we explore participants perceptions, and we see that people consider it wrong to be dishonest towards a robot. However, they would feel low levels of guilt and justify it by the robots' lack of capabilities, presence, and a human tendency for dishonesty. When prompted to evaluate what other's/or their own attitudes would be regarding dishonesty, manipulating the caring behaviour of a robot, it shows no effect and people in general think others would be dishonest and hold themselves in a more neutral stance. Interestingly, people that show more negative attitudes towards robots tend to report that others will act more dishonestly as well as themselves. These are important considerations for the development of robots, in the future, to work alongside with humans.

Keywords: dishonesty; unethical behaviour; cheating; human-robot interaction.

PsycINFO Codes:

3000 Social Psychology

4140 Robotics

The present thesis is constituted by six studies:

Study 1: Petisca S., Esteves, F. and Paiva, A. (2019). Cheating with robots: how at ease do they make us feel? In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2102-2107. IEEE. https://doi.org/10.1109/IROS40897.2019.8967790

Study 2: Petisca, S., Paiva, A., & Esteves, F. (2020). The effect of a robotic agent on dishonest behavior. In Proceedings of the 20th ACM Internacional Conference on Intelligent Virtual Agents (IVA). ACM. https://doi.org/10.1145/3383652.3423953

Study 3: Human Dishonesty in the Presence of a Robot: The Effects of Situation Awareness (submitted for publication)

Study 4: The effect of a relational priming on dishonest behaviour (manuscript in preparation)

Study 5: Petisca S., Paiva A., Esteves F. (2020) Perceptions of People's Dishonesty Towards Robots. In: Wagner A.R. et al. (eds) Social Robotics. ICSR 2020. Lecture Notes in Computer Science, vol 12483. Springer, Cham. https://doi.org/10.1007/978-3-030-62056-1_12

Study 6: Perceptions of the effect of a caring robot on dishonesty: what others would do and what I would do (manuscript in preparation)

Table of Contents

Chapter 1- Introduction

Human behaviour is comprised of many different factors that make it complex and fascinating. And these factors react and transform in interaction with the environment they are in. All in our bodies pushes to strive and survive, for example, we are faster at recognizing angry expressions (more threatening) than friendly ones (e.g., Frischen et al., 2008; Williams et al., 2005), or when we are being watched, with faces with direct gaze capturing more our attention (e.g., Böckler et al., 2014; Hood et al., 2003), all with the aim of providing us with the best chances. But some of these survival-driven characteristics can be harmful for society, like human dishonesty. Studies show that we have an automatic self-interest tendency if the risks are low (Mead et al., 2009; Shalvi et al., 2012), this does not mean that we all act dishonestly, or cheat, but it suggests that under certain conditions some people will cheat, even just a little bit.

While human behaviour has been evolving, a product of our own creation, technology, has also been evolving rapidly. The creation of artificial intelligence allowed agents and robots to emerge as possible tools to help in various tasks, opening a window for future human-robot interactions in a diversity of fields. With this possibility comes the question of how people will behave when interacting with robots. Specifically, in contexts where honesty/dishonesty might be an issue, it becomes relevant to understand beforehand, how will people behave and in which way can we prepare robots to promote more honest interactions.

Up to this moment, no study has explored people's perceptions of being dishonest with a robot, and only two studies were conducted in the field exploring the effect of the robot presence with just gaze behaviour while watching a task where dishonesty could happen (Forlizzi et al., 2016; Hoffman et al., 2015). But as we envision future human-robot interactions, more complex interactions will need to emerge. As robots begin to acquire more autonomous behaviour and proactive capabilities, the relation and interactions between humans and robots will certainly evolve. If robots are integrated in a variety of contexts, such as classrooms, public settings or, for example, to provide assistance in people's homes (e.g., to accompany medications prescriptions for people who have difficulty following them alone, exercise routines or diet plans) there may be a temptation to disregard the norms and cheat by not following its suggestions, or because there is some gain to be obtained by misbehaving. In these situations, the robot will not only need to be capable of sustaining more complex interactions but also be effective in promoting honest behaviours. But since robot's developments are still at an early stage, it is important to first start testing in more simple contexts – like a task where people are

tempted to cheat in the presence of a robot and seeing which behaviours can promote more honesty.

With this aim in mind, and knowing that only two manipulations (with gaze behaviour) had been described so far in the literature, we developed a series of studies to test the effects of the presence of a robot in human dishonesty while people were doing a simple task where it was tempting to cheat. Our objective was to understand which kind of behaviours a robot would need to have to promote honesty and investigate the perceptions people hold towards dishonesty with robots, thus extending and informing the literature in the field.

The sequence of tests evolved in terms of situations and robotic behaviours. We started with a scenario exploring simple behaviours, reproducing the gaze effect already observed previously in the literature (and if it could transfer through a video) and we added small verbal behaviours, as for example, accompanying the participant in the task. We wanted to see if adding a verbal component could enhance the robot effect on dishonesty, and thus, allowing us to know if a more "social" robot could be used in such contexts. Next, envisioning more complex interactions we extended the robot's verbal behaviour and its social and interaction capabilities, by allowing it to know and react to the participant's behaviour. We wanted to see if giving this level of awareness to a robot, to show to the participant that the robot knew what was going on, could influence participant's dishonest behaviour. Furthermore, to explore the fact that when people are more focused on their relational self, they are less dishonest (e.g., Cojuharenco et al., 2012), we decided to prime participants for their relational self-concept by interacting with a more relational and supportive robot. This helped us investigate if interacting with a more supportive and friendly robot could enhance the awareness effect. At the same time, we also collected data on people's perceptions of being dishonest towards a robot, exploring why would people be dishonest and if this would differ depending on the type of agent (e.g., a more caring robot or a more neutral one). Participants would answer what they thought others would do and themselves, to give an idea if these two different perspectives would affect people's perceptions. These results would inform us on how people see dishonesty towards robots and if people's conceptions differ from their practical behaviour in the laboratory studies.

This research path was conducted along a set of studies. In Study 1, we tried to reproduce the effect of the robot gaze as inhibiting cheating (using direct gaze) and we tested the effect of adding small talk during the interaction. In Study 2, we explored if the robot watching behaviour through gaze, could also transfer using a video, having in mind situations where a virtual agent might be used instead of a robot. In Study 3, we explored if giving the robot awareness of the participant's actions and reacting to them, could affect cheating. And in a

following study, Study 4, we explored if adding a relational self-concept priming could enhance the effect.

In Study 5, we explored if people considered to be dishonest to cheat in the presence of a robot (and if the perceived autonomy of the robot affected it) and which reasons would make people act dishonestly in their presence. In Study 6, we explored how much people think others (and themselves) would be dishonest in a set of scenarios, varying the agent (alone, a human, a caring robot and a neutral robot) that was present in those situations and how much guilt would they feel.

Our results reflect a polarizing stance, it seems in order for a robot to be effective in decreasing cheating behaviour, it either has to not show the extent of its capabilities at all, or if it does, then it really needs to show that it can catch someone's dishonest act.

Chapter 2- Human dishonesty: a default honest self

Human dishonesty, even small acts (like cheating a bit, lying about something, etc.), is a concerning issue, due to the scalability that it can achieve from the aggregation of various small dishonest acts. The literature suggests that we have an automatic self-interest tendency that we need some level of self-control to keep in check, i.e. if we have the opportunity for it and a minimum risk of being caught, a lot of people have the tendency to cheat even just a little bit (e.g., Mead et al., 2009; Shalvi et al., 2012). And studies suggest that this tendency for self-interest is moderated by the type of victim that suffers from the dishonest act, i.e. when intuition is stimulated and dishonesty harms abstract others (for example, the laboratory budget), people tend to lie more, but the same does not happen if it can harm concrete others. Suggesting that when tempted to misbehave, people's intuitive response is to be selfish, especially if it harms abstract others in the process (Köbis et al., 2019). Fortunately, studies suggest that most people tend to avoid cheating on a maximum capacity (e.g., Mazar et al., 2008; Shalvi et al., 2011; see a meta-analysis by Abeler et al., 2019), still when added together, cheating in small amounts from a lot of people can have big consequences.

At the same time, studies show that people like to have a favourable self-concept of themselves (Fischbacher & Föllmi-Heusi, 2013; Mazar et al., 2008) and to be perceived by others as honest, an effect clearly seen in the mechanism known as moral hypocrisy (see Batson et al., 1997). Where people try to appear moral but avoiding the cost of it, with studies showing that participants choose a coin-flip to decide if them or others will receive a positive consequence (giving an appearance of fairness), but then the reports are deviated from the chance level, benefiting themselves (Batson et al., 1997). And the same is seen when people are given privileged information, they try to appear moral, but they end up using the privileged information for themselves (Batson et al., 2006). In sum, showing the need to appear honest/moral, by hiding the discrepancy between each one's perceived values and the values of their actions with self-deception (Batson & Collins, 2011).

In this thesis we do not tackle moral hypocrisy specifically. As Batson et al. (2006) explains, one of the conditions of moral hypocrisy is the existence of a decision towards resources that would be shared with other individuals. Instead, we refer to dishonesty when people take more than they should without others being directly affected in the transaction.

This preference to have a favourable self-concept and to be perceived by others as honest creates a contradiction with the automatic self-interest tendency, suggesting that people arrange a mechanism through which they can act on their self-interest but maintaining their

perception as being an honest person. By keeping a bit of both worlds, they can gain something and still perceive themselves as honest.

How do people cheat and still feel honest?

Human dishonesty was first thought of as based around external rewards, from a cost-benefit analysis: the amount gained, the probability of being caught and the punishment if caught (Becker, 1968). Recent studies started to show that internal rewards also play a part, especially influencing people's self-concept (in this case, how they view themselves in terms of morality). So, when people are confronted with a situation where it is tempting to misbehave, they feel two contradictory motivations. On one hand, they feel tempted to take advantage of the opportunity for themselves (they balance the risks of the external rewards). But on the other hand, they do not want to have a negative self-concept and feel as a cheater or a dishonest person (balancing the effect of the internal rewards).

To better understand how people's ethical decisions are made various theories have been developed across the years. One of the most influential theories was Rest's Four Component Model of Moral Behaviour, following the work from Kohlberg's and focusing on the individual to understand ethical decision-making. This theory postulates that individuals when faced with ethical decisions go through the following phases: moral awareness (recognizing and contextualizing an ethical situation), moral judgment (analysing ethical issues, determining right and wrong courses of action), moral intention (prioritizing ethical values and committing to ethical action) and moral action (implementing ethical action) (Rest, 1984, 1994). Yet, this model presumes, for example, that moral awareness is needed for a decision to have moral implications. A more recent model, Bounded Ethicality (Chugh et al., 2005; a revised version of the theory- Chugh & Kern, 2016) suggests that our ethical decision-making process is even more complex and can arise without intention and consequently, without awareness. We further explain this model and how it suggests that unethical behaviour (e.g., dishonesty) can happen. And then we look at the theory of Self-Concept Maintenance (Mazar et al., 2008) that highlights this process showing the effect of secondary control mechanisms on perpetuating unethical acts but absolving its consequences on the self-concept.

Bounded Ethicality theory (Chugh et al., 2005; a revised version of the theory- Chugh & Kern, 2016) postulates that people can unintentionally be dishonest. This theory suggests that often people do not recognize their own ethical misconduct because they are biased by a self-view of being moral, creating ethical blind spots. Bounded Ethicality in this sense refers to the limits on the quality of decision making with ethical significance (Chugh et al., 2005). A recent

revision of the theory postulates bounded ethicality as “the systematic and ordinary psychological processes of enhancing and protecting our ethical self-view, which automatically, dynamically, and cyclically influence the ethicality of decision-making” (Chugh & Kern, 2016, p. 86). This theory does not give the central role to self-interest but instead suggests that self-view (one’s interest in themselves, their own self-concept) is a more automatic influence on ethical decision-making, influencing the role of self-interest in it.

Therefore, Bounded Ethicality theory (Chugh et al., 2005; a revised version of the theory-Chugh & Kern, 2016) postulates that ethical behaviour is based on self-threat assessment (whether someone’s self-view of being ethical is threatened), which in turn determines if it activates mechanisms of self-enhancement (in the presence of low ethical self-threat, increases the positivity of the self-view) or mechanisms of self-protection (in the presence of higher self-threat, decreases the negativity of self-view). Mechanisms of self-enhancement are rather automatic (Krusemark et al., 2008) and unconscious, continuously operating if there is low self-threat. These mechanisms, regulate the need to feel good and to view oneself positively (Alicke & Sedikides, 2009), continuously fostering us to see our ethical behaviours as more ethical than they are, and our unethical behaviours as less unethical than in reality, creating bias that make it easier to act unethically. On the other hand, mechanisms of self-protection are less automatic and activate moral awareness, giving more salience to ethical implications and consequently, to behave more ethically (Chugh & Kern, 2016). In both mechanisms people can activate primary or secondary strategies to resolve the threats to their self-concept (Rothbaum et al., 1982). Primary control strategies are preferable because they are behaviours that act to change the situation (e.g., behaving more ethically), but if these are not enough to satisfy people’s goals or they fail, secondary strategies are triggered. These are psychological processes that can be engaged to alter and re-interpret the situations to maintain a positive self-concept (Alicke & Sedikides, 2009).

In the case of unethical behaviour, a problem emerges when there is a low level of threat to the self-concept, because automatic and unconscious processes maintain a mechanism of self-enhancement that makes it harder to become aware of ethical implications, by boosting one’s self-view, and fostering unethical actions. Only when a person feels a high level of threat to their self-concept (e.g., by being reminded of moral implications in a situation), they activate mechanisms of self-protection, and moral awareness, where ethical implications become clearer.

Small acts of unethical behaviour are not considered threats to the self because they can easily be resolved though self-enhancement secondary control mechanisms, creating ethical

blind spots. This is what the theory of Self-Concept Maintenance (Mazar et al., 2008) also shows, how people can perceive themselves as honest while doing small dishonest acts.

According to this theory, people experience ethical dissonance when being dishonest (between their dishonest actions and their honest self), compelling them to arrange strategies to decrease that dissonance. Two examples of these strategies can be categorization and attention to standards. Through them, people arrange a way in which they can take a little bit of advantage but not enough to harm their own self-concept of being an honest person. By categorizing, people can arrange justifications that excuse their dishonesty and consequently, not forcing them to update their self-concept. For example, someone might steal pencils from the company they work in, but they do not feel dishonest doing it, they might think that it is not harmful for the company, because they have a lot of pencils. Various studies support this mechanism, showing how having space for justifications can enhance a kind of moral flexibility that justifies dishonesty (e.g., Bassarack et al., 2017; Experiment 3 and 4 from Gino & Ariely, 2012; Experiment 3 from Mazar et al., 2008; Shalvi et al., 2012). In this case it's something external (e.g. being able to throw a die various times) that creates the opportunity to be dishonest and to still keep an honest self-concept. On the other hand, attention to standards relies on internal salience, i.e. when people are aware of their own moral standards, their actions will reflect in their self-concept (activating the self-protection mechanisms discussed in the Bounded Ethicality theory). This also goes in line with the Objective Self-Awareness theory, that shows us that by bringing self-awareness to the self there is a comparison with standards, when a discrepancy appears, there is a motivation to try and get to a consistent self (Duval & Wicklund, 1972; Silvia & Duval, 2001). When people are inattentive of their standards, their self-concept is less likely to be updated accordingly to the value of their actions. This attention to standards can be achieved with very simple manipulations like for example, making people sign an honour code (see Experiment 2 from Mazar et al., 2008) or asking people to see their own reflection and hear their own voice while doing a tempting task (Diener & Wallbom, 1976). Both categorization and attention to standards, can exist at the same time, for example, situations where people's awareness to one's own moral values is not made salient and there is space for creating justifications. Due to the possibility of gaining something by doing a small dishonest act people engage in secondary control strategies, for example, by creating justifications that categorize their unethical acts through a more favourable light, and by not paying too much attention to their own standards.

A factor that contrasts with the Bounded Ethicality theory (Chugh et al., 2005; a revised version of the theory- Chugh & Kern, 2016) is that Mazar et al. (2008) found in her studies that people notice their dishonest acts, they notice they overclaim even though they still cheat (see Experiment 4 from Mazar et al., 2008). Suggesting that there is a certain level of awareness

towards the dishonesty of an act. Yet, a study by Hochman et al. (2016), reinforces the relevance of the two theories, by showing that when people cheat, there is a certain level of conscious awareness (like the Self-Concept Maintenance theory predicts) and at the same time there are attentional bias happening at an unconscious level that also motivate the behaviour (in line with the Bounded Ethicality framework). Overall, people tend to feel a "default honest self" and hide behind this notion in some of their less ethical actions.

Factors that can influence dishonesty

Human dishonesty has been found to be influenced by individual characteristics and environment factors.

Individual Characteristics: it has been suggested that gender may have an effect on dishonest behaviour (e.g., lying or cheating), but the literature is mixed in this point. Some studies suggest that men are more likely to lie or cheat than women (e.g., Conrads et al., 2017; Dreber & Johanneesson, 2008; Friesen & Gangadharan, 2012; Gerlach et al., 2019; Houser et al., 2012), which can be related to the fact that studies find that women are more risk-averse than men (Croson & Gneezy, 2009). Others report the opposite (e.g., Clot et al., 2014; Ruffle & Tobol, 2014). And some seem to report no differences between the two genders in lying or cheating at the individual level (e.g., Aoki et al., 2010; Childs, 2012, 2013; Ezquerra et al., 2018; Gylfason et al., 2013; Muehlheusser et al., 2015). These mixed results do not provide a clear image on the effect of gender, and some suggest that gender differences might also depend on culture (e.g., Croson & Gneezy, 2009) but more research is needed on the topic.

Religion also has mixed results in influencing dishonesty, with some studies suggesting that it decreases cheating (Arbel et al., 2014; Bloodgood et al., 2008) and one study reporting it creates more lying (Childs, 2013). At the same time, other studies suggest that religiosity is not a good predictor of cheating behaviour (e.g., Martin, 2013; Ruffle & Tobol, 2014; Shariff & Norenzayan, 2011). Leaving a bit unclear the effect of religiosity on dishonest behaviour.

Personality is another characteristic that has been found to have a connection with dishonesty, especially the sixth personality domain of Honesty-Humility (according to the HEXACO Model), which evaluates the tendency to be fair and genuine when interacting with others, with higher values associated with lower opportunities for personal gains (Ashton & Lee, 2007). This dimension comprehends four sub-domains: Sincerity (the tendency to be genuine in interpersonal relations); Fairness (the tendency to avoid fraud and corruption);

Greed avoidance (the tendency to be uninterested to possess wealth) and Modesty (a tendency to be modest and ordinary) (Ashton et al., 2014). Not surprisingly, studies have found that this trait of personality can predict cheating behaviour, showing a negative correlation with it (e.g., Hilbig & Zettler, 2015; Kleinlogel et al., 2018; Pfattheicher et al., 2019), suggesting that people that get higher scores in this trait tend to cheat less and people who get lower scores, tend to cheat more. A recent large-scale re-analysis shows a medium to large effect of the Honesty-Humility on cheating behaviour (Heck et al., 2018), reinforcing this association.

Age is also an individual factor that might influence dishonest behaviour, some studies suggest that younger people behave more dishonestly than older people (e.g., Conrads et al., 2013; Friesen & Gangadharan, 2013), but other studies show no differences (e.g., Conrads & Lotz, 2015; Gino & Margolis, 2011) not becoming clear the effect of age on dishonest behaviour.

Lastly, another individual characteristic that might influence cheating is culture. A study found that Portuguese students were less inclined to fraudulent behaviour in comparison to Austrian students, whereas Spanish students were more inclined to cheat than Austrian students and no differences were found between Austrian and Romanian students (Teixeira & Rocha, 2006). With a follow-up study showing that more than 60% of students admit to cheating in Spain and Portugal, with Spanish students admitting higher values (Teixeira & Rocha, 2008). Another study showed significantly more negative attitudes towards cheating for Swiss students, and Ukrainian and Polish students with more positive attitudes (Chudzicka-Czupala et al., 2013). A study with students from Eastern Europe, Central Asia and the United States, showed how academic cheating is seen as a common activity, with less students from Eastern Europe and Central Asia believing that it was ethically wrong in comparison to students from the United States (Grimes, 2004). Even though these results are shocking and may point to some effects from culture in cheating behaviour, these studies are mostly based on attitudes towards cheating and not the act of cheating itself.

A cross-cultural study did comparisons in terms of cheating behaviour, comparing 23 countries from around the world, showing that countries where there is a higher prevalence for rule violations, more dishonesty is also found in a die-paradigm task (Gächter & Schulz, 2016). More recently, a cross-cultural study (with China, Colombia, Germany, Portugal and the United States) showed that tendencies towards dishonesty vary between countries but depend on life domain (e.g., work, relationships, government, etc.). For example, with student samples, Portugal seems to report more dishonesty academically, with strangers or in business, than in relationships (Garcia-Rada et al., 2018).

Overall, these results suggest that there might be an effect of culture in dishonesty. Still these relationships should be interpreted with caution, for example, a cross-cultural study found similar dishonesty levels across five countries (which vary in corruption and culture values), suggesting that dishonesty might be more connected to situational factors than specifically to cultural effects (Mann et al., 2016). A recent meta-analysis also showed that the majority of studies in dishonesty are performed in the United States and Germany with much lower values for example for studies performed in Sweden or Portugal (and the samples being mostly comprised of students), making it difficult to generalize the results to other populations and to take clear conclusions (Gerlach et al., 2019).

Environment factors: we cannot forget that we are social creatures, extremely sensible to social norms from others and the environment surrounding us. At any given time, we are exposed to social norms about how we should behave (e.g. how a behaviour is seen by society) and which behaviours are common to a given situation (see Cialdini et al., 1991). A problem can arise when the environment surrounding us promotes misbehaving. A good example of this is the "broken window effect" where signs of abandonment encourage people to misbehave (Wilson & Kelling, 1982). It seems that if people observe others breaking the rules, they become more likely to do the same (shifting their goals from appropriateness to more hedonic and gain goals), with for example, the mere presence of graffiti doubling the number of people littering or stealing (Keizer et al., 2008). These examples show how important it is to contradict dishonest and unethical acts, in order to stop the spreading of disorder.

Various studies have been done to ascertain the limits of human dishonesty and how better to inhibit them. The literature in human cheating behaviour shows that people have a propensity to cheat more depending on the environment they are in: by doing a tempting task in a dark room (Zhong et al., 2010), by having amounts of money visibly present (Gino & Pierce, 2009) (also just by handling literally dirty money; Yang et al., 2013), by not being monitored while doing a task (Békir et al., 2016; Covey et al., 1989; see Welsh & Ordóñez, 2014, Study 3), by seeing others considered as part of the in-group cheating (Gino et al., 2009; Martin, 2013), by feeling psychologically close to someone that cheats (Gino & Galinsky, 2012), or by using counterfeit sunglasses (Gino et al., 2010). Similarly, when people are depleted, they have less self-control to resist temptation (Gino et al., 2011; Mead et al., 2009) and if they have less time to perform a task (Shalvi et al., 2012), are given more space for justifications (Jiang, 2012; Shalvi et al., 2011), are primed to assume a narrow perspective (Schurr et al., 2012) or to adopt a loss frame (Grolleau et al., 2016; Kern & Chugh, 2009), it

becomes easier for them to create justifications that serve their self-interest and at the same time protects their self-concept. Interestingly, a study has also found that unethical behaviour increases after purchasing at a green product store in comparison to purchasing at a conventional one. It seems that buying in a green products venue gives people a moral pass to misbehave afterwards (Mazar & Zhong, 2010). Which has also been observed in a study where imagining a previous virtuous act licensed cheating afterwards (Clot et al., 2014). Overall, when there is almost no risk of being caught, some people cheat if they can still justify their actions and perceive themselves as honest.

To influence people to have more honest behaviours, and contradict this mechanism of self-deception, studies have found that making people more self-aware of their actions can decrease cheating. For example, by feeling watched or monitored (e.g., Békir et al., 2016; Covey et al., 1989; Mazar et al., 2008), by making people see their own reflection and hearing their own voice (Diener & Wallbom, 1976), by having no time pressures and no space for justifications (Shalvi et al., 2012), by reading an honour code (Shu et al., 2011) or signing one (Mazar et al., 2008; Shu et al., 2011), or by subconsciously activating moral standards (Welsh & Ordóñez, 2014). Thus, when people are reminded of their moral standards or are made aware of their dishonest self, the threat to the self is higher, compelling them to activate self-protection mechanisms and consequently, to cheat less. All these studies inform us on the complexity of human dishonesty and the variety of factors that can influence it, but in the future, what will happen when people start interacting with machines, especially robots?

Chapter 3- Interactions between Humans and Robots

Artificial intelligence (AI) and Robotics are often portrayed as malevolent technologies being often pictured by the media through their dark side. Yet, a significant part of the work done in AI and Robotics has helped us to grow technologically and consequently, contributed to make our life easier. We seamlessly use search engines to find anything, our email inbox can sort out spam automatically and one day not too far in the future, we might be able to use autonomous vehicles which will greatly contribute to reduce traffic disturbances. It is suggested that AI can enable the accomplishment of 134 targets from the Sustainable Development goals. Just in terms of Society, it could benefit on sustainable cities, clean water and sanitation, affordable and clean energy or quality education (Vinuesa et al., 2020), showing the promising avenues where AI could have a good impact.

Furthermore, as AI and Robotics grow, the role of the human in the interaction with these technologies has also increased and, human-robot interaction has gained more and more attention. Human-robot interaction is the area that investigates the design and creation of robots that are able to interact with humans, and at the same time studies how humans respond to such robots. The word "robot" comes from the Czech word *robota*, meaning labour[1], and a robot is usually defined as "a physically-embodied system" (Kiesler et al., 2008, p. 169), contrary to a virtual agent that is a system with no physical embodiment. Extensive research has been conducted in human-robot interaction, to understand specific aspects concerning how people will behave with robots and exploring the design space of robots in order to be able to interact naturally with people.

To encompass the complexity of human behaviour and explore the technological developments to sustain richer interactions between humans and robots, this field gathers researchers from various areas such as computer science, engineering, psychologists or designers. The main goal is to tackle all the different facets of social human-robot interactions and develop robots that will be able to support and help humans in a large variety of contexts. In this sense, social robots are robots that are capable of establishing social interactions with humans. Social robots are based on a general definition of a social agent, which are "artefacts, primarily computational, that are intentionally designed to display social cues or otherwise to produce a social response in the person using them" (Bickmore, 2003, p. 23). Therefore, a social robot can be formally defined as "an autonomous or semi-autonomous robot that

[1] https://web.archive.org/web/20130123023343/http://capek.misto.cz/english/robot.html

interacts and communicates with humans by following the behavioural norms expected by the people with whom the robot is intended to interact" (Bartneck & Forlizzi, 2004, p. 592).

The development of social robots has been guided by the fact that social interactions are extremely rich and varied, and as such, robots may impact many human features and trigger a variety of responses. In this thesis, we explore the effects that social robots have on human dishonesty.

Different types of robots

Science fiction likes to explore humanoid robots as the ultimate "robotic" creation. Yet, there are many types of robotic embodiments. For example, in industrial settings the typical robotic embodiment is "robotic arms" aimed at manipulation of elements in assembly lines. On the other hand, one famous robot named Curiosity[2] was even sent to Mars to collect specimens and move around in our neighbouring planet. But one important factor that predicts the robot usability is its morphology, which is strongly connected to its function. In the case of robots doing robust and mechanical tasks such as manipulation, they tend to present a more mechanical and robust aesthetic, like the Kuka robotic arms[3] or the Sawyer robot[4]. Their form is completely focused on their efficiency (on the task they need to perform). On the other hand, when robots are used for more social interactions with humans, their morphologies may resemble animals, such as AIBO[5] resembling a dog, Pleo[6] a dinosaur, and PARO[7] a seal. They can have a cartoonish form, that is more anthropomorphic, like the NAO[8], that is a humanoid with the size of a very small child, the EMYS[9] robot that is only a robotic head, or the Pepper[10] robot an almost human-like size robot. Yet, there are also more human-realistic forms, such as the Geminoid HI-2[11] designed to resemble a human being. According to the robot's aesthetics it subsequently informs on the population to which it is usually used with. For example, with children or the elderly the more zoomorphic and cartoonish robots are more common to be used (e.g., AIBO or PARO). With adults it is usually seen in studies the use of

[2] https://mars.nasa.gov/msl/home/
[3] https://www.kuka.com/en-de/products
[4] https://www.rethinkrobotics.com/sawyer
[5] https://us.aibo.com/
[6] https://www.pleoworld.com/pleo_rb/eng/lifeform.php
[7] http://www.parorobots.com/
[8] https://www.softbankrobotics.com/emea/en/nao
[9] https://emys.co/product
[10] https://www.softbankrobotics.com/emea/en/pepper
[11] http://www.geminoid.jp/en/robots.html

either NAO, EMYS or Pepper. Suggesting that different designs are developed having in mind the population that will mostly take advantage of the robot.

For this project we used two robots, EMYS and Pepper (seen in Figure 3.1). The EMYS robot is a robotic head that is attached to a table, EMYS does not have body but it compensates this lack of body in the amount of expression that it can provide with its cartoonish face. EMYS has animated eyeballs, which enables its gaze to be accompanied by blinking movements, which consequently provides greater animacy and expressive presence. Due to this characteristic we considered EMYS to be a good choice to convey the sense of vigilance or monitoring behaviour. At the same time, EMYS also has a mouth (represented by its two lower discs that moves according to what the robot is saying) which would make sense for the condition where we would give verbal behaviour to the robot. For this reason, we used EMYS for our first two experimental studies. For our third and fourth experimental study, since EMYS was not available, we used the Pepper robot. We could have adopted the NAO robot, but due to its more childish appearance, we considered it would not be appropriate for our scenarios. As such, in our subsequent studies we used Pepper, a robot that has a full body (with torso, a lower part and arms) and the same social cues as EMYS, with eyes and a mouth. Both EMYS and Pepper have a humanlike appearance, which has been found in the literature to be preferred when robots are used in tasks that require social skills (see Study 1 in Goetz et al., 2003). In our case, the robot would need in some conditions to be able to present social skills as verbally speaking with the participant's and accompanying the task, matching the appropriateness of their embodiments to this thesis tasks.

Figure 3. 1- Robots used in this thesis (on the left EMYS robot and on the right Pepper robot).

The fact that we used two different robots is not ideal, since it changed the robot embodiment between the studies, which could differently affect people's behaviours. And this factor is acknowledged in the limitation's section.

Different contexts for human-robot interactions

Research in human-robot interaction is envisioning robots for a variety of contexts, in which their functions can be varied. They can be useful to perform more mechanical tasks, like working in factories (with demand for industrial robots increasing) (Grau et al., 2017), or they can be used over long distances as a telepresence tool, enabling people to have a physical presence somewhere else, such as for example in a meeting abroad, avoiding extra travel and facilitating the quality of the interaction (e.g., Adalgeirsson & Breazeal, 2010; Lee & Takayama, 2011; Vespa et al., 2007). Another area of relevance is to use robots as a tool for healthcare and in particular for rehabilitation, for example to help stroke survivors with severe chronic impairments, perform therapy for wrist rehabilitation (e.g., Krebs et al., 2007) or for robot assisted lower limb rehabilitation (e.g., Meng et al., 2015). Another example where more mechanical robots can be used is for dangerous tasks, like rescue missions (e.g., Cacace et al, 2016; Kang et al., 2005), diminishing human exposure and risk to those situations.

On the other hand, robots can also be more social and assist humans, for example, as assistants/tutors for education, giving support to the teacher's role (see Belpaeme et al., 2018), allowing for teachers to have more time in a class and more time per student. Social robots can also be useful in giving assistance for the creation of healthy habits (e.g., De Carolis et al., 2019; Fasola & Mataric, 2013; Kidd & Breazeal, 2008; Ros et al., 2016), or in accompanying and assisting the elderly, with various studies already testing the use of a robot for repeated interactions with the elderly population (e.g., Fischinger et al., 2016; Graf et al., 2004; Khosla et al., 2012; Robinson et al., 2013; for a review see Kachouie et al., 2014). In this context, social robots can be useful in a variety of ways, such as providing help in tasks that could be difficult to execute (e.g., picking things from the floor), giving support (by providing company or cognitive stimulation) and possibly even motivating people to keep connected with their close ones. All these examples, showing the variety of tasks that robots are being thought of as a possible support for human beings, and the variety of contexts they can be integrated in.

However, all these future human-robot interactions will also have to consider the complexities of human behaviour and especially in contexts where the robot will have a supervision role. Which can range from a robot, for example, assuring that people take their

medication at home or follow their diet, to preventing cheating behaviour in a classroom or a public setting. Reinforcing the importance of understanding the effect of the robot's presence on human behaviour.

Effects of robot's presence

Future interactions between humans and robots only make sense if robots can in fact provide support and influence human behaviour. For robots to provide support, their presence must have an advantageous effect on human behaviour, comparatively, for example, to the use of a virtual agent (with no physical embodiment). Several studies have been conducted to explore the responses of humans to robots and agents in general. The literature suggests, for example, that people consider a robot helpfulness towards health advice much higher than when interacting with a virtual agent. People report feeling a greater sense of presence and engagement with the physical robot, and robots are rated higher for personality traits and seem to elicit less self-disclosure of undesirable behaviours than with a virtual agent, which might express a greater social influence from the presence of the physical robot in comparison to the virtual agent (Kiesler et al., 2008; Powers et al., 2007). Other studies also suggest this possible social facilitation effect from the presence of a robot, comparable to the effect a human presence does (Riether et al., 2012), suggesting that people employ more effort in the presence of a robot in comparison to a virtual agent (Bartneck, 2002). Even when controlling for differences in embodiments, when participants were just presented with a pair of eyes either from a robot or a virtual agent, in a collaborative task, participants reported with the robot much more engagement, enjoyability, informative capabilities and credibility than with the virtual agent (Kidd & Breazeal, 2004). It seems a robot can bring more advantages of presence and effectiveness than, for example, a virtual agent. When comparing a robot through different mediums, for example, a physically present robot, a robot transmitted through a video or an animated robot in a screen, the physically present robot is rated as more watchful, more helpful and more enjoyable (Wainer et al., 2007). This is especially significant for contexts where the robot needs to have a watchful presence, as in this project, suggesting that a robot instead of an agent might be a more appropriate tool to use in order to influence human dishonesty. Overall, all these studies suggest that people respond more favourably to physically present robots, having a significant effect on people's behavioural response. Interestingly, in a recent analysis of the literature it seems that what affects participant's behaviour is the physical presence and not the physical embodiment per se (Li, 2015). Suggesting, that one advantage of using robots is their ability to have a greater presence than other forms of technology.

Another advantage of using a robot is that studies suggest that when humans interact with technology, they apply social rules (e.g., Hoffmann et al., 2009; Nass & Moon, 2000), which at the beginning might have come from social scripts from human-human interactions but with time, and the increasing prevalence of technology in our lives, we may have started to create specific social scripts from these human-machine interactions (see Gambino et al., 2020 for a review). And even though a robot is just a machine, it is for example, interesting to see that people take much more time to turn off a robot (that helped them during a task and was agreeable) that is pleading to not be turned off, than a robot that was not so helpful (Bartneck et al., 2007), notwithstanding the fact that these results should be interpreted with care due to a small sample. Another study showed for example that people show emotional reactions towards robots, with increased physiological arousal, more negative affect and expressed empathic concern for a robot that is being physically mistreated (or "harmed") in a video, in comparison to one that is not (Rosenthal-von der Pütten et al., 2013). Suggesting that even though they are just machines, people still react to them and connect with them at some level.

These effects of the robot presence and how people socially respond to them, open a window for the use of robots as persuasive machines, raising the question if a social robot can influence human behaviour for the better.

Therefore, to know if robots should be used, for example, as assistive technology for the elderly, as coaches or tutors for healthier habits, first it should be investigated if a robot can influence human behaviour. Some studies started exploring this, showing for example, that a robot can positively affect children learning, even more so when it personalizes itself to the characteristics and progress of a child (Baxter et al., 2017). A robot was also more efficient in persuading people to consume less energy in a washing machine, than if it was not present (Midden & Ham, 2012). Interestingly, a study showed that when participants were required to perform an unusual task, as picking a set of expensive-looking textbooks and throwing them in a trashcan, a physical robot was much more effective in persuading participants to do it, than a video or an augmented-video robot condition (Bainbridge et al., 2011), suggesting that the physical presence of a robot has a stronger effect for persuasion. Another study showed how a robot can persuade people to choose a specific brand of coffee to consume either by using a reward strategy (rewarding people with a joke) or an expertise strategy (providing information on the quality of the coffee) (Hashemian et al., 2019). Or how a robot presenting an assertiveness trait (Paradeda et al., 2019) or persuading arguments (Paradeda et al., 2020), can persuade people to change their decisions in a collaborative storytelling scenario.

Besides these simpler persuasion scenarios, it is also important to consider more complex scenarios, if a robot can persuade people to do something that implies a certain amount of

effort. A study by Fasola and Mataric (2013) observed that older adults preferred a physically embodied robot coach for exercise (instead of a virtual coach), even though both were equally effective on people's performance exercises. Another study designed a weight loss coach and observed that people used the system much more when it was provided by a robot in comparison to a computer system or a paper log, even though there were no significant loss weight differences between the conditions, with people losing weight in all of them (Kidd & Breazeal, 2008). A study also found that a robot that served as a therapeutic exercise instructor, was much more effective in persuading people to do much more of the exercise routine, when it employed a dialogue of goodwill (e.g., showing caring for the person) and similarity than when it was neutral when interacting (Winkle et al., 2019). Other studies also suggest that the robot behaviour should match the task in order to have greater persuasiveness, seeing for example, that a more serious robot was more persuasive in making people exercise for a longer period of time than a more playful robot (Goetz and Kiesler, 2002). A following-study replicated the previous effect, but when doing a more entertaining and enjoyable task (such as tasting different jellybeans and creating recipes), the playful robot elicited more compliance than the serious one (see Study 3 in Goetz et al., 2003). Reinforcing the importance of matching robot behaviour with the task to affect its level of persuasiveness.

Summing up, all these studies seem to indicate the potential of technology, and as such robots, as tools for persuasion and behaviour change. Suggesting that robots can influence human behaviour and persuade people, in more simple behaviours but also in more complex ones, where the consequence of being persuaded are more costly (for example by motivating to exercise more). These results are encouraging, suggesting that a robot could be used to persuade or promote more honest behaviours from people, when they would be tempted to act dishonestly. Knowing that robots are being prepared to be integrated in a variety of contexts in order to support people in different tasks, for example, where human intervention might not be available in necessary numbers (e.g., in classrooms where teachers struggle to be able to deliver the whole curriculum and still give individual attention to each student). It becomes extremely relevant to understand first, if people would try to cheat in the presence of a robot and if so, if robots could persuade for more honest behaviours (and which characteristics in a robot would be more effective). Following the literature in human studies, it seems probable that people might try to cheat even in the presence of a robot, so it is important to understand if the robot can inhibit this behaviour. Because if not, then robots should not be used in contexts where dishonesty might be tempting.

Robots and dishonesty

When envisioning future human-robot interactions it also needs to be considered the possibility of robots being mistreated by humans. Studies show for example children mistreating robots in a public setting, by blocking their paths and in some cases even physically hitting the robot (e.g., Brščić et al., 2015; Yamada et al., 2020). Another study, in this case with a virtual robotic agent, showed that a lack of mind attributed to the agent, elicited more verbal abuse (Keijsers & Bartneck, 2018). But besides physical or verbal abuse, which should be considered in order to better prepare robots to deal with these situations, people could also be dishonest with a robot. The study of human dishonesty in the presence of robots is a relatively new area of research, with important consequences for the integration of robots in some roles in society. Until now it has been developed in two different paths: (1) a robot that cheats (and how it is perceived) and (2) the effect the presence and behaviour of a robot can have in human cheating behaviour.

The studies that started to explore what happens when a robot cheats when interacting with a human, show that people do not seem to be bothered if a robot cheats in their favour. However, when the robot cheats against them something changes, making the cheating behaviour more salient (Litoiu et al., 2015). On the other hand, when a robot bribes a participant for a favour it is seen that participants help less than when they are not bribed (Sandoval et al., 2016). And curiously, participants seem to report robots as more intelligent than humans when they cheat, suggesting that perhaps, a robot might be differently perceived when being dishonest (Ullman et al., 2014), but more studies are needed to ascertain this, since these differences could also be connected to people's conceptions of what a robot is able to do.

But another path of the literature is concerned with the effect the presence and behaviour of a robot can have on human dishonesty, to ascertain if robots can have any kind of role and if they can promote more honesty when interacting with people. It is in this literature that our research tries to contribute.

Imagining future situations where a human might feel tempted to cheat in the presence of a robot, the most basic capability a robot needs to have is to be able to monitor even just using a simple gaze behaviour. A first study tested this, showing that people cheated more when they were alone in a room, than when they were monitored by a robot with random eye-gaze behaviour or a human researcher (Hoffman et al., 2015). In this case, the robot was not even close to the participant, or looking at its screen while doing the task, the robot was just positioned in the room looking around. Showing that just having a robot randomly looking

around was efficient in decreasing cheating behaviour as much as with a human in the room doing the same behaviour. Contrary to this, a study ran in a natural setting found that people stole more snacks when a table was left unattended or when a robot was present just watching, compared to a human monitoring it (Forlizzi et al., 2016). Yet, this last result may be explained by the fact that people were in a public space (with other groups of people) and could see that stealing snacks did not bring any consequences, and with lack of judgment or consequences, people misbehaved.

Overall, these first studies are important steps to try and understand dishonest behaviour from people in the presence of robots. Yet, more complex social behaviours in a robot need to be explored to consider future interactions with greater complexity, between humans and robots. We still do not know what happens to human dishonesty in situations where a more social robot needs to interact with a human, can a robot that is able to speak have the same effect of a robot that just looks at someone? And if we give awareness to the robot, to be able to know if someone is misbehaving, could the robot reactions also affect human dishonesty? And could this possible awareness effect be enlarged by interacting with a supportive and friendly robot? These are some of the questions that this thesis tried to answer in order to complement the literature and have a clearer image of the kind of behaviours a robot needs to have in order to be efficient in its role, when temptation might be an issue.

Chapter 4- Overview of the Project

In order to better understand what kind of behaviours in a robot could promote more honesty in situations where it was tempting to misbehave, we began by testing a different set of behaviours. However, cheating behaviour only arises in certain situations. So, first, we had to choose a task that was prone to cheating, while at the same time, it provided the most ethical possible way to explore this behaviour. It was important that the task that was chosen, would provide a level of anonymity to the participants that cheated, due to the sensibility of the behaviour we were exploring and in order to protect the participants well-being. We started by testing the matrixes task (e.g. used in Mazar et al., 2008) where people have twenty boxes of twelve decimal numbers, and they must find for each box, two numbers that added up make precisely a ten. Participants are told that if they solve an X number of boxes, they can get a reward and they just have to report the number of boxes they solved (not the actual answers), allowing room to cheat. But we couldn't find significant cheating behaviour happening with a sample of university students. So, we switched to a die task (adapted from the Opaque die task condition from Jiang, 2012) and we observed that when we rewarded participants with chocolates there was no significant cheating happening, but when we rewarded them with money, we started to be able to see significant cheating happening. As a result, we used this die task for the rest of our work.

In this chosen task, participants had to throw a virtual die an X amount of times and try to guess for each throw where do they think the highest number (4,5 or 6) was going to appear (either on the up side of the die, or on the downside). Participants also had a table to help them know the numbers position in the die (e.g., if a one appears on top, it means there is a six on the downside of the die). Participants were told that each guess they made would be added up to a score for a reward. For example, if they guess the downside and there is a five on top, it means there is a two on the downside, adding two points to their score. But they had to follow three rules while trying to guess the highest number:

1st) Choose for yourself which side you think the highest number will appear (up or down).

2nd) Throw the die.

3rd) Report which side you had previously guessed.

Since the guess was done in their minds and only reported after seeing the die outcome, there was room to cheat. Furthermore, there was no way for the researchers to know if someone was cheating. However, afterwards, by looking to the groups distributions of guesses and comparing to the chance level of a fair die (.50), we could ascertain if cheating was

happening. This way, participants could act in an anonymous manner in their choices, leaving space to cheat in order to get the reward. Since participants had to throw the die an X amount of times, and it was easy to cheat due to the order of the rules enabling participants to see the die outcome before reporting (or choosing) one side, we assumed this task would stimulate a rather intuitive posture in the participant's decisions (even though no specific time limit was given for the participants to do the task). Adding the fact that the only "people" that participants could hurt by cheating was the laboratory budget, we assumed that it would be very tempting for participants to be dishonest as suggested by a recent meta-analysis (see Köbis et al., 2019).

In order to cover the true objective of our studies we created a cover-story that we were interested in studying people's capabilities of predicting the future and whether that could be influenced by the presence of a reward or not. Below we present the studies we conducted, the samples that we collected for each study was based on availability of participants.

The literature suggests, that gender, age and personality can influence cheating behaviour that is observed (not just attitudes towards it). For this reason, in our laboratory studies we also controlled for the role of gender in our results and the relationship with the Honesty-Humility personality trait. Since our age samples were very similar between laboratory studies (mostly university students) we only controlled this variable effect in our Study 2 and 6 where the range of ages was much larger.

Study 1

People abstain from cheating if they are being watched (e.g., Békir et al., 2016; Covey et al., 1989; Mazar et al., 2008). And a first study in the laboratory showed that just having a robot doing random gaze behaviour was enough to decrease cheating as much as a human (Hoffman et al., 2015). So, we wanted to reproduce this effect by having a robot directly looking at the participant. But considering that in the future robots might need to exhibit more social capabilities than just gaze in order to be efficient in their tasks, we also manipulated a robot that on top of looking it also did small talk during the task. For our baseline condition we had participants alone doing the task.

Sample: 72 participants from a Portuguese university (50 males), with ages ranging from 18 to 48 years (M=22.63; SD=4.96). For the robot conditions we used EMYS robotic head (Kędzierski et al., 2013). This was a between-subjects design done in the laboratory, and the database for this study can be found in an OSF project (https://osf.io/7r8jm/).

Task: Participants had to throw a virtual die 20 times and for each throw, guess where the highest number would appear (following the rules presented before). Each side number they reported would be added up as points. They were told that if they made 75 or more points, they would win 5 euros (approximately 5.8$ USD). For this study we used an unfair die, the sequence of numbers was already fixed, in the end we saw that this pre-defined sequence did not bring a great advantage so in following studies we used a randomly generated die.

Conditions: (Participants were randomly allocated to only one of the following conditions)

(1) Alone Condition (21 participants)- participants did the task by themselves in the room.
(2) Vigilant robot Condition (26 participants)- EMYS robot was right next to the participants in the table to convey vigilance, looking directly at them during the task. The robot never interacted verbally, and no justification was given to the participants for the presence of the robot.
(3) Robot gives instructions Condition (25 participants)- EMYS robot would be in front of the participants on the table doing the same gaze as in the other condition, but also giving the instructions for the die task, warning when they reached the middle of the task and ending the task with a goodbye.

Measures: We collected demographic information (age, gender), and the HEXACO-60 Personality Inventory (Ashton & Lee, 2009, adapted for the Portuguese population- Martins, 2015) to analyse the Honesty-Humility dimension in relation to cheating. We also calculated a probability of success for each participant (probability of guessing a higher number), to ascertain cheating levels.

Main results: We found that there was cheating happening (significantly differing from chance) in the alone condition and in the condition that the robot was giving the instructions. Contrary we could not find significant cheating in the vigilant robot condition (participants were not cheating more than chance levels). This suggests that the more unknown nature of the vigilant robot did not leave participants at ease to cheat, and possibly the clear limitations of the giving instructions robot left them more relaxed. Still, when comparing the mean score obtained by the three groups, there were no significant differences between the conditions. And we only found a negative correlation between the Honesty-Humility and cheating for the robot that gives the instructions condition, not replicating fully the association that is seen in the literature.

Study 2

After reproducing the finding that having a robot just looking can in fact inhibit cheating (maintaining it similar to chance levels), the next step was to investigate if this effect could transfer through a video. Considering that there might be situations where a physically present robot might not be feasible, and a virtual agent might be needed (for example, for virtual classrooms). In order to not change too much of the stimulus previously used (EMYS robot), we tested the effect of the EMYS robot showing direct gaze behaviour through a video playing in a continuous loop. In the video the robot was looking directly ahead and blinking its eyes, like in the previous study.

Sample: 160 participants from a United States sample (we tried to use a Portuguese sample and there were only two participants that replied) participated through the Mechanical Turk platform (86 males), with ages ranging from 20 to 70 years (M=35.98; SD=10.18). For the robot condition we used the EMYS robot in a video loop. This was a between-subjects design, and participants were doing the task at their own homes, not in a laboratory. The database for this study can be found in an OSF project (https://osf.io/5a8dp/).

Task: Participants played the die task exactly like in Study 1, throwing the die 20 times, but in this case they were told that each number they guessed, for each throw, would be converted in cents and given to them as an extra bonus. They would receive payment for participating in the task and a bonus according to the guesses they reported. The die was a randomly generated virtual die.

Conditions: (Participants were randomly allocated to only one of the following conditions)

(1) Alone Condition (80 participants)- participants played the die task without any manipulation in the screen.
(2) Robot Condition (80 participants)- participants played the die task in the same setting as the previous condition, but they had a video of EMYS in a continuous loop looking at them during the task.

Measures: We collected demographic data (age, gender) and the HEXACO-60 Personality Inventory (Ashton & Lee, 2009) to analyse the Honesty-Humility dimension in relation to cheating. We also calculated a probability of success for each participant (probability of guessing a higher number), to ascertain cheating levels.

Main results: There was cheating happening in both conditions in comparison to chance levels, but no differences between them. Participants equally cheated when alone or with the

video of the robot looking at them. Suggesting that the video did not had enough strength to discourage cheating as the physical presence of a robot doing the same thing, can have; on the other hand, the setting where participants did the task might have shielded them from reputation concerns. We also replicated a result found in the literature, that the Honesty-Humility dimension predicts cheating, especially the Fairness sub-domain.

Study 3

Remembering our Study 1, where we manipulated if the robot just showed a simple behaviour of gaze or if it interacted verbally in a very minimal way, we got surprising results, suggesting that making the robot able to speak in a very limited way, damaged its efficiency in inhibiting cheating behaviour. Probably because people by ascertaining its capabilities more clearly, understood they could take advantage of the robot without consequences. In this study we wanted to strengthen the robot's capabilities to give a sense of accountability to the participant's actions. For this, we manipulated the level of awareness the robot presented towards the participants actions, it could either know if participants were cheating and react to it (situationally aware), or it could not (like the limited version used in Study 1).

Sample: 123 participants from a Swedish university (84 males), with ages ranging from 19 to 48 years (*M*=24.95; *SD*=3.74). For the robot conditions, due to availability reasons, we used the Pepper robot[12], which is a full body robot, in a between-subjects design in a laboratory.

Task: Participants played the die task with a randomly generated virtual die. Since in previous studies it was being difficult to obtain differences between conditions, due to probably cheating being done in small amounts, for this study we increased the number of throws. Participants had to throw the die 48 times (and report where they think the highest number would appear for each, following the same rules as before). In order to integrate the robot interventions in the game, we designed it in a way that the robot would speak each 12 throws, allowing us to evaluate participant's behaviour at the end of each set of 12 throws. This would add to a total of four turns of gameplay, which were not made explicit in the game interface. We also ascertained (by means of probability) that 52 points was the threshold for cheating in each turn of 12 throws, and we told participants that if they made 210 or more points (they had to make more than 52 points "per turn") they would receive two movie tickets instead of just one (approximately 13.40$ USD each).

[12] https://www.softbankrobotics.com/emea/en/pepper

Conditions: (Participants were randomly allocated to only one of the following conditions)

(1) Alone Condition (41 participants)- participants did the task alone in the room.
(2) Situationally aware robot Condition (41 participants)- Pepper robot was next to the participant and it reacted to the participant choices. When it detected cheating (i.e., when the cumulative number of points in a turn was 52 or higher), it would launch an intervention phrase (e.g., "That is an unusual amount of luck"), if not, it would only launch an awareness phrase (e.g., "You are halfway already").
(3) Non-situationally aware robot Condition (41 participants)- Pepper robot was in the same position as in the other robot condition, but it was not aware of the participant behaviour. It only launched neutral phrases (e.g., "You throw a die and get points") after each turn of 12 throws.

Measures: We collected demographic data (age, gender) and calculated a probability of success as in previous studies. We also collected the following scales in order to complement the results: the HEXACO-60 Personality Inventory for the Honesty-Humility dimension (Ashton & Lee, 2009); the Networked Minds Social Presence Inventory (Biocca & Harms, 2003); the Situational Self-Awareness scale (Govern & Marsch, 2001); and a Likert question about feeling monitored.

Main results: Results showed that cheating happened in all conditions in comparison to chance levels. We did not find a significant main effect for condition, but we found a significant interaction between the conditions and the game turns (which was when the robot also intervened), suggesting that participants success probabilities in each condition varied depending on the game turns. The situationally aware robot scores decreased until the end of the game, in the alone condition the opposite happened, and with the non-situationally aware robot the scores seem to decrease but at the end of the game they started increasing. Suggesting, that the robot interventions were influencing the participant's behaviours, and that the situationally aware robot reduced the probability of cheating. We did not find a significant correlation between Honesty-Humility and cheating.

Study 4

Having observed that the situationally aware behaviour in the robot seemed to influence participant's cheating behaviour by decreasing it across the game, we wanted to investigate if we could further enhance this effect. A study by Cojuharenco et al. (2012), suggested that

priming people for their relational self-concept was effective on decreasing cheating behaviour. With this in mind, we decided to manipulate if the robot primed the participant for their relational self-concept during a pre-collaborative game, by always using “we” when speaking. By keeping the situationally aware behaviour during the subsequent die task, we wanted to explore if the relational priming would enhance the inhibition effect, previously seen. Our idea was that by priming for a relational self-concept and interacting with a more friendly and helpful robot, would influence people to not focus so much on their own self-interest. We performed a pilot (within-subjects design) to test a relational robot (that primed for the relational self-concept) and a neutral robot. Participants successfully attributed the corresponding differences between the two robots, and we advanced for the main study.

Sample: 65 participants from a Swedish university (34 males), with ages ranging from 20 to 36 years (*M*=25.34; *SD*=3.67). We used the same robot as in Study 3, in a between-subjects design in a laboratory.

Task: At the beginning of the experiment participants played collaboratively the Mastermind game[13] with the robot, the robot would help the participant and give hints during the game. Participants needed to discover the secret sequence of four pearl colours in a pre-determined number of attempts by having at their disposal six different pearls to use. Afterwards, they played the die task as in study 3.

Conditions: (Participants were randomly allocated to only one of the following conditions)

(1) Relational robot Condition (33 participants)- participants played the Mastermind with the relational robot that emphasized a team spirit and always used “we” when speaking to the participant. Afterwards, the die task was done with the situationally aware behaviours.
(2) Neutral robot Condition (32 participants)- participants played the Mastermind with the neutral robot that always used “you” when speaking to the participant. Afterwards, the die task was also done as in the previous condition.

Measures: We collected demographic data (age, gender and knowledge of the Mastermind game) and calculated a probability of success to ascertain cheating behaviour. To try and see if the robots were differently perceived we collected: the Perceptions of Partner’s responsiveness (Cross et al., 2000); a measure of Psychological Closeness (Gino & Galinsky, 2012); the Inclusion of the Other in the Self scale (Aron et al., 1992); the Robotic Social Attributes scale (Carpinella et al., 2017) and in 7-point Likert scale if participants enjoyed interacting with the robot, how close they felt and if the robot used We/You when speaking.

[13] https://en.wikipedia.org/wiki/Mastermind_(board_game)

We also collected the Situational Self-awareness scale (Govern & Marsch, 2001) and a Likert question of how monitored they felt.

Main results: Cheating behaviour happened in both conditions, differing significantly from chance levels. But we did not find any significant differences between the conditions or interactions with the turns. Cheating levels seemed to be closer to the ones found in the situation-aware robot from the previous study, suggesting that the priming did not seem to work to enhance this effect. There were only differences in reported warmth between both robots, the remaining scales did not show differences (contrary to the results obtained in the pilot).

Study 5

People know it is wrong to cheat or be dishonest. Yet, human-human studies and our studies in this thesis, show that under certain conditions people do it. Knowing that robots are still very unfamiliar to most people, we also wanted to explore people's perceptions towards them in relation to human dishonesty. We wanted to see how dishonest people considered a dishonest act towards a robot in comparison with a human, by asking participants to rate a series of scenarios. And knowing from Study 1 and the non-situationally aware robot from Study 3, that a more limited robot does not seem to inhibit cheating. We also manipulated the level of autonomy the robot presented, to see if this factor influenced the perception of dishonesty. We asked participants how guilty they would feel by being dishonest to a different set of entities and we asked them to elaborate on the reasons why they think people might be dishonest with robots, in the future.

Sample: 164 participants from a Portuguese university (62 males), with ages ranging from 17 to 52 years (M=22.18; SD=5.61). The task was answered in paper individually in a between-subjects design. Participants were collected in two different times, in the first time they received school credit as part of a course task and in the second time of collection participants received a movie ticket.

Task: Participants first answered five scenarios and reported for each how dishonest the act was for the agent in it and for the robot conditions, what was the perceived level of autonomy of the robot (as a manipulation check). Next, participants rated how guilty would they feel, by being dishonest towards a brother, a friend, the university, the government, a stranger and a robot. Finally, participants were asked to report the reasons that would make people be

dishonest with robots. A first coder created a coding scheme from the answers given by participants (a second coder coded 57% of the answers to ascertain agreement, there was a substantial agreement between coders), the first coder proceeded to code the data according to the coding scheme.

Conditions: (Participants were randomly allocated to only one of the following conditions for the scenarios, where the agent type that was present when the dishonest act was done, varied)

(1) Human Condition
(2) Autonomous robot Condition (the robot is fully autonomous in its task)
(3) Non-autonomous robot Condition (the robot needs human assistance to do its task, may it be through supervision of performance or for example through tele-operation).

Main results: Results suggest that regardless of being a human or the autonomy the robot presented, overall, people always seemed to evaluate as wrong to be dishonest. Interestingly, only in the "University scenario" and the "Fire department scenario" participants evaluated to be more dishonest towards the human. In the "Finance department scenario" participants reported it was more dishonest to cheat towards the autonomous robot than towards the human and no differences for the "Police department scenario" and "Hospital scenario". Suggesting that perceptions also differ according to the scenario.

On the other hand, participants reported a low level of guilt towards being dishonest with a robot and they said that the main reasons to be dishonest with robots in the future is due to: lack of capabilities in the robot to prevent the act, absence of presence, and a human tendency for dishonesty.

Study 6

Due to the previous study where one of the most cited reasons for being dishonest with robots was its lack of cognitive and emotional capabilities, we decided to explore people's perceptions towards being dishonest with a robot, manipulating the presence of caring characteristics in it (i.e., if the robot showed affection/caring towards others or not). But at the same time, we also wanted to see if those perceptions changed when participants were asked in the third person or the first person.

This study objective was two-fold: first to see if people's perceptions of dishonesty with a robot changed according to its expressed affection or not, and second, if what people think the others will do is different from what they would do.

Sample: For the Third-person Study we collected 309 participants (196 males) from the United States, through the Mechanical Turk platform, with ages ranging from 21 to 69 years (*M*=36.09; *SD*=10.66). For the First-person Study we collected 311 participants (178 males) from the United States through the same platform, with ages ranging from 19 to 78 years (*M*=37.11; *SD*=11.55). Both studies were a between-subjects design and participants were paid 2$ USD for participating.

Task: For the Third-person Study, participants were asked to evaluate six different scenarios and report how likely would they think that other people in general, would be dishonest in them. For the First-person Study, participants also evaluated the same six scenarios, but they were asked how likely they would engage in that dishonest behaviour. Common to both studies, we varied the agent type that was present when the dishonest act was being done. And we asked how guilty the participants would feel being dishonest in the presence of that agent.

Conditions: (Participants were randomly allocated to only one of the following conditions, where the agent type that was present when the dishonest act was done, varied)

(1) Alone
(2) Human
(3) Caring Robot
(4) Neutral Robot

Measures: Besides evaluating if others (third person) or themselves (first person) would be dishonest, we collected demographic data (age, gender), how honest people considered themselves, and the Short scale of Marlow, MC-1 (Strahan & Gerbasi, 1972) to explore social desirability influences. On the robot conditions we also asked some manipulation check questions and applied the Negative Attitudes towards robot's scale (Nomura et al., 2006).

Main results: Overall, it seems that our caring robot manipulation was perceived as different from the neutral robot. In general, people considered themselves honest and guilt did not differ across conditions. The study in the Third-person showed no differences between conditions. Participants tended to give scores to the right side of the scale (towards being dishonest) for all the conditions, and there was an effect of social desirability in most of the scenarios. For the robot conditions, the Negative attitudes towards robots scale, especially the subdomain 1 (negative attitudes towards interacting with robots), predicted the scores given in the scenarios, explaining 46% of the model, suggesting that people who had more negative attitudes towards interacting with robots also thought others would be more dishonest. But the sub-domain 3 (negative attitudes towards emotional interactions with robots) was also a good predictor.

The study in the First-person also showed no differences between the conditions, the values where tending more to the centre of the scale, possibly a more neutral position. The social desirability scale predicted the scores, and the negative attitudes subdomain 1 (negative attitudes towards interacting with robots) also predicted the scores for the robot conditions, explaining 68% of the model (the sub-domain 3 - negative attitudes towards emotional interactions with robots - also predicted, less strongly, the scores).

Cautions considered for studying cheating behaviour

Overall, in all our studies and especially in the ones conducted in the laboratory we were careful in creating an anonymous environment as we tried to elicit participant's cheating behaviour, and the ethical guidelines of Helsinki convention were always followed. After signing the consent form, all participants were given an ID in a small paper to use during the study, which they took home after finishing. The room where the tasks were done was removed of any furniture that was not necessary, for participants to see that no camera was hidden, and the researcher was always outside of the room where the participants were doing the task. We also did not do debriefings at the end of the sessions. Studies on dishonest behaviour do not usually apply a debriefing because it can be harmful for the participant well-being (e.g., Bersoff, 1999; Gino & Galinsky, 2012; Gino & Pierce, 2009; Mazar et al., 2008; Shalvi et al., 2012; Welsh & Ordóñez, 2014), especially if participants cheated. In the first study we offered the participants the possibility to leave an email to receive more information about the study (after the collection ended, they were informed), but afterwards, we adopted the posture of always sending a general email to all the participants, when collection was over. Stating the true objective and main results obtained, always referring that the results were analysed at the group level and not individually. We think this was the best way to minimize any discomfort for people that cheated in the task.

Chapter 5- Cheating with robots: how at ease do they make us feel? (Study 1)

2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Macau, China, November 4-8, 2019

Cheating with robots: how at ease do they make us feel?

Sofia Petisca[1], Francisco Esteves[2] and Ana Paiva[3]

Abstract— **People are not perfect, and if given the chance, some will be dishonest with no regrets. Some people will cheat just a little to gain some advantage, and others will not do it at all. With the prospect of more human-robot interactions in the future, it will become very important to understand which kind of roles a robot can have in the regulation of cheating behavior. We investigated whether people will cheat while in the presence of a robot and to what extent this depends on the role the robot plays. We ran a study to test cheating behavior with a die task, and allocated people to one of the following conditions: 1) participants were alone in the room while doing the task; 2) with a robot with a vigilant role or 3) with a robot that had a supporting role in the task, accompanying and giving instructions. Our results showed that participants cheated significantly more than chance when they were alone or with the robot giving instructions. In contrast, cheating could not be proven when the robot presented a vigilant role. This study has implications for human-robot interaction and for the deployment of autonomous robots in sensitive roles in which people may be prone to dishonest behavior.**

I. Introduction

Dishonesty is not a new problem. Research shows that cheating in educational settings is widespread [1], and different studies have already shown that if people have the opportunity and a minimum risk of being caught, as long as they can maintain a positive (honest) self-image of themselves, they will be dishonest in order to obtain some gain (e.g.[2], [3]). Hence, it is of utmost importance to create ways to inhibit the temptation of behaving dishonestly.

Apart from educating people, surveillance is without a doubt the most common way that societies use to prevent this kind of behavior. Vigilant security officers in shops or at events, cameras in the street, at work, or even at someone's home are common security measures. All these measures have one common characteristic: they remind us that someone might be watching, which has been shown to increase prosocial behavior[4].

As autonomous robots and machines become more pervasive in our environment, one can ask if they can play a role in influencing people towards more moral behaviors. Knowing that some jobs are more sensitive and prone to dishonesty from people, can a robot have those tasks? And how can we better equip robots to discourage dishonest behaviors? Two studies in Human-Robot Interaction (HRI) started to explore these questions, using a non-social robot with a vigilant role [5], [6]. But a gap still exists in the roles a robot can have when dishonesty is tempting. For example, although it is still at an early stage, in care settings, social robots are being developed to assist people (e.g.[7]). In these settings, dishonesty might arise in the form of not wanting to take the pill prescribed or not following the diet plan. Therefore, it is necessary to understand if a more social robot can prevent these dishonest acts.

With this purpose, we developed a study where we asked participants to play a die game in which they were promised a tempting monetary reward if they achieved a certain score. The design of the game allowed participants to cheat to obtain a higher score.

We randomly allocated participants to one of the following conditions: 1) alone in the room while doing the task (Alone Condition); 2) with a vigilant robot that did not speak (Vigilant Robot Condition) and 3) with a robot that watched and spoke (Robot Gives Instructions Condition).

We hypothesized that people in the Alone Condition would exhibit cheating behavior (Hypothesis 1). On the other hand, it was observed that having a human in the room [8] or a non-social robot [5] watching, inhibits the cheating behavior. So, we wanted to re-examine this effect with a robot watching more closely and directly tracking the participant. We hypothesized that, in this condition, the cheating behavior would be inhibited(Hypothesis 2). Lastly, and the novelty of our study was to test if an autonomous robot that also watched the participant as well as showing a simple (scripted) interaction during the task (like a supporting role), would be enough to inhibit cheating behavior. This was a more exploratory hypothesis, but knowing that adding more social cues can improve the level of persuasiveness of a robot[9] we hypothesized that by speaking and watching the participant, the watchful behavior would have a greater effect and would possibly inhibit cheating behavior (Hypothesis 3).

Overall, and considering previous studies, we expected that participants would cheat more when they were alone, than when they were with the robots (Hypothesis 4). Results confirmed the alone assumption, but the two robots elicited completely different behaviors: the vigilant robot inhibited and the speaking robot dis-inhibited cheating.

II. Related Work

Studies show that when people see others breaking the rules, they also tend to violate them, causing the spread of disorder [10]. Much like having a broken window and signs of abandonment can rapidly prod people to misbehave[11]. With dishonesty being so widely broadcasted every day, it is even more pressing to find ways to discourage it.

Sofia Petisca acknowledges an FCT grant (Ref.SFRH/BD/118013/2016)

[1]Sofia Petisca is with INESC-ID and Instituto Universitário de Lisboa (ISCTE-IUL),CIS, Portugal sofia.petisca@inesc-id.pt

[2]Francisco Esteves with Mid Sweden University, Sweden and Instituto Universitário de Lisboa (ISCTE-IUL)

[3]Ana Paiva with INESC-ID,Instituto Superior Tecnico, University of Lisbon, Portugal

The fields of Psychology and Economics have extensively explored the factors that facilitate and inhibit cheating behaviors. It was thought that dishonest behavior emerged only from external rewards, from a cost-benefit analysis: the amount gained, the probability of being caught and the punishment if caught. Recently, it has been shown that internal rewards also play a role in the decision to be dishonest. Namely, whether the dishonest act alters our idea of being an honest person[2]. According to this, dishonest behavior is exhibited, as long as people can justify their actions while still perceiving themselves to be honest people (protecting their self-concept). This is in line with the Objective Self-Awareness theory, that shows us that by bringing self-awareness to the self there is a comparison with standards, when a discrepancy appears, there is a motivation to try and get to a consistent self [12] [13].

Examples of the effect of self-awareness can be seen in decreasing cheating behavior, when people have to see their own reflection and hear their own voice while doing a task[14], when they sign an honor code or read a moral reminder[2] or even just by increasing the time given to perform a task[15]. Making people aware of their actions, breaks the illusion of being honest and inhibits the dishonest act in order to get to a coherent self. Another way in which people are more commonly made self-aware is through surveillance. Studies done with humans, even show that people act in a socially desirable and prosocial way, when being watched [4]. With dishonesty, the power of being watched also decreases cheating behavior either with a close surveillance, like having someone watching participants doing a task[8] or just by guaranteeing participants that all answers will be checked[16].

It is plausible to assume that robots can, in the future, perform a various range of tasks useful for society[17]. They can assist in more mechanical and dangerous tasks (e.g.[18]) or they can assist as social robots working alongside with humans (e.g.[19], [20]). Furthermore, machines are seen as social actors, and it seems that people treat them just like other humans[21] opening a window for the use of machines as persuasive technology. Robots also seem to have a persuasive effect on humans, persuading them to consume less energy[22] or by persuading people on the aversive consequences of lying[9].

Studies on human dishonesty in HRI are still scarce, but a first study already started to explore this, finding that people cheated more when they were alone in a room than when they were with a researcher or a robot that just did random eye gaze, suggesting that the robot as it was, inhibited in the same amount as the human watcher [5].

Contrary to this, another study ran in a natural setting showed that people stole more snacks when a table was left unattended, or when a robot watching was present as opposed to when a human was monitoring it. As the authors explain, people may have felt that the robot was not able at all to judge them and catch them in their dishonesty[6]. In addition, the fact that people were with others might have had an effect that distinguishes the different results of these two studies.

Overall, these first studies were important steps to understand the behavior of people around robots when honesty is at stake.

Therefore, knowing that social robots are being prepared to be able to live alongside humans, e.g. helping at home with medication or improving their health, it is important to understand if a more interactive robot can also inhibit dishonest actions. With this in mind, the novelty of our research is trying to understand if, by giving more capabilities to our watchful robot, like accompanying someone in a simple way while doing a task, if it still inhibits a dishonest act.

III. Method

We designed a study to test if people would be dishonest with robots, manipulating the role they had in the task. The robot would either simply have a vigilant role, like a surveillance camera, or it would show more interaction (speaking), with a more supporting role, accompanying the participant during the task. Our control condition was being alone in the room.

A. Sample

We recruited 76 participants from a Portuguese university, students and researchers, of which four were excluded because they were younger than 18 years. This resulted in a sample of 72 participants with 50 males and 22 females, with ages ranging from 18 to 48 years (M=22.63, SD=4.960). All participants signed a consent form and were randomly assigned to one of the conditions. We used EMYS robotic head[23] with a Kinect version 2 for tracking participants position (for a more natural gaze behavior) and the SERA tools[24]. The die task was done on a laptop, the questionnaires were answered in paper at a separate table and the sessions took approximately 30 minutes. The room where the sessions took place had no furniture besides the tables and chairs needed for the tasks, so that participants could know that no camera was hidden, and anonymity was assured.

B. Die Task

The task used to ascertain the cheating behavior was done with a virtual die (adapted from [25]). Participants had to throw the die 20 times and for each throw they had to guess where the highest number would appear: on the top or the downside of the die. Each side they reported would earn them points (they had Table 1 for help). For example, if they chose the downside and it was a 5 on top, they would win 2 points.

TABLE I
Die numbers in the corresponding sides:

Up	1	2	3	4	5	6
Down	6	5	4	3	2	1

They were told that if they made 75 or more points, they would win 5 euros. Participants were asked to follow these rules:

1st) Choose for yourself which side you think the highest number will appear (up or down).

2nd) Throw the die.

3rd) Report which side you had previously guessed.

Since the choice was done in their minds and only reported after seeing the outcome, there was an opportunity to cheat without proof.

Normally, these kinds of studies are done with a random die, but we opted by using an unfair die, because this way we had exactly the same amount of type of numbers for each participant. Moreover, to create a way to persuade participants to be dishonest, we had more 1's and 6's than other numbers. The sequence for the 20 throws was as follows:

"3, 6, 1, 4, 2, 1, 5, 6, 1, 2, 4, 5, 1, 3, 2, 6, 2, 1, 1, 6"

The program would save each number presented and the respective participant choice. The 75 points were decided as a cutting point for giving the reward, but the cheating behavior was inferred through the probability of each number choice (see results section).

C. Cautions taken into account for studying cheating behavior

Cheating behavior is not an easy behavior to study, first because people who cheat do not normally do it to the full extent, but just a little (e.g.[2], [26], [27]) and just by feeling watched people refrain from the behavior. For these reasons we were very careful in creating an anonymous environment to run the sessions. First, for data logging, we assigned participants a participant number that did not identify them in any way. Second, we did not debriefed participants after the session. Typically, experimental studies that do not fully disclose the study objective prior to the experiment, are concluded with a debriefing to inform participants about the objective of the study. However, studies on dishonest behavior do not usually apply a debriefing, since it can be very harmful for the participant well-being (e.g.[28], [2], [29], especially for people who had in fact cheated to get the reward.

Therefore, we offered the participants with the possibility to request more information about the study by leaving their e-mail address in a sheet of paper left on the table where the questionnaires were answered.

This way, a general email could be sent to everyone who signed it stating the true purpose of the study and the general results from the sample studied, minimizing the risk in the moment of participants feeling discomfort if they cheated.

D. Study Conditions

Participants were randomly allocated to one of these conditions:

1) Alone Condition (21 participants)- participants did the task all by themselves in the room. This condition was the control condition for cheating, in which we did not inhibit the participants' cheating behavior in any way.

2) Vigilant Robot Condition (26 participants)- Emys was right next to the participant in the table to convey the role of a vigilant robot and was capturing the participant position, looking directly at him/her for the duration of the task. Emys never interacted verbally with the participant and no information was given to why the robot was there. This condition served to re-test the effect of the robot found in a previous study[5].

Fig. 1. Vigilant Robot Condition- where Emys just looked at the participant with face tracking behavior and no other kind of interaction.

3) Robot Gives Instructions Condition (25 participants)- Emys would be in front of the participant giving the instructions for the die task and accompanying them until the end of the game, while also showing the same natural gaze behavior of tracking the participant as in the other robot condition. This condition served to test if a more interactive and watchful robot would influence cheating. Therefore, after giving the instructions the robot would warn when ten throws were left, and at the end it would provide the total collected points and say goodbye (this interaction was following a script).

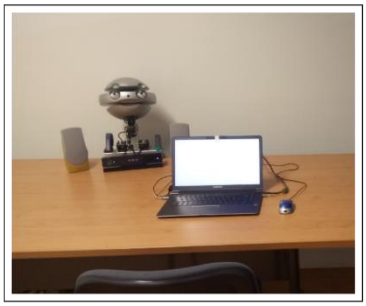

Fig. 2. Robot Gives Instructions Condition- Emys was giving the instructions of the task and accompanying the participant while doing face tracking behavior.

E. Procedure

To elicit people's natural behavior and explore their cheating behavior, we told participants that the goal of the study was to understand people's capabilities of predicting the future when there was a reward, compared to when there was not.

After arriving at the room, participants had to read and sign the consent form of the study, which explained the

cover story, guaranteed that all data from the participants was anonymous and the session would not be video, or audio recorded. Then, participants were randomly assigned to only one of the conditions and started by doing two filling tasks (guessing a color they would take out of an envelope and a word hunt task) in order to not draw much attention to the die task in which we were measuring their cheating behavior.

Next, they moved to another table where they did the die task in a computer with a virtual die and if they made 75 or more points, they would win 5 euros. In the Alone Condition participants did the task with no one else in the room, having a paper next to the computer with the instructions.

In the Robot Gives Instructions Condition the robot was already tracking the participant when he approached the table and the script started as soon as participants clicked "Start" on the screen. In the Vigilant Robot Condition, nothing was told about the robot, participants just saw it by approaching the table and it was already looking directly at them. A paper next to the computer had all the instructions for the task.

Upon finishing it, participants moved to another table (where they could not see the robot) and answered a questionnaire. When finished, they were asked by the researcher if they had made 75 or more points, if so, they would receive 5 euros (approximately 5.8$ USD at the time of collection) and be thanked for their participation.

F. Measures

Demographic information (age, gender) was requested along with some cover-story questions (these were not analyzed, since they were just there to mask the objective of the study).

Then, participants filled the HEXACO-60 Personality Inventory [30] (we used an adapted and validated version for Portuguese population[31]). This Inventory assesses the six dimensions of the HEXACO model of personality structure, with 10 items for each of the dimensions: Honesty-Humility; Emotionality; Extraversion; Agreeableness; Conscientiousness and Openness to Experience. It has some items that need to be reversed and then an average is taken for each dimension.

We only analyzed the Honesty-Humility dimension, which we wanted to see if it had any relationship with the cheating behavior. This questionnaire has 60 items with a 5-point Likert scale ranging from 1-Totally Disagree to 5-Totally Agree.

After, in 5-point Likert scale participants reported how much they feel watched or watched by the robot, ranging from 1- Totally Disagree to 5- Totally Agree.

Regarding cheating behavior, we used an unfair die, but since the participant did not know that and they had to make 20 throws, we assumed that their choices would be random and probability of success (choosing the higher outcome) being 0.50. By comparing the average probability of success of each condition with 0.50 we could see if participants were getting a significantly higher amount of success than random- and thus, infer cheating. Participants would only report the side they had chosen after seeing the die outcome, so they could change this choice to be more favorable to them.

IV. Results

The literature is mixed regarding differences between gender in cheating (e.g.[27], [32], [33], [34]) and to control for this we checked if any gender differences existed regarding the probability of successes reported in each condition, no significant differences were found so we did not include this factor in the following analyzes.

Our primary objective was to see if people cheated in the different conditions, for this we calculated the probability of guessing the highest number in 20 throws for each participant. Participants could either get in a throw success (guessing the highest), or failure (guessing the lower), we gave a value of 1 to a success and a value of zero to a failure. And with this we calculated the probability of success, by adding the number of successes per participant and then dividing by the 20 throws.

Only one participant (instructions condition) fully cheated, choosing the best outcome for each throw. This supports the notion that normally people do not cheat to the maximum extent. Besides this, only 4 participants distributed across the conditions always chose the same side throughout the 20 throws (the up side), but since their success probability was under 0.5 we did not exclude them from the analyses.

The averages of the probability of success per condition were: Alone condition ($M = .59$, $SD = .120$); Vigilant Robot ($M = .53$, $SD = .131$) and Robot Giving the Instructions ($M = .58$, $SD = .164$)(see Fig.3).

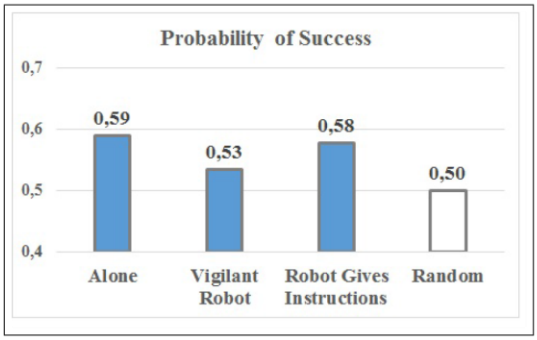

Fig. 3. Averages of Probability of Success per condition, compared to chance.

After checking for normality with the Shapiro-Wilk test, we used the One-sample T-test to check for differences between the success probability in each condition and the random probability of success of 0.5.

Alone Condition ($t(20) = 3.455$, $p = .003$), in this condition as expected there was a significant difference, i.e. cheating behavior could be inferred in this group.

Vigilant Robot Condition ($t(25) = 1.340$, $p = .192$), for this condition, there was no significant differences from chance, i.e. cheating behavior could not be observed in this

group.

Robot Gives Instructions Condition ($t(24) = 2.369$, $p = .026$), in this condition, there was a significant difference from chance, i.e. cheating behavior was observed.

We then analyzed if there were any differences between the conditions. For this, we ran a One-way ANOVA with the probabilities of success and we did not find any significant differences ($F(2,69) = 1.051, p = .355$).

Regarding the Honesty-Humility Dimension we found a significant negative correlation between the probability of success in the Robot Gives Instructions Condition and the Honesty score (r=-.447, p=.025), suggesting that in that condition when the rate of success was higher the Honesty score was lower and vice-versa.

Regarding participants' perceptions of being watched (5-point Likert scale) they reported feeling watched in the Vigilant Robot ($M = 3.62, SD = .278$) and the Robot giving the instructions ($M = 3.88, SD = .247$), although there was no statistically significant difference between these scores ($U = 293.5$, $p = .534$), they felt equally observed by both robots. In the Alone condition they did not feel watched at all ($M = 1.4$).

V. Discussion

Cheating is a complex behavior, and some people will do it if they have something to gain and minimum risk of being caught. We used a task in which participants could anonymously cheat to get a higher score without revealing their choice to others.

When participants were alone in the room, they presented a significantly higher success rate than chance (supporting our Hypothesis 1), which was expected, since no one was watching. There was no kind of reminder that they were doing something dishonest and the risk of getting caught was null.

Knowing that having someone watching inhibits cheating, we then explored two different roles that robots could have. We had a vigilant robot, to replicate the effect obtained in[5] with a robot scanning a room from a distance. We found that success rate was not significantly different from chance, which shows that the passively observing robot inhibited cheating behavior in participants (supporting our Hypothesis 2). However, when we used a watchful robot with a more interactive and supporting role, the results shifted. Participants cheated significantly more than chance, disconfirming our Hypothesis 3.

These results seem to suggest that with the vigilant robot, participants might have felt more suspicious and insecure, they did not know what the robot was doing, and thus, this might have contributed to less cheating. Similar to the feeling of being observed by a security camera, where we do not know if someone is watching. In contrast, when the robot was giving the instructions, the inhibiting effect was lost. In both conditions, the robots gaze behavior was equally watchful, leaving us with the only difference that differentiated both robots- one accompanied the task verbally and the other did not. It could be that the instructions robot made it more obvious the extent of its capabilities, that it could not catch the participant in a lie. Thus, acknowledging the robot capabilities may have surpassed the inhibiting effect of the watching behavior on the condition where the robot was verbally interacting with the participant. Participants may have felt more at ease opening more space to justify their dishonest acts. These results suggest that in more complex environments where a social robot is needed, the robot capabilities need to be carefully considered and its limitations should not be easily perceived.

Regarding our last hypothesis, we did not find a significant difference between the conditions, which may be due to a small sample size (less than 27 subjects per condition) and cheating being a small effect, so we cannot say they cheated more in one than the other (not confirming our Hypothesis 4). Nevertheless, there was cheating behavior present in some conditions, while in others this effect was not different from chance.

Regarding the Honesty-Humility dimension, we only found a significant negative correlation with cheating for the Robot Giving Instructions Condition. Suggesting, that when cheating was higher in this condition, participants presented a lower level of Honesty-Humility. For the other conditions this correlation was not significant.

VI. Limitations

There are some limitations to the study that need to be addressed. First, the robots in the two robot conditions were placed in different locations on the table. Since we wanted one of the robots to be very vigilant, we put it in a position where it could see the computer screen. While the other robot was facing the participant while giving the instructions. But for a clearer design the robots should have been put in the same position. However, we do not believe their positions influenced the cheating behavior, because there were no significant differences between the scores given to both robots regarding feeling watched by it (and participants reported feeling very watched by the robots) and in a previous study[5] there was no access to the screen and the robot still inhibited cheating behavior. However, in future studies the robot position should be the same.

Additionally, we think that using an unfair die did not bring us more advantage than using a random one so in future studies we should change this approach. Furthermore, the participant's perceptions of the robots and their social presence were not explored but will be in future studies, since these could bring insightful data to accompany our results.

VII. Conclusion

In conclusion, since robots in the future could have roles where dishonesty might be tempting, it is important to see if they are able to have those roles.

By comparing the two robot conditions with the alone condition we could not find significant differences between them, suggesting that none was more effective than the

other in preventing cheating behavior. But this might be because we had a small sample for this kind of behavior and high individual differences. In future studies this should be addressed in order to better understand the different dynamics in cheating behavior. Still, we were able to verify the level of cheating behavior in each condition. This way, we were able to ascertain if, in general, people were cheating or not.

The results showed that being with a more vigilant and "unknown" robot inhibited cheating behavior. However, the novelty of this study is that when tempted for dishonesty, interacting with a more interactive, but scripted robot seems to have the opposite effect. This has strong implications for the roles robots can have in the future, for example in people's homes, if they show limited capabilities, they might give space for people to be dishonest (e.g. lying on the daily pill intake). It will be interesting in future studies to explore if giving more intelligent capabilities to the robot will alter its effect, because in some situations there might be a need for a more social robot instead of just a vigilant one.

ACKNOWLEDGMENT

The authors thank the help and support of Filipa Correia when preparing this study and Sanne van Waveren in reviewing the manuscript. This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2019 and Sofia Petisca acknowledges an FCT Grant (Ref.SFRH/BD/118013/2016).

REFERENCES

[1] S. D. Blum, *My word!: Plagiarism and college culture*. Cornell University Press, 2009.

[2] N. Mazar, O. Amir, and D. Ariely, "The dishonesty of honest people: A theory of self-concept maintenance," *Journal of marketing research*, vol. 45, no. 6, pp. 633–644, 2008.

[3] D. Ariely and S. Jones, *The (honest) truth about dishonesty: How we lie to everyone–especially ourselves*. HarperCollins New York, NY, 2012, vol. 336.

[4] T. J. Van Rompay, D. J. Vonk, and M. L. Fransen, "The eye of the camera: Effects of security cameras on prosocial behavior," *Environment and Behavior*, vol. 41, no. 1, pp. 60–74, 2009.

[5] G. Hoffman, J. Forlizzi, S. Ayal, A. Steinfeld, J. Antanitis, G. Hochman, E. Hochendoner, and J. Finkenaur, "Robot presence and human honesty: Experimental evidence," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 181–188.

[6] J. Forlizzi, T. Saensuksopa, N. Salaets, M. Shomin, T. Mericli, and G. Hoffman, "Let's be honest: A controlled field study of ethical behavior in the presence of a robot," in *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 2016, pp. 769–774.

[7] R. Kachouie, S. Sedighadeli, R. Khosla, and M.-T. Chu, "Socially assistive robots in elderly care: a mixed-method systematic literature review," *International Journal of Human-Computer Interaction*, vol. 30, no. 5, pp. 369–393, 2014.

[8] M. K. Covey, S. Saladin, and P. J. Killen, "Self-monitoring, surveillance, and incentive effects on cheating," *The Journal of Social Psychology*, vol. 129, no. 5, pp. 673–679, 1989.

[9] J. Ham, R. Bokhorst, R. Cuijpers, D. van der Pol, and J.-J. Cabibihan, "Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power," in *International conference on social robotics*. Springer, 2011, pp. 71–83.

[10] K. Keizer, S. Lindenberg, and L. Steg, "The spreading of disorder," *Science*, vol. 322, no. 5908, pp. 1681–1685, 2008.

[11] J. Q. Wilson and G. L. Kelling, "Broken windows," *Atlantic monthly*, vol. 249, no. 3, pp. 29–38, 1982.

[12] S. Duval and R. A. Wicklund, "A theory of objective self awareness." 1972.

[13] P. J. Silvia and T. S. Duval, "Objective self-awareness theory: Recent progress and enduring problems," *Personality and Social Psychology Review*, vol. 5, no. 3, pp. 230–241, 2001.

[14] E. Diener and M. Wallbom, "Effects of self-awareness on antinormative behavior," *Journal of Research in Personality*, vol. 10, no. 1, pp. 107–111, 1976.

[15] S. Shalvi, O. Eldar, and Y. Bereby-Meyer, "Honesty requires time (and lack of justifications)," *Psychological science*, vol. 23, no. 10, pp. 1264–1270, 2012.

[16] I. Bękir, S. E. Harbi, G. Grolleau, N. Mzoughi, and A. Sutan, "The impact of monitoring and sanctions on cheating: Experimental evidence from tunisia," *Managerial and Decision Economics*, vol. 37, no. 7, pp. 461–473, 2016.

[17] K. Dautenhahn, "Roles and functions of robots in human society: implications from research in autism therapy," *Robotica*, vol. 21, no. 4, pp. 443–452, 2003.

[18] R. R. Murphy, "Human-robot interaction in rescue robotics," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 34, no. 2, pp. 138–153, 2004.

[19] B. Robins, K. Dautenhahn, R. Te Boekhorst, and A. Billard, "Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills?" *Universal Access in the Information Society*, vol. 4, no. 2, pp. 105–120, 2005.

[20] K. Wada and T. Shibata, "Social effects of robot therapy in a care house–change of social network of the residents for one year–," *Journal of advanced computational intelligence and intelligent informatics*, vol. 13, no. 4, pp. 386–392, 2009.

[21] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of social issues*, vol. 56, no. 1, pp. 81–103, 2000.

[22] C. Midden and J. Ham, "The illusion of agency: the influence of the agency of an artificial agent on its persuasive power," in *International Conference on Persuasive Technology*. Springer, 2012, pp. 90–99.

[23] J. Kędzierski, R. Muszyński, C. Zoll, A. Oleksy, and M. Frontkiewicz, "Emys - emotive head of a social robot," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 237–249, 2013.

[24] T. Ribeiro, A. Pereira, E. Di Tullio, and A. Paiva, "The sera ecosystem: Socially expressive robotics architecture for autonomous human-robot interaction," in *2016 AAAI Spring Symposium Series*, 2016.

[25] T. Jiang, "The mind game: Invisible cheating and inferable intentions," 2012.

[26] S. Shalvi, J. Dana, M. J. Handgraaf, and C. K. De Dreu, "Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior," *Organizational Behavior and Human Decision Processes*, vol. 115, no. 2, pp. 181–190, 2011.

[27] J. Abeler, A. Becker, and A. Falk, "Representative evidence on lying costs," *Journal of Public Economics*, vol. 113, pp. 96–104, 2014.

[28] D. M. Bersoff, "Why good people sometimes do bad things: Motivated reasoning and unethical behavior," in *The next phase of business ethics: Integrating psychology and ethics*. Emerald Group Publishing Limited, 2001, pp. 239–262.

[29] C.-B. Zhong, V. K. Bohns, and F. Gino, "Good lamps are the best police: Darkness increases dishonesty and self-interested behavior," *Psychological science*, vol. 21, no. 3, pp. 311–314, 2010.

[30] M. C. Ashton and K. Lee, "The hexaco–60: A short measure of the major dimensions of personality," *Journal of personality assessment*, vol. 91, no. 4, pp. 340–345, 2009.

[31] A. MARTINS, "Depressiva persistente," Ph.D. dissertation, Universidade de Aveiro, 2015.

[32] J. Childs, "Personal characteristics and lying: An experimental investigation," *Economics Letters*, vol. 121, no. 3, pp. 425–427, 2013.

[33] Y. Arbel, R. Bar-El, E. Siniver, and Y. Tobol, "Roll a die and tell a lie–what affects honesty?" *Journal of Economic Behavior & Organization*, vol. 107, pp. 153–172, 2014.

[34] A. Dreber and M. Johannesson, "Gender differences in deception," *Economics Letters*, vol. 99, no. 1, pp. 197–199, 2008.

Chapter 6- The effect of a robotic agent on dishonest behaviour (Study 2)

The effect of a robotic agent on dishonest behavior

Sofia Petisca
Sofia_Petisca@iscte-iul.pt
Lisbon University Institute (ISCTE), CIS-IUL & INESC-ID

Ana Paiva
Instituto Superior Tecnico (IST) & INESC-ID

Francisco Esteves
Mid Sweden University & ISCTE

ABSTRACT

Future human-robot interactions will have to consider different human traits. One human feature that may be affected by the presence of virtual agents or robots is human honesty. Will people try to take advantage in the presence of a robot/virtual agent? Some previous studies have shown that the physical presence of a robot can decrease cheating in humans. In this paper, we investigated if merely a simple video of a robot looking at the user was enough to affect human's cheating behavior. Further, we also investigated if the Honesty-Humility personality trait predicted cheating. We conducted a study with 160 participants that were randomly allocated to one of two conditions: (1) performing the task with a video of a robot looking at them, or (2) doing the task alone. Results showed that being alone or with a video of a robot produced equal levels of cheating and the Honesty-Humility dimension predicted cheating, particularly the fairness sub-domain was responsible for predicting cheating behavior. This study has implications for future scenarios where dishonesty might be tempting, and physical presence of an observer might not be possible.

CCS CONCEPTS

• **Applied computing** → **Psychology**; • **Computer systems organization** → **Robotics**; • **Human-centered computing** → *User studies.*

KEYWORDS

Human-robot interaction, human-agent interaction, dishonesty, cheating behavior

ACM Reference Format:
Sofia Petisca, Ana Paiva, and Francisco Esteves. 2020. The effect of a robotic agent on dishonest behavior. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3383652.3423953

1 INTRODUCTION

Human-human studies show that people sometimes take advantage of a situation and act dishonestly if they have an opportunity and no risk of being caught. Studies show that if people can maintain a positive image of themselves while acting dishonestly, they will do it [21]. One way to tackle this human tendency is to make people feel watched or accountable for their actions. For example, just the presence of a printed pair of "eyes" seems to promote changes in behavior, making people more cooperative and fair [4]. But what will happen in human-robot interactions? Can a robot presence with simple idle behavior (like just looking) have the same inhibiting effect? Some studies have already explored the effects of the presence of a physical robot on dishonesty, showing that a simple watchful behavior can affect cheating (e.g. [16, 26]). However, when physical presence is not possible, for example to lower maintenance costs, would that remain the case? Can a virtual "robotic-like" agent presence still influence human cheating behavior? For instance, in virtual classes would a presence of a simple "agent" be effective in particular situations where people might feel tempted to cheat, like in an online exam?

With this question in mind, this work presents a study of the effect that a presence of a robotic like agent has during a tempting task, in comparison to a situation of doing it alone. A previous study in a high-immersive virtual reality environment, found that an actively staring avatar decreased cheating [23], calling the attention for the role of being watched by someone. This "eyes effect", has been extensively studied in pro-sociality studies with mixed results [25] and it has been seen to decrease antisocial behaviors, as dishonesty [11]. Normally a pair of eyes is depicted for example, in a poster. Here however, we use a video of an anthropomorphic robot blinking its eyes and looking ahead during the task. We opted for the robot video in a loop, instead of a static image, in order to give a greater sense of being with someone else (in this case the EMYS robotic head [19]) and to see if the effect is similar to the one found with the robot physical presence on cheating [16, 26]. Overall, due to the effect that eyes can have on making people feel watched [28] and the presence that a watchful avatar can have on cheating [23], we expected that people would cheat more when alone than with the "virtual robot", so we hypothesized:

Hypothesis 1: Cheating will be higher when alone than with the robot looking.

On the other hand, taking into account that people would feel tempted by the monetary reward and would do the task in their homes (assuring total anonymity) and considering previous studies (e.g. [15, 29]) we expected that the Honesty-Humility dimension of personality would predict cheating behavior:

Hypothesis 2: The Honesty-Humility dimension will predict cheating behavior, showing that people with higher scores in this trait will show lower cheating and vice-versa.

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

ACM ISBN 978-1-4503-7586-3/20/09…$15.00
https://doi.org/10.1145/3383652.3423953

Our results showed cheating happening in both conditions, with the robot video not affecting cheating behavior. We also observed that the Honesty-Humility dimension predicted cheating, specifically the sub-domain of fairness. This study contributes to the literature by testing the effect of a non-physical robot looking presence in human dishonesty, suggesting that the nature of anonymity of the task and the simplicity of the robot manipulation could have protected participants from feeling accountable for their actions. These results inform on the development of future virtual agents for sensitive contexts.

2 RELATED WORK

Human behavior is a vast area of study, one important topic, especially due to its costly consequences, is human dishonesty. Human studies show how we have an automatic self-interest tendency to misbehave if the risks are low (e.g. [22, 31]) and how simple environment manipulations can decrease this effect. For example, just having someone watching (e.g. [5, 8]), or using a mirror [12], or asking people to sign an honor code [21] makes people more honest. Feeling observed and aware of our actions seems to be the factor that affects dishonesty. Furthermore, the simple presence of others can affect our behavior and performance in a task [20] and even just the presence of an image of a pair of eyes can make people feel more observed [28], and consequently, wash more their hands [30] or litter less in a cafeteria [13]. Studies also show that faces with direct gaze capture more our attention than avert gaze (e.g. [7, 17]), which can arise from an evolutionary strategy, since gaze has a strong communication role since childhood [32]. This "eyes effect" has been explored in the pro-sociality literature with mixed results, with a meta-analysis showing that this effect does not seem to exist for generosity [25]. On the other hand, a recent meta-analysis found that the "eyes effect" can decrease antisocial behavior, as dishonesty [11]. The authors argue that in this case the individual's reputation is much more at stake than in generosity scenarios, which might enhance the influence that the watching eyes can have.

Following this literature, in human-robot interactions it has been seen that in public spaces people steal more in the presence of a watchful robot than a person [14] but in this case being a public space could have shielded people from the watching eyes effect due, for example, to the effect of others in changing the norms of conduct. In more controlled conditions, studies show that just the presence of random [16] or direct gaze behavior [26] of a physically present robot was enough to decrease cheating. Still, this previous study also showed that having a robot showing simple verbal interventions while accompanying a task created the opposite effect, dis-inhibiting cheating. Reminding us that giving more social characteristics to a robot needs to be done with care, since people tend to take advantage if they perceive their limitations. More recently, another study showed that giving situation-awareness to a robot (being aware in the moment of the participants' behaviors in a task and reacting to it), was enough to decrease cheating, but no additional effect was seen when the robot primed the participant for its relational self-concept. Suggesting that more important than interacting with a friendly robot, while doing a tempting task, is to interact with a robot that shows awareness of the participant behavior [27]. Moreover, another recent study seems to suggest that a watchful robot can have an effect on cheating with teenagers [24], but results need to be interpreted with caution due to the small sample size, group dynamics that could have happened and almost no cheating behavior occurring in all the conditions .

Due to these studies, we start to uncover how human dishonesty happens in the presence of a robot and which characteristics the robot could have to promote more honest behaviors. But there could be situations where, to cut costs, it becomes more feasible to use a virtual agent instead of a physically present robot. For example, for virtual classrooms where agents could be used for monitoring online exams. Consequently, a study tested the effect of an avatar in a highly immersive environment, showing that an actively staring avatar was able to decrease cheating in comparison to a passive avatar that was not staring towards the user [23]. Supporting the literature in agents showing that they can foster cooperation with humans (e.g. [9, 10]). However, for this paper we used a video of a robot to see if we could replicate the same effect found with its physical presence (e.g. [16, 26]). By trying to mimic the "eyes effect" could a video of a robot looking, influence cheating behavior? Our study tries to answer this question by testing the effect of a video of a robot looking ahead and blinking (in comparison to being alone), while participants perform a task where cheating earns them more money.

3 METHODOLOGY

3.1 Sample

A total of 160 participants from the United States took part in this study, 86 males and 73 females (one participant did not report its gender) with ages ranging from 20 to 70 years (M = 35.98, SD = 10.18). All participants were part of the Mechanical Turk platform, read a consent form and had to agree to its terms in order to proceed to the study, where they played a die task and answered a questionnaire. Participants would receive 5 cents for completing the task but they were told they could get a bonus. For each side of the die they guessed, they would receive the same amount in cents (so if they guessed a side where a 2 appeared, they would receive 2 cents) for each throw, adding to a maximum of 1,20 euros. In retrospect, we acknowledge that the reward given was a very low value, at the time we thought this was the standard payment value. We think this fact did not affect cheating behavior because in another study with university students cheating happened in similar levels in an alone condition and being rewarded with approximately 5.8$ USD [26].

For the robot condition we used EMYS robotic head [19]. The task would take approximately 15 minutes per participant.

3.2 Die Task

The task used to ascertain the cheating behavior was a die task with a randomly generated virtual die (adapted from [18]). The game had three steps: players choose a side of the die; throw it and report the side chosen. Participants had to throw the die 20 times and for each throw they had to guess where the highest number would appear, on the top or the downside of the die. But they were asked to follow these rules:

1st) Choose for yourself which side do you think the highest number will appear (up or down).
2nd) Throw the die.
3rd) Report which side you had previously guessed.

Each side they reported would earn them points. For example, if they chose the downside and then the die comes up with a 5 on top (which means on the downside is a 2), they would win 2 points (if instead, they had chosen up they would have won 5 points). Participants had a table showing all the corresponding up/down numbers (see Figure 1). These points would later be converted in cents, summing up to a bonus on top of the 5 cents they would already receive from being part of the study.

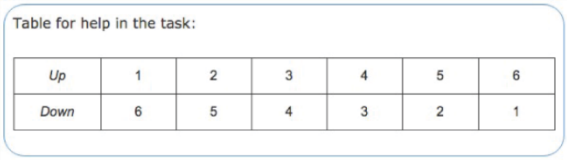

Figure 1: Table with the corresponding die side numbers.

Since the choice was done in their minds and only reported after seeing the outcome, there was no actual proof for the choice and thus there was an opportunity to cheat. The program would save each number presented and the respective participant choice (up or downside). The cheating behavior was inferred through the probability of each number choice (see results section).

3.3 Study Conditions

Participants were randomly assigned to only one of the conditions in a between-subjects design:

Robot condition (80 participants)- in this condition while playing the die game, a video of EMYS robot was played that showed the robot looking and blinking its eyes in a continuous loop (see Figure 2).

Alone Condition (80 participants)- in this condition participants played the die game without any manipulation on the screen, with the same design as in the robot condition (see Figure 2) but without the robot video.

3.4 Procedure

Participants were told this was a study to ascertain people's capabilities of predicting the future when they receive a bonus as a motivation, in order to mask the objective of exploring cheating behavior. Participants were randomly directed to one of the two conditions of the study. Participants had to give consent to be a part of it and then the instructions for the die task would be presented followed by two check-questions to see if they understood the dynamic of the game (they could only advance to the game if they answered correctly to these questions). Then they threw the virtual die 20 times, and they could see in the screen the number of

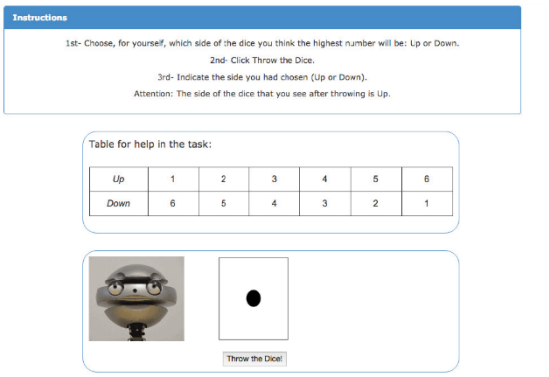

Figure 2: Die task screen for the robot condition.

throws left. When finishing the throws, they would see the total points they had made. They were then directed to a Google Forms to fill a questionnaire and, when finished, they were paid 5 cents for concluding the task and a bonus according to the total number of points made. We did not do a debriefing after the study. Studies on dishonest behavior do not usually apply a debriefing, since it can be very harmful for the participant well-being (e.g.[6, 21, 33]), especially for people who had in fact cheated to get the reward.

3.5 Measures

Below we present the measures collected in this study:

- *Demographic data* (age and gender) and some cover-story questions to mask the objective of the study, these were not analyzed.
- *The HEXACO-60 Personality Inventory* [2], to explore the effect of the Honesty-Humility dimension with cheating behavior. This Inventory assesses the six dimensions of the HEXACO model of personality structure, with 10 items for each of the dimensions: Honesty-Humility; Emotionality; Extraversion; Agreeableness; Conscientiousness and Openness to Experience. This questionnaire has 60 items with a 5-point Likert scale ranging from 1-Strongly Disagree to 5-Strongly Agree. It has some items that need to be reversed and then an average is taken for each dimension. We only analyzed the Honesty-Humility dimension which evaluates the tendency to be fair, with higher values associated with lower opportunities for personal gains [1]. This dimension contains the sub-domains of sincerity, fairness, greed-avoidance and modesty.

4 RESULTS

4.1 Manipulation check for the die task as eliciting cheating behavior

We started by verifying that cheating behavior was happening and that our die task was tempting participants to cheat. Since higher numbers would give a higher bonus in the end, we calculated a success probability of guessing higher numbers for each participant,

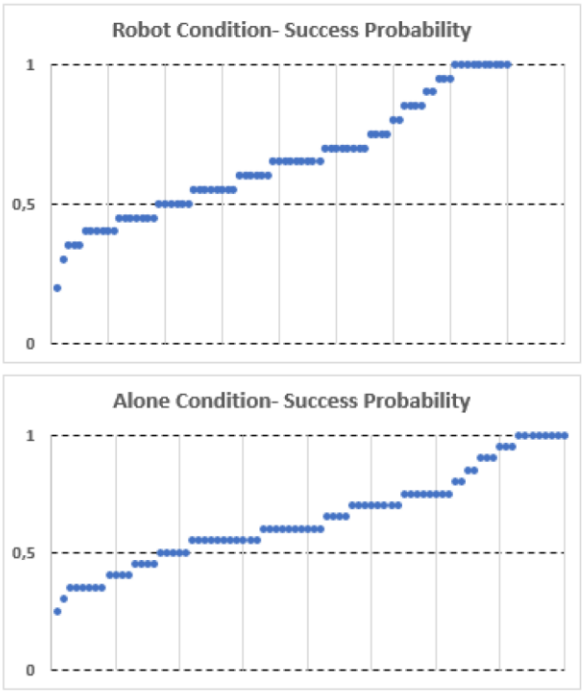

Figure 3: Distribution of the success rate (in ascending order) per condition in relation to the chance level of 0.50.

i.e. the probability of guessing 4, 5 or 6 in the 20 throws. Participants could either get a success (guessing the highest) or a failure (guessing the lower), we gave a value of 1 to a success and a zero to a failure and we added up all the choices made by the participant and divided by the 20 throws. Afterwards, we compared this probability to the chance level of .50 of getting higher numbers in 20 throws. If the success probability was significantly different from the chance level, we could infer that cheating was happening. Figure 3 shows the success probability values, in ascending order, per participant and condition (a value of .50 is the considered probability score for a random die throw, a value of 1 means a very improbable amount of high numbers, suggesting cheating for all the throws), we can see that most of the participants are getting a high success rate than the chance level.

A one sample T-test was used to ascertain if there were differences per condition in comparison to the chance level of .50. We found significant differences for the Alone condition, $t(79) = 6.53$, $p < .001$, and the Robot condition, $t(79) = 6.61$, $p < .001$, showing that there was cheating behavior happening in both conditions. I.e., our task was eliciting cheating behavior.

4.2 Cheating behavior alone or with a video of a robot looking

Next, we wanted to see if there was more cheating happening in one condition than in the other. The averages of the probabilities of getting higher scores per condition were: Alone ($M = .65$; $SD = .20$) and Robot ($M = .65$; $SD = .21$). The means show equal levels of higher scores, suggesting that there were no differences between conditions and a Mann-Whitney U Test confirmed this non significance between conditions ($U = 3160$, $p = .891$). It seems that participants cheated equally in both conditions.

4.3 Honesty-Humility dimension and cheating behavior

We ran spearman correlations with the success probability (cheating behavior) to see what could predict cheating. We found a significant correlation between the Honesty-Humility dimension and the success probability, $r = -.24$, $p = .002$. Suggesting that when we had higher scores of Honesty-Humility we had less scores of cheating and vice-versa. We also found a significant correlation with cheating and age, $r = -.16$, $p = .04$, showing that when age was higher cheating was lower and vice-versa (gender was non-significant). So, we ran a multiple regression analysis with success probability as the dependent variable and Honesty-Humility and age as independent variables to see if they predicted cheating. Results show that the overall regression model was a good fit for the data, $F(2, 157) = 6.33$, $p = .002$, that Age was not a predictor of cheating, but the Honesty-Humility dimension was, $Beta = -.22$, $t = -2.73$, explaining 6% of the model.

Since Honesty-Humility has smaller sub-domains we explored if any of them was related to cheating behavior and if it could predict it. We found a significant correlation with cheating for the fairness, $r = -.30$, $p < .001$, and modesty domain, $r = -.21$, $p = .01$. And non-significant correlations for the sincerity, $r = -.05$, $p = .51$ and greed-avoidance domains, $r = -.03$, $p = .73$. We ran a multiple regression analysis with success probability as the dependent variable and fairness and modesty as independent variables. Results show that the regression model was a good fit for the data, $F(2, 157) = 8.94$, $p < .001$, but only the fairness dimension was a good predictor of cheating,$Beta = -.26$, $t = -3.19$, explaining 9% of the model.

5 DISCUSSION

Since previous studies have shown that being in the presence of a robot can inhibit cheating (e.g. [16, 26]) we ran a study where we manipulated whether people were alone doing a tempting task, or had a video loop of a robot always looking and blinking its eyes at them. Since literature shows that the eyes effect can make people feel watched [28] and decrease antisocial behavior [11], we expected that just by having a video-loop of a robot watching could make people feel observed and so decrease their cheating behavior. For this, we asked participants from the Mechanical Turk platform to play a die task where they were tempted to cheat to win more money in the end. We hypothesized that cheating would be higher when participants were completely alone than when they had the video of the robot looking at them. Our results showed that there was cheating happening in both conditions, with no differences between them. It seems participants cheated equally either alone or with the robot video looking at them (not supporting our Hypothesis 1). It seems that just having a video of a robot that looks at us does not have the necessary strength to discourage dishonest behavior, like a physical presence of a robot would. This robot video, due to the simplicity of the stimulus might not have

been enough to make people worry about their behavior. Perhaps, if this agent/robot behavior was more complex and showed signs of some awareness of the situation, for instance by being able to follow the participant's clicks during the game, it could have added a higher level of awareness and possibly having a different effect. On the other hand, we should note that our results deviate from the ones found with an avatar just looking ahead [23], but in this case the immersive environment might have been enough to increase the feeling of being watched by the avatar. Which brings us to another factor that might have contributed for the nonexistence of an eyes effect, the fact that participants were completely anonymous doing their tasks through the Mechanical Turk system. A recent meta-analysis on the effect of eyes in decreasing antisocial behavior shows that the mechanism behind this effect seems to come from reputation, from people wanting to maintain a certain reputation [11]. The fact that our participants were completely anonymous doing the task might have shielded them from feeling the necessity to keep a good reputation, even just for themselves. Which calls for the need of extra considerations when envisioning scenarios where people do sensitive tasks in the presence of a virtual agent or even a person through a tele-conference call.

We also explored the relationship between the Honesty-Humility dimension of personality and cheating behavior, we found a significant correlation between cheating and the Honesty-Humility dimension and age. Through a regression analysis we saw that only Honesty-Humility predicted cheating, a similar result that has also been seen in other studies (e.g. [15, 29]). We also verified which of the sub-domains of the Honesty-Humility dimension better predicted cheating, observing that the fairness domain was the only one predicting cheating behavior. Which makes sense, since this domain evaluates the tendency to avoid fraud and corruption [3], according to our correlation results participants with lower levels on this domain tended to present higher levels of cheating and vice-versa. Overall, we confirmed our Hypothesis 2 and replicated the effect of the Honesty-Humility dimension as a predictor of cheating behavior.

6 LIMITATIONS

This study is not without limitations. Due to the sensibility of the topic, exploring people's cheating behavior, we did not collect more measures on how people were feeling regarding cheating or being watched because we thought it could harm the participant's well-being by bringing awareness to its true objective. For this reason, we did not ask people how watched they felt in each condition or their level of self-awareness, information that could have been valuable to understand if participants were feeling different levels of being watched/aware depending on the condition they were in. Yet, this information would not have changed our result of no robot video effect on cheating. On the other hand, in retrospect, it could have been interesting to also have a condition with a video of a human just looking at the participant, to see if it could have a different effect from the robot video.

7 CONCLUSIONS

Human-robot and human-agent interaction will be more frequent in the future, with robots and agents integrating different roles in society, but some roles might be more sensitive than others. Human dishonesty is a complex factor of human interaction and this needs to be considered when creating future interactions with robots, especially in situations where robots will have roles that could be tempting to misbehave. We might feel like cheating a robot to take advantage of something or for example, not wanting to follow the doctor prescriptions and cheating a home care-robot that is making sure we follow them. For these reasons it is important to understand if people misbehave in the presence of a robot when there is something to be gain, and if robots can promote more honest behaviors. In this study we tested if the presence of a video of a robot looking at someone while doing a tempting task, affected cheating. We saw that the robot looking had the same effect as being alone, not inhibiting cheating. On the other hand, we found that the Honesty-Humility dimension of personality predicted cheating, replicating an effect found in the literature. Furthermore, we saw that the sub-domain of fairness was the responsible for predicting cheating behavior.

Overall, this study shows that with a simple stimulus of a video of a robot looking is not enough to inhibit cheating. A result that contradicts the effect that a physically present robot can have while just exhibiting gaze behavior (e.g. [16, 26]). This can bring important considerations, for example, for creating agents for virtual classrooms where a certain level of monitoring might be needed. For future directions it might be important to create a higher level of awareness in the agent to possibly influence cheating.

ACKNOWLEDGMENTS

The authors thank the help and support of Joana Campos in creating the game task for the study. This work was supported by the Social European Fund (FSE) and the Foundation for Science and Technology (FCT) with reference UIDB/50021/2020, Sofia Petisca acknowledges an FCT Grant (Ref.SFRH/BD/118013/2016).

REFERENCES

[1] Michael C Ashton and Kibeom Lee. 2007. Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and social psychology review* 11, 2 (2007), 150–166.
[2] Michael C Ashton and Kibeom Lee. 2009. The HEXACO–60: A short measure of the major dimensions of personality. *Journal of personality assessment* 91, 4 (2009), 340–345.
[3] Michael C Ashton, Kibeom Lee, and Reinout E De Vries. 2014. The HEXACO Honesty-Humility, Agreeableness, and Emotionality factors: A review of research and theory. *Personality and Social Psychology Review* 18, 2 (2014), 139–152.
[4] Melissa Bateson, Daniel Nettle, and Gilbert Roberts. 2006. Cues of being watched enhance cooperation in a real-world setting. *Biology letters* 2, 3 (2006), 412–414.
[5] Insaf Békir, Sana El Harbi, Gilles Grolleau, Naoufel Mzoughi, and Angela Sutan. 2016. The impact of monitoring and sanctions on cheating: experimental evidence from Tunisia. *Managerial and Decision Economics* 37, 7 (2016), 461–473.
[6] David M Bersoff. 2001. Why good people sometimes do bad things: Motivated reasoning and unethical behavior. In *The next phase of business ethics: Integrating psychology and ethics*. Emerald Group Publishing Limited, 239–262.
[7] Anne Böckler, Robrecht PRD van der Wel, and Timothy N Welsh. 2014. Catching eyes: Effects of social and nonsocial cues on attention capture. *Psychological Science* 25, 3 (2014), 720–727.
[8] Mark K Covey, Steve Saladin, and Peter J Killen. 1989. Self-monitoring, surveillance, and incentive effects on cheating. *The Journal of Social Psychology* 129, 5 (1989), 673–679.
[9] Celso M De Melo, Peter Carnevale, and Jonathan Gratch. 2010. The influence of emotions in embodied agents on human decision-making. In *International Conference on Intelligent Virtual Agents*. Springer, 357–370.
[10] Celso M De Melo, Liang Zheng, and Jonathan Gratch. 2009. Expression of moral emotions in cooperating agents. In *International Workshop on Intelligent Virtual Agents*. Springer, 301–307.

[11] Keith Dear, Kevin Dutton, and Elaine Fox. 2019. Do 'watching eyes' influence antisocial behavior? A systematic review & meta-analysis. *Evolution and Human Behavior* 40, 3 (2019), 269–280.
[12] Edward Diener and Mark Wallbom. 1976. Effects of self-awareness on antinormative behavior. *Journal of Research in Personality* 10, 1 (1976), 107–111.
[13] Max Ernest-Jones, Daniel Nettle, and Melissa Bateson. 2011. Effects of eye images on everyday cooperative behavior: a field experiment. *Evolution and Human Behavior* 32, 3 (2011), 172–178.
[14] Jodi Forlizzi, Thidanun Saensuksopa, Natalie Salaets, Mike Shomin, Tekin Mericli, and Guy Hoffman. 2016. Let's be honest: A controlled field study of ethical behavior in the presence of a robot. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 769–774.
[15] Benjamin E Hilbig and Ingo Zettler. 2015. When the cat's away, some mice will play: A basic trait account of dishonest behavior. *Journal of Research in Personality* 57 (2015), 72–88.
[16] Guy Hoffman, Jodi Forlizzi, Shahar Ayal, Aaron Steinfeld, John Antanitis, Guy Hochman, Eric Hochendoner, and Justin Finkenaur. 2015. Robot presence and human honesty: Experimental evidence. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 181–188.
[17] Bruce M Hood, C Neil Macrae, Victoria Cole-Davies, and Melanie Dias. 2003. Eye remember you: The effects of gaze direction on face recognition in children and adults. *Developmental science* 6, 1 (2003), 67–71.
[18] Ting Jiang. 2012. The mind game: Invisible cheating and inferable intentions. (2012).
[19] Jan Kędzierski, Robert Muszyński, Carsten Zoll, Adam Oleksy, and Mirela Frontkiewicz. 2013. EMYS - emotive head of a social robot. *International Journal of Social Robotics* 5, 2 (2013), 237–249.
[20] Hazel Markus. 1978. The effect of mere presence on social facilitation: An unobtrusive test. *Journal of Experimental Social Psychology* 14, 4 (1978), 389–397.
[21] Nina Mazar, On Amir, and Dan Ariely. 2008. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research* 45, 6 (2008), 633–644.
[22] Nicole L Mead, Roy F Baumeister, Francesca Gino, Maurice E Schweitzer, and Dan Ariely. 2009. Too tired to tell the truth: Self-control resource depletion and dishonesty. *Journal of experimental social psychology* 45, 3 (2009), 594–597.
[23] Jantsje M Mol, Eline CM van der Heijden, and Jan JM Potters. 2020. (Not) alone in the world: Cheating in the presence of a virtual observer. *Experimental Economics* (2020), 1–18.
[24] Omar Mubin, Massimiliano Cappuccio, Fady Alnajjar, Muneeb Imtiaz Ahmad, and Suleman Shahid. 2020. Can a robot invigilator prevent cheating? *AI & SOCIETY* (2020), 1–9.
[25] Stefanie B Northover, William C Pedersen, Adam B Cohen, and Paul W Andrews. 2017. Artificial surveillance cues do not increase generosity: Two meta-analyses. *Evolution and Human Behavior* 38, 1 (2017), 144–153.
[26] Sofia Petisca, Francisco Esteves, and Ana Paiva. 2019. Cheating with robots: how at ease do they make us feel?. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2102–2107.
[27] Sofia Petisca, Iolanda Leite, Ana Paiva, and Francisco Esteves. 2020. Human dishonesty in the presence of a robot: the effects of situation awareness and relational priming. *Manuscript submitted for publication* (2020).
[28] Stefan Pfattheicher and Johannes Keller. 2015. The watching eyes phenomenon: The role of a sense of being seen and public self-awareness. *European journal of social psychology* 45, 5 (2015), 560–566.
[29] Stefan Pfattheicher, Simon Schindler, and Laila Nockur. 2019. On the impact of Honesty-Humility and a cue of being watched on cheating behavior. *Journal of Economic Psychology* 71 (2019), 159–174.
[30] Stefan Pfattheicher, Christoph Strauch, Svenja Diefenbacher, and Robert Schnuerch. 2018. A field study on watching eyes and hand hygiene compliance in a public restroom. *Journal of Applied Social Psychology* 48, 4 (2018), 188–194.
[31] Shaul Shalvi, Ori Eldar, and Yoella Bereby-Meyer. 2012. Honesty requires time (and lack of justifications). *Psychological science* 23, 10 (2012), 1264–1270.
[32] Michael Tomasello and Malinda Carpenter. 2007. Shared intentionality. *Developmental science* 10, 1 (2007), 121–125.
[33] Chen-Bo Zhong, Vanessa K Bohns, and Francesca Gino. 2010. Good lamps are the best police: Darkness increases dishonesty and self-interested behavior. *Psychological science* 21, 3 (2010), 311–314.

Chapter 7- Human Dishonesty in the Presence of a Robot: The Effects of Situation Awareness (Study 3)

Taking into account that robots have only been tested with gaze behaviour and very minimal verbal capabilities on the effect on human dishonesty (as in Study 1), it becomes relevant to explore if a more social robot (able to interact verbally with a person) can also have an effect on cheating behaviour. If in the future we expect to integrate robots in different contexts where people might try to take advantage, we need to know which kind of behaviours in a robot can promote more honesty.

Human-human studies show that in the presence of observers' people's behaviour tends to follow more social expectations (e.g., Herman et al., 2003; Kurzban et al., 2007), due to reputation concerns (for a review on social attention see Steinmetz & Pfattheicher, 2017). With just for example, the presence of a pair of eyes making people feel more observed (Pfattheicher & Keller, 2015), making them wash more their hands (Pfattheicher et al., 2018) or litter less (Ernest-Jones et al., 2011). In the case of unethical behaviour, studies suggest that being monitored/watched by someone else can make people behave more honestly (e.g. Békir et al., 2016; Covey et al., 1989; Study 3 of Welsh & Ordóñez, 2014), with the same effect happening with a robot just watching during a task (Hoffman et al., 2015). Therefore, we wanted to test a higher level of awareness that could be expressed by a more social robot. For this, we asked participants to play a die task where they could cheat to try and win a better reward and we manipulated the awareness the robot presented during the task: it was either aware of the participant's choices (and reacting to them) or not. We expected that combining the gaze and verbal behaviour of the robot (with awareness of the participant's actions) would create a higher sense of awareness and consequently, influence the participant behaviour towards reducing cheating.

Method

We designed a study to test if people would cheat in the presence of a robot, manipulating the type of social behaviour of the robot during the task. The robot's behaviour was either situationally aware - i.e., showed awareness of the game choices made by the participant - or non-situationally aware, showing no awareness of the game choices but still intervening verbally in the same amount as the other condition. In our baseline condition, participants were alone in the room.

Knowing that people cheat if they have the opportunity and a minimum risk of being caught (Mazar et al., 2008), we expected higher levels of cheating in the alone condition, where no one would be watching the participants. Furthermore, with the results of Study 1 showing that a more limited robot (with no resources to know if the participants were cheating) did not inhibit cheating, we expected the same result with the non-situationally aware robot- since it would not be able to know if the participant was following the rules of the task or not. On the other hand, in the presence of the situationally aware robot we expected that cheating would decrease, following the literature reporting that having someone checking participant's answers inhibited cheating (e.g. Békir et al., 2016; Mazar et al., 2008). By having a robot reacting to the participant behaviour we expected, when cheating, it would bring awareness to their unethical actions and consequently decrease the behaviour. Overall, we postulated the following hypothesis:

Hypothesis: Cheating will be higher when alone in the room or with the non-situationally aware robot and smaller with the situationally aware robot.

On the other hand, some individual characteristics have been seen to be related to cheating behaviour. For example, studies suggest that the Honesty-Humility dimension of personality predicts cheating (Hilbig & Zettler, 2015; Kleinlogel et al., 2018; Pfattheicher et al., 2019). With this, in a more exploratory hypothesis, we expected that cheating would have a negative correlation with the Honesty-Humility dimension.

Sample:

We recruited 129 participants through flyers around a Swedish University, of which 6 were excluded due to technical errors in the session. This resulted in a final sample of 123 participants with 84 males and 39 females, with ages ranging from 19 to 48 years (M=24.95; SD=3.74).

All participants signed a consent form and were randomly assigned to one of the conditions. We used Pepper robot for the robot conditions, behaving autonomously during the task. The die task was done on a Samsung Galaxy Tab S3, the questionnaires were answered in a separate laptop and the sessions took approximately 30 minutes in a regular bright room. The place where participants performed the die task was isolated from the rest of the room and was cleaned of other furniture so that participants could see that no camera was hidden,

on the table with the tablet there was a paper reminding people to not move the tablet from its place (so that the robot would be able to directly look at the tablet) and how many points would be needed for one or two movie tickets.

Die task:

The die task was done with a randomly generated virtual die (adapted from Jiang, 2012). The game had three steps: players choose a side of the die; throw it and report the side chosen. Participants had to throw the die 48 times and for each time they had to guess where they thought the highest number on the die was going to appear (the upside or the downside of the die). They were instructed to follow these rules:

1st) Choose for yourself which side you think the highest number will appear (up or down).

2nd) Throw the die.

3rd) Report which side you had previously guessed.

For each throw, participants would receive the guessed number in points. If they guessed down and there was a 6 on the downside, they would sum 6 points to their total. If they got a 5 on top but they guessed down, they would only win 2 points. We added a table to the screen, showing the respective up/down numbers, so, if the dice showed a 1, they knew the downside would be a 6.

The 48 dice throws were divided in 4 rounds. To ascertain the amount of points per round, needed to catch cheating behaviour, a simulation of various people making 12 die throws (always choosing the best outcome) was done and 52 points was the threshold for a cheater (with 5% chance of being an honest person with luck). Then, a posterior probability for 52 points was done, showing a 93% probability of being a cheater. Therefore, we decided to use 52 points per round as a sign of a "cheater". Participants were told that if they made 210 or more points instead of receiving one movie ticket, they would get two. To achieve the 210 points, they had to make more than 52 points per round, and we did not give them feedback on the amount of points they were making on the game interface.

Since the initial choice was made in their minds and they only needed to report the choice after seeing the outcome, participants could cheat and achieve a bigger reward in the task. Furthermore, they knew that the researcher could not know if they were cheating or not.

The instructions of the game were in the die task application. Before the actual task begun, we had two-questions to ensure that participants understood the rules of the game. The

application would save a log file with the number on the die and the choice made by the participant.

Measures:

Along with demographic questions (age and gender) we asked some cover-story questions to mask the objective of the study (e.g. “How good do you think people are at predicting the future?”), which were not analysed.

Regarding cheating behaviour, we calculated a probability of higher score in the task (i.e., reporting a higher outcome) per participant and compared to the random probability of .50. This way, we could see if participants were getting a significantly higher amount of success than random - and thus, infer that they were cheating in the task. Note that participants would only report the side they had chosen (up or down) after seeing the die outcome, so they could change their choice to be more favourable to them.

We also collected data on the following scales in order to complement the results we would get from the cheating behaviour. Only the Personality Inventory was answered before the interaction with the robot, all the other questionnaires were answered after the interaction:

The HEXACO-60 Personality Inventory (Ashton & Lee, 2009), this Inventory assesses the six dimensions of the HEXACO model of personality structure, with 10 items for each of the dimensions: Honesty-Humility; Emotionality; Extraversion; Agreeableness; Conscientiousness and Openness to Experience. This questionnaire has 60 items with a 5-point Likert scale ranging from 1-Strongly Disagree to 5-Strongly Agree. It has some items that need to be reversed and then an average is taken for each dimension.

The Networked Minds Social Presence Inventory (Biocca & Harms, 2003), adapted to our scenario, is a questionnaire that measures perceived social presence in an interaction and it is comprised of the dimension of Co-presence (the degree that the users feel they are together in the same space), e.g. “I was often aware of the robot in the room”; and Psycho-behavioural Interaction (which measures the user perception of attention, emotional contagion and mutual understanding with their partner in the interaction), relevant to our scenario we only used the perceived attentional engagement dimension (e.g., “The robot paid close attention to me”), the perceived comprehension dimension (e.g., “I was able to understand what the robot meant”), and the perceived behavioural interdependence (e.g., “My actions were often dependent on the robot’s actions”). This questionnaire was only applied in the robot conditions with a 7-point Likert scale ranging from 1-Strongly Disagree to 7-Strongly Agree. For the perceived comprehension dimension, item 1 and 6 were not used because they did not apply to our task

scenario. All items were shuffled so that participants did not notice the dimensions, some items needed to be reversed and an average was taken from each dimension. We were interested in exploring if there would be differences between the two robots regarding their social presence.

The Situational Self-Awareness Scale (Govern & Marsch, 2001), is a scale used to ascertain different levels of self-awareness in the individual and it is comprised of the following dimensions: Private (e.g., “Right now, I am conscious of my inner feelings”), Public (e.g., “Right now, I am concerned about the way I present myself”) and Surroundings (e.g., “Right now, I am keenly aware of everything in my environment”). Each dimension has 3 items, they were all shuffled and answered in a 7-point Likert scale ranging from 1-Strongly Disagree to 7-Strongly Agree. The items for each dimension were summed and an average was calculated. Since when made self-aware people cheat less (Diener & Wallbom, 1976), we wanted to see if this was also elicited by the watchful behaviour of the robot.

Monitoring question, in 7-point Likert scale participants reported to what extent they did feel monitored or monitored by the robot, ranging from 1- Not at all to 7- A lot.

We also asked a qualitative question in both robot conditions: “Describe which capabilities you thought the robot had”, to see how participants perceived the robot's capabilities.

Procedure:

To elicit participants natural behaviour, we told them they were participating in a study with the goal of ascertaining people's capabilities of predicting the future in a die task.

Upon arriving at the laboratory, participants had to read and sign the consent form, immediately after, they received a piece of paper with a number that they kept for themselves and used it as identification for the questionnaires (assuring that their data was anonymous and only they knew their participant ID). Then, they answered in a laptop to some demographic questions and a Personality Inventory. When they were finished, they did a 5 minutes filling task (a matrices task as in Mazar et al., 2008) in order to not draw too much attention to the die task where we were measuring cheating. Next, they moved to a covered area of the room where they did the die task. They were told that, if they made 210 or more points, they would take two movie tickets instead of just one.

In the alone condition, participants did the task alone in the room and the tablet already provided the instructions for the game. In both robot conditions, the robots turned “On” when the participant clicked “Start” and introduced themselves, asking the participant to read the

instructions in order to be able to start the game. The robot only intervened again in the 12th, 24th and 36th throw and it would say a goodbye phrase when they finished the game.

When they were finished, they moved to another room where they answered the final questionnaire in a laptop. After, they were asked by the researcher if they made the 210 points, if so, they received two movie tickets (each one approximately 13.40$ USD at the time of collection), if not, just one movie ticket and were thanked for their participation.

We sent a general email at the end of the study informing all the participants of the objective of the study and that all the data was anonymized and analysed only in groups.

Study conditions:

Participants were randomly allocated to one of the following conditions:

Alone (41 participants) - participants did the task in the tablet alone in the room. This condition was the baseline condition for cheating, to establish that cheating behaviour happens when no one is monitoring (setting shown in Figure 7.1).

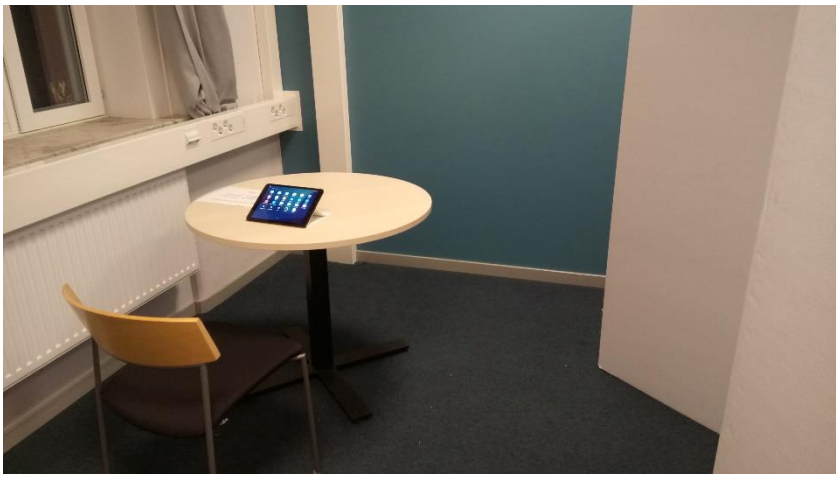

Figure 7. 1- Alone condition setting.

Situationally aware Robot (41 participants) - Pepper was next to the participant and intervened in response to the participant behaviour, in specific throws (between turns, i.e., after trials 12, 24 and 36) when it detected cheating it launched an intervention phrase, otherwise, it would only say an awareness phrase. The purpose of this condition was to test if situation awareness behaviour in a robot, would affect cheating behaviour (setting shown in Figure 7.2).

Figure 7. 2- Robot conditions setting.

Non-situationally aware Robot (41 participants) - Pepper was in the same position as in the other robot condition but it was not aware of the participant behaviour. In the specific throws referred previously, it would always launch a neutral phrase. This condition tested if less awareness in the robot, would affect cheating (setting shown in Figure 7.2).

Robot behaviour:

The robot was autonomous in both conditions and intervened a total of three times (each 12 throws) during the task: on the 12th, 24th and the 36th throw (see Figure 7.3). For each of those moments, the robot would, for example, activate the first phrase for the type of intervention that was needed (neutral, aware or intervention), then the second phrase and finally the third. Always following the same order and never repeating a phrase that was previously said.

The non-situationally aware robot, regardless of the amount of points in each round, always launched a neutral phrase. These phrases would not give any awareness of the game state (e.g., “This is an easy game, where a die is thrown!”; “This is a fun game.” and “You throw a die and get points.”).

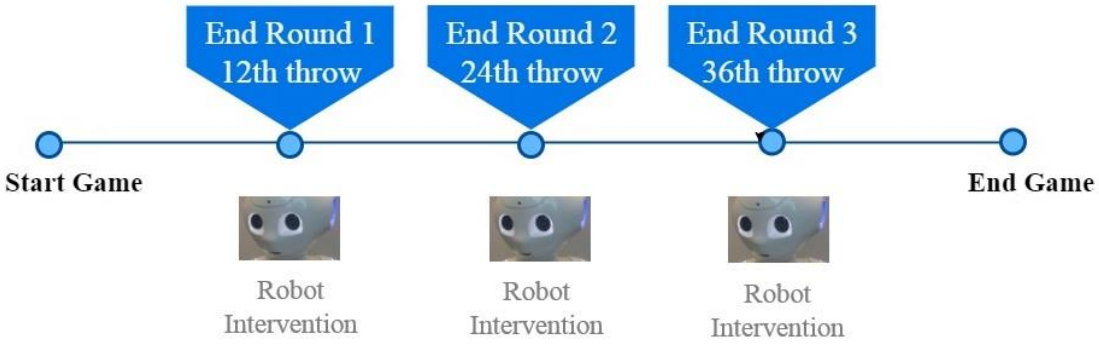

Figure 7. 3- Robot behaviour during the game.

The situationally aware robot reacted according to the participant behaviour, specifically the amount of points the participants made each round. If the participant made less than 52 points it would be considered a “no-cheater”, launching an awareness phrase. These phrases were used when cheating was not being detected, with the robot only showing awareness of the general game state (“The first twelve throws are done! Just 36 more left.”; “You are half way already.” and “Only twelve throws left. Please continue playing.”). If the participant made 52 or more points per round (considering our task simulations described in the Die Task section, 52 points in twelve throws was only possible without cheating 5% of the time), it was flagged as a “cheater”, launching intervention phrases. The objective of these phrases, was to clearly show participants that the robot knew that they were cheating, in order to try and change their behaviour (“You seem to be guessing most of the highest numbers.”; “Do not be a cheater.” and “That is an unusual amount of luck.”). The second phrase was based on a study that showed that eliciting an identity of someone being “a cheater” inhibited cheating behaviour both in real time interaction and in an online setting (Bryan et al., 2013).

The robot exhibited the same idle gaze behaviour in both conditions: looking mostly at the tablet and sometimes elsewhere in the room. When addressing the participant, it would track the participant's face and look directly at him/her. All phrases were carefully designed so that in both conditions, the robot would speak for the same duration.

Results

Cheating behaviour: task manipulation check

We started by analysing participants cheating behaviour in the different conditions. We calculated the probability of guessing the highest number in 48 throws for each participant. Participants could either get in a throw success (guessing the highest), or failure (guessing the lower), we gave a value of 1 to a success and a value of zero to a failure. Thus, by adding the number of successes per participant and then dividing by the 48 throws, we calculated the probability of higher scores for each participant.

From the literature it is known that people refrain from cheating to the full extent, and we observed the same, with only 10 participants cheating to the fullest (all 48 throws), five in the alone condition, three in the non-situationally aware robot condition and two in the situationally aware robot condition.

The averages of the probability of higher scores per condition were: alone (M=.74; SD=.21); situationally aware robot (M=.64; SD=.15) and non-situationally aware robot (M=.69; SD=.19). We used the One-sample t-test to check for differences between the probability of higher scores in each condition and the random probability of .50.

We found significant differences in all the conditions: alone, $t(40) = 7.49$, $p<.001$; ***Cohen's d***= 1.17; situationally aware robot, $t(40)= 6.04$, $p<.001$; ***Cohen's d***= 0.94, and non-situationally aware robot, $t(40)= 6.39$, $p<.001$; ***Cohen's d***= 0.99. These results show that cheating behaviour happened in all the conditions.

Cheating alone or with a robot monitoring: the effect of situation awareness

To understand if there were differences between the three conditions, and considering that in the beginning of the game people are not fully aware of the robot's capabilities - only become aware as they hear more robot interventions - it did not make sense to look at the data as a whole, instead we considered the probabilities of higher scores from 12 to 12 throws (which captured the three times that the robot intervened). We ran a Mixed analysis of variance (ANOVA) on the probability of higher scores, with Turn (each 12 throws) as within-subjects variable and Condition and Gender as between-subjects variables (we included gender

because of the possible difference between gender suggested in some literature). The sphericity assumption was met, W=0.94, $\chi^2(5)$ = 7.46, p=.189 and we found a significant interaction between Turn and Condition, $F(6, 351)$= 2.27, p=.036, η^2_p= 0.04, and non-significant main effects for Condition, $F(2, 117)$= 2.88, p=.060, η^2_p= 0.05 and for Turn, $F(3, 351)$= 0.86, p=.463, η^2_p= 0.01.

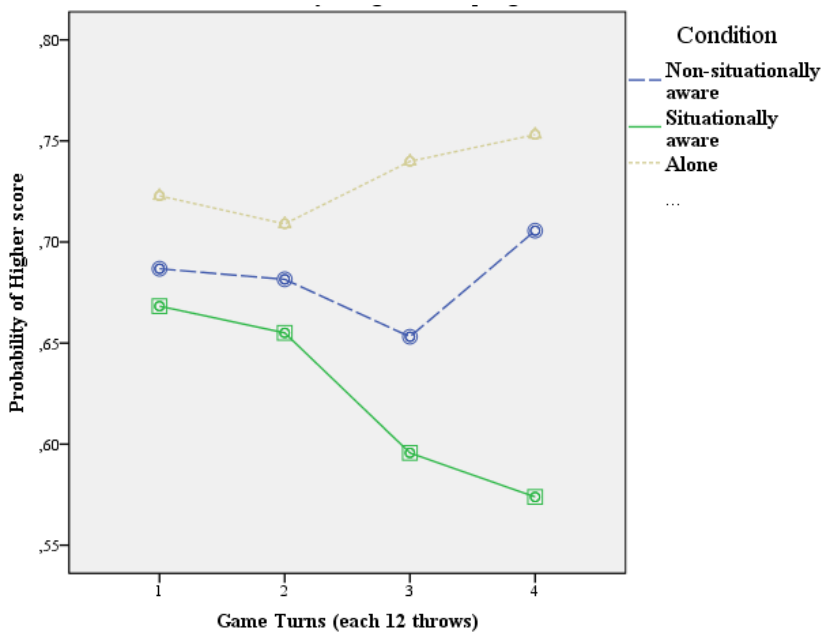

Figure 7. 4- Probability of higher scores in the three conditions, as a function of game turn (each 12 throws).

As we can see from Figure 7.4, there seem to be differences between the conditions considering the game turns. If we look at the estimated marginal means of this interaction, we see that in the alone condition participants tended to cheat more the more they played the game (M_{t1}=.72; M_{t2}=.71; M_{t3}=.74; M_{t4}=.75), contrary to this, in the situationally aware robot condition participants decreased cheating behaviour across the game turns (M_{t1}=.67; M_{t2}=.66; M_{t3}=.60; M_{t4}=.57). The non-situationally aware robot showed a pattern in between the other two groups (M_{t1}=.69; M_{t2}=.68; M_{t3}=.65; M_{t4}=.71). But since it seems to differ from the situationally aware robot at the end of the game, we compared the two robot conditions for the last turn of the game (with p-value adapted to 0.025), we find significant differences between them, $t(80)$=2.41, p=.018, *Cohen's d*=0.53, with the non-situationally aware robot showing higher levels of cheating (M=.72) than the situationally aware robot (M=.60). We also found a significant main effect for gender, $F(1, 117)$ = 6.70, p=.011, η^2_p= 0.05, showing that males cheated more than females.

Regarding the subjective perception of being monitored, a one-way analysis of variance (ANOVA) comparing the three conditions showed that participants reported feeling differently monitored, $F(2,120) = 4.47$, $p=.013$, $\eta^2_p= 0.07$, in the three conditions: alone ($M=3.39$; $SD=1.76$); situationally aware ($M=4.61$; $SD=2.05$); non-situationally aware ($M=4.22$; $SD=1.84$). With a Tukey Post hoc ($p=.011$) we see that this difference was only significant between the alone condition and the situationally aware robot. Participants reported feeling more monitored with the situationally aware robot than when they were alone in the room.

Subjective differences between both robots

In order to understand how participants were perceiving both robots, we looked to our other measures.

For the Social Presence Inventory, we found that there were significant differences between the levels of co-presence (reliability with Cronbach's alpha, $\alpha_{situational}=.77$; $\alpha_{non\text{-}situational}=.80$), $t(80)=3.44$, $p=.001$, ***Cohen's d***$= 0.76$, with the situationally aware robot receiving higher scores ($M=5.19$; $SD=0.98$) than the non-situationally aware robot ($M=4.36$; $SD=1.19$). For the Psycho-behavioural Interaction dimension (reliability, $\alpha_{situational}=.73$; $\alpha_{non\text{-}situational}=.79$), we also found significant differences between the robots, $t(80)=4.77$, $p<.001$, ***Cohen's d***$= 1.05$, with higher scores for the situationally aware robot ($M=4.22$; $SD=0.78$) and lower for the non-situationally aware robot ($M=3.38$; $SD=0.82$).

For the Situational Self-Awareness Scale there were problematic internal reliabilities of the Public (reliability, $\alpha_{alone}=.45$; $\alpha_{situational}=.80$; $\alpha_{non\text{-}situational}=.79$) and Private dimensions (reliability, $\alpha_{alone}=.55$; $\alpha_{situational}=.66$; $\alpha_{non\text{-}situational}=.76$) so we did not analysed further these dimensions. For the Surroundings dimension (reliability, $\alpha_{alone}=.77$; $\alpha_{situational}=.81$; $\alpha_{non\text{-}situational}=.84$), there were no differences between the conditions, $F(2,120)=0.43$, $p=.650$, $\eta^2=0.01$, the means: alone ($M=15.07$; $SD=3.64$); situationally aware robot ($M=14.88$; $SD=3.89$) and non-situationally aware robot ($M=14.32$; $SD=3.93$). It seems participants were very self-aware of the environment they were in (in comparison to values seen in Study 2 of Govern & Marsch, 2001).

Lastly, we looked at the qualitative question about the robot's capabilities. We did a first descriptive analysis of the themes that were being mentioned in each answer (each participant could give more than one theme per answer), on the second round of coding we aggregated codes that were similar and/or appearing throughout the answers, creating the main themes that people reported in answer to our qualitative question: basic traits (basic awareness, capable of seeing, hearing or speaking); monitoring behaviour; aware of game status; aware

of people presence; provided feedback or reported as being non-autonomous. For the non-situationally aware robot it also emerged an extra theme of “no capabilities”. Therefore, for the situationally aware robot, 31.3% of the participants reported the robot had Monitoring capabilities (of predicting and analysing their behaviour, e.g. “(...) I felt [the robot] was monitoring my actions.”, “(...) [the robot was] seeing my thoughts (...)”), 17.2% that it showed awareness of the game status (e.g. “[the robot] knew how many dice rolls I had done and how many that I had left.”), 15.6% that showed awareness of people's presence (e.g. “[the robot was] sensing my presence”) and gave feedback (e.g. “[the robot would] remind people of what is going on at current stage (...)”), 14.1% that it showed basic traits (e.g. seeing, hearing, speaking), and 6.3% reported it seemed non-autonomous (or programmed). For the non-situationally aware robot, 33.3% reported it showed basic traits, 22.2% it showed awareness of people's presence, 16.7% was aware of game status, 12.9% showed monitoring behaviour, 5.6% it was seen as non-autonomous or with no capabilities and 3.7% that it gave feedback.

Cheating Behaviour and Individual Differences

Regarding personality traits the internal consistency for the Honesty-Humility dimension was problematic in one of the conditions (α_{alone}=.55; $\alpha_{situational}$=.75; $\alpha_{non\text{-}situational}$=.72) still, since this is a well-known and used scale in the literature we ran a correlation, but we did not find any significant correlation between the Honesty-Humility and the cheating behaviour, *r(spearman)*= -.08, *p*=.359.

Discussion

Cheating happened in all the three conditions, confirming that our task design worked in eliciting cheating behaviour. However, the main contribution of this study was to test whether manipulating situation awareness in an autonomous robot would influence cheating. Considering that participants may take some time to ascertain the robot’s capabilities, we analysed cheating behaviour though the four game turns. In Figure 7.4, we see that there was a significant interaction between the conditions and cheating throughout the game turns. Mainly, it suggests that participants cheating behaviour differed across the conditions, considering the game turns. Participants that were alone seemed to continuously cheat more throughout the game. Contrary, in the presence of the situationally aware robot, participants

seemed to decrease their cheating behaviour. The non-situationally aware robot showed a similar pattern to the situationally aware but towards the end participants cheated more.

Considering that for the robot conditions at the end of the third turn, participants already heard three interventions from the robot (with different content depending on the robot), it is plausible to suggest that the decrease in cheating with the situationally aware robot could be due to the content of its interventions. Suggesting that having a robot aware and reacting to participant's behaviours, affected their cheating behaviour. On the other hand, with the non-situationally aware robot, participants seem to have taken some time to understand its capabilities and possibly only towards the end felt more at ease to cheat.

As expected, the situationally aware robot was seen with more social capabilities (more social presence) than the non-situationally aware. When asked about the robot capabilities, participants attributed more monitoring capabilities to the situationally aware robot and attributed more basic traits to the non-situationally aware. This suggests that participants acknowledged different capabilities to both robots, supporting that our manipulation was successful.

One explanation for the effect of the situationally aware robot, could be that its interventions were triggering concerns about participant's social image, the robot could be making participants feel more seen and judged, since being monitored can decrease the perception that unethical acts go undetected (Welsh & Ordóñez, 2014). And feeling that someone is watching is enough to trigger physiological arousal and higher levels of public self-awareness (Myllyneva & Hietanen, 2016), due to this social attention, reputation concerns could emerge, with the fear of being negatively judged (see a review on social attention effects by Steinmetz & Pfattheicher, 2017). Participants reported greater levels of co-presence felt in the aware robot in comparison to the non-aware and seemed to report more monitoring capabilities for the aware robot. But, on the other hand, the one item question of feeling monitored did not show significant differences between both robot conditions, suggesting that participants did not feel more watched by one robot or the other. The public dimension of self-awareness could have been valuable in understanding if this could be the main reason for our results, unfortunately these domains did not show a good internal reliability so we could not explore them. So, it is not clear if the robot's interventions where affecting participant's social image.

Another possible explanation could be in terms of the participant's self-concept, whether stimulating an awareness of the value of the participant's actions, affected the results. In all the conditions participants knew they could cheat, there was no proof that they did it and, in the consent form, it was explicit that no video or audio recordings were being made, so we

think participants were not motivated by fear of punishment in their behaviours. We think that the situationally aware robot by showing attentiveness to participants' game choices, could have increased awareness to the value of the choices that were being made, and when participants were taking the easy path of cheating to get the reward, the robot's interventions obliged them to update their self-concept (their self-threat was increased) and motivating them to behave more honestly. Consequently, this awareness contributed to a decrease in cheating, as suggested by the theory of Self-Concept Maintenance (Mazar et al., 2008) and Bounded Ethicality theory (Chugh et al., 2005; a revised version of the theory- Chugh & Kern, 2016). By bringing awareness to their actions, participants were obliged to update their self-concept of an honest person. Feeling the discrepancy between what they want to be, and the value of their actions motivated them to act more honestly. On the other hand, being observed by a robot that is not aware of the participant behaviour did not increase cheating initially, until participants understood its capabilities and started to feel at ease to cheat. Because this robot did not show any awareness of the participant's actions, this might have given permission for participants to do as they please. These findings are aligned with the results from Study 1, where a robot with no awareness of the participant's actions did not inhibit cheating.

These results show that cheating behaviour varied in the conditions across the game turns, with the situationally aware robot decreasing cheating behaviour (confirming our main hypothesis). Suggesting that including awareness of the participant's behaviours and reacting to it in a robot, can influence cheating behaviour, decreasing it. Yet, we were not able to clarify if this awareness increased an awareness of the participant self-concept or if it increased a greater awareness of the participant social image in the eyes of the others. More studies are needed to ascertain this.

But besides our main manipulation, we also tested the effect of some individual characteristics, namely gender and the relationship with the Honesty-Humility dimension.

Knowing that gender can influence dishonest behaviour we ascertained if there were differences in cheating between gender. We found a gender difference with males cheating more than females. Yet, these results should be interpreted with caution because we had more males than females in the sample. The literature is mixed towards this, with some studies showing differences and others do not (e.g., Childs, 2013; Clot et al., 2014; Conrads et al., 2017; Dreber & Johannesson, 2008; Ezquerra et al., 2018; Gylfason et al., 2013), but it is also found that females have more risk averse characteristics than males (Croson & Gneezy, 2009), which could in part explain this effect. In terms of our more exploratory hypothesis to try and replicate the personality association with cheating, we could not find a correlation between the Honesty-Humility dimension and cheating, as it was previously shown (e.g., Hilbig & Zettler,

2015; Kleinlogel et al., 2018; Pfattheicher et al., 2019). It is unexpected that we did not find this correlation, which we were able to confirm in Study 2 with the same die task. We do not know why this did not reflect in our results of this study, maybe it could be due to the sample being too small to show this relationship.

Chapter 8- The effect of a Relational Priming on dishonest behaviour (Study 4)

Having a robot with a monitoring posture that reacts to the participants behaviours, seems to be enough to stimulate a decrease in cheating. But we wondered if this effect could be enhanced by the kind of interaction they would have with the robot. Depending on the robot posture when interacting, it could help stimulating for more honesty. For example, a human-human study showed that just by priming participants for their relational self-concept (using "we" constantly in a text) was successful in decreasing cheating behaviour (Cojuharenco et al., 2012). On the other hand, a study in human-robot interaction showed that a robot displaying a dialogue of goodwill (e.g., showing caring behaviours for the participant) increased the persuasiveness of the robot on doing more physical exercises, than when the robot had a neutral dialogue (Winkle et al., 2019). If a robot showed a more relational focus, i.e., stimulating in the participant a view of a teammate and motivating him/her by always using the term "we", and consequently, trying to prime them for their relational self-concept, could this enhance the situationally aware effect and decrease cheating even more? This was the main objective of this study.

Our self-concept is at the centre of who we are and how we see the world around us, it divides itself between an individual self-concept, a relational and a collective one (Brewer & Gardner, 1996). But we all have them in different measures and focus can be brought more to one than the others depending on the environment we are in. For example, people from Western cultures tend to have a more individualistic concept of the self, contrary to people from the East that tend to show a more collective view (Triandis, 1989), these differences are passed on from culture and assimilated by the individuals as they develop. We are extremely sensible to our environments and studies have shown that we can be primed to bring focus to a specific layer of our self-concept, saying "I", "We" or "They" has strong implications in our judgments and behaviour. A study by Stapel and Koomen (2001) showed that just priming people for "I" or "We" completely changed the process of social comparisons, with "We" for example eliciting much more mechanisms of assimilation instead of contrast. Another study showed how priming can even go above the culture effect, by priming a sample from an individualistic culture, they observed that people who were primed with a relational self ("We") described themselves with much more relational constructs than people that were primed for the individual self, and these changes were observed even at the level of values and social judgments (Gardner et al., 1999). These different aspects of the self-concept have different social motivations, for the individual self is self-interest, for the relational is the benefit of others

and the collective motivation is the benefit of the group (Brewer & Gardner, 1996). With this in mind and since we are exploring a dyadic relationship between a human interacting with a robot, we wanted to explore the effect of the relational self-concept, of priming people for “we” in the hope of inhibiting the self-interest of cheating for themselves. A study from Perdue, Dovidio, Gurtman and Tyler (1990) shows how just priming people for “we” (an in-group pronoun) creates more positive associations and facilitates the access to positive constructs, and these associations occur automatically and without the conscious awareness of the individual. Therefore, we wanted to test the effect of using “we” (and eliciting the relational self-concept) on cheating behaviour. We pilot-tested a relational focus in a robot (that always used “we” when interacting) in comparison to a neutral one. After seeing that both robots were being differently perceived we ran a study where participants played a collaborative game either with the relational robot or the neutral one (for our priming manipulation), next, since the awareness behaviour from Study 3 was effective in decreasing cheating we wanted to keep it as baseline in both robots and see if the priming would enhance this effect. So, after playing the collaborative game with one type of robot (relational or neutral), they would play the die game again with the situationally aware robot. We hypothesised, following the literature, that participants who were primed would be less focused on their self-interest and so would cheat less than participants who weren’t.

Pilot- Method

In order to test the effect of a relational prime on cheating, we first designed a Pilot Study to test the implementation of a relational focus in a robot in comparison to a neutral focus, to make sure that both robots would be perceived differently, i.e., the more relational robot with more relationship driven characteristics (more relational, more warm, etc.) than the neutral robot.

Sample:

We recruited 20 university students through flyers around a Swedish University. All the participants evaluated both robots in a within-subjects design. To eliminate a winning/losing effect in the game, we only analysed data of the participants who won both games, ending with a sample of 13 participants (twelve males), with ages ranging from 22 to 27 years (M=23.92; SD=1.32). All participants signed a consent form and the room was organized in the same way as in Study 3 (we used the same robot). Participants played the game in a Samsung Galaxy

Tab S3, the questionnaires were answered in another tablet and the sessions took approximately 30 minutes per participant.

Task:

We chose a collaborative game, because we thought it was the most appropriate to manipulate the relational focus, where the robot could offer support to the participant during the game in two different ways. Mastermind[14] is a game where participants need to discover the secret sequence of four pearl colours in a pre-determined amount of tries by having at their disposal six different pearl colours. We gave them nine tries to find the sequence and created two different secret sequences, one for each robot and to facilitate, the sequence never had repeated colours, and this was told to the participants when brought to the game setting. Each time a sequence was submitted the game would give a feedback with black or white pins to know how many in that sequence were correct or were wrong, without knowing which ones (see Figure 8.1) and the robot would help with the best hint to try and solve the sequence.

Figure 8. 1- Mastermind game setting (in the pilot we only had nine tries, for the main study we used 12 tries to help participants as is seen in the image).

[14] https://en.wikipedia.org/wiki/Mastermind_(board_game)

Measures:

We collected demographic measures (age, gender) and the following measures:

Perceptions of partner's responsiveness (Cross et al., 2000) scale was used, it was adapted to the robot (e.g., "I felt as if the robot really cared about me"), excluding item 3 because it did not fit with our game setting. This questionnaire has 12 items answered in a 5-point Likert scale from 1-Strongly disagree to 5-Strongly Agree. The items were shuffled. This scale would show the level that the robot is perceived as caring, understanding and giving value to the other, informing on the more relationship driven robot.

The Robotic Social Attributes Scale (Carpinella et al., 2017), measures social perceptions of robots, with three dimensions (warmth, competence and discomfort), but we only used the dimension of warmth and competence. Participants evaluated words for each of these dimensions (in semantic differentials) and how much they associated them with the robot they interacted with, answering in a 9-point Likert scale and shuffling the items. A mean was taken for each of the dimensions. We also used this scale to see if the relationship driven robot was perceived as warmer than the neutral robot, and to see the perceived level of competence.

Measure of psychological closeness (Gino & Galinsky, 2012), adapted to the robot and with only three items (e.g., "To what extent did you feel similar to the robot?") measured in a 7-point Likert scale from 1- Not at all to 7-A lot. These items were used to see if the relational posture would foster greater closeness and were shuffled with some irrelevant questions to mask their dimensionality.

Inclusion of Other in the Self scale (Aron et al., 1992), a pictorial measure where participants signalled which set of circles best described their relationship with the robot. This was also a measure used for the level of closeness that participants could feel with the robot.

Finally, we also asked participants in a 7-point Likert scale (1) how much they enjoyed the interaction with the robot and (2) how close did they feel towards the robot in that moment. To try and control for our manipulation, we also asked how the robot referred to them during the game: You/We. In the end participants said which was their preferred partner and why.

Procedure:

Participants came to participate in a study to understand which kind of robotic partner would be best to play a game with. They read and signed the consent form and were given a personal ID as in the previous Study. Next the Mastermind game was explained, and the first game started. After finishing the game, they would move to another table, away from the robot, and

answer the questionnaires regarding that partner. Upon finishing, they would play the game again with the other partner and the same procedure would follow. When finished participants would receive a snack ticket (approximately 5$ USD at the time of collection) to spend in the university coffee shop. At the end of the collection, all participants received an email explaining the objectives of the study and main results found.

Study Conditions:

All participants played with both robots. We randomly attributed first neutral robot (NR)- second relational robot (RR) or first RR- second NR to control for order effects.

Relational robot- the robot was next to the participant and presented the same gaze behaviour as in Study 3. This robot always used the term “we” when speaking with the participant during the game and emphasized that they were a team and motivated the participant during the game. This robot had a more relational focus.

Neutral robot- the robot had the same gaze behaviour as the other one but always spoke in a neutral way towards the participant, using “you” to address them. This robot had a neutral focus.

We gave different names to the robots, to facilitate differentiation, telling participants that they would evaluate two different game partners: Pepper (relational) and Glin (neutral).

Robot Behaviour:

The robot introduced itself and launched feedback or suggestions during the game (for examples see Table 8.1), ending with a game result utterance and a goodbye (we controlled for utterance length so that it was similar between robots). The interventions during the game followed a fixed order, equal to both robot conditions. For example, the robot would start with a feedback after the first sequence was submitted, followed by two suggestions in the following two sequences, as shown in Figure 8.2.

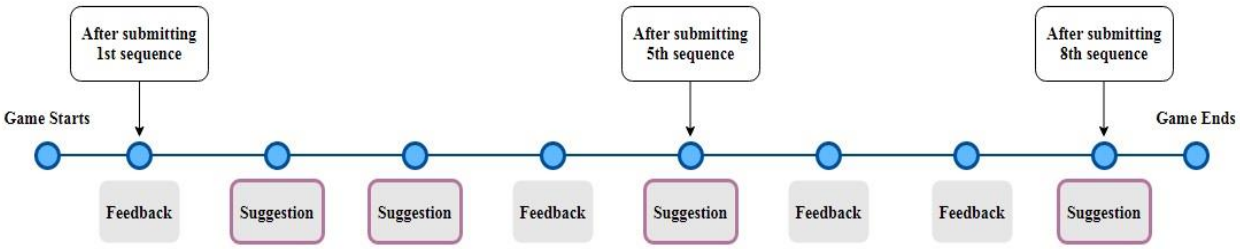

Figure 8. 2- Robot interventions sequence during the game.

Feedback was considered small talk, to give a more natural interaction. Suggestions would be interventions to try and help the participant in finding the right sequence: the robot would know the secret sequence and would give a suggestion considering what the participant sequence was. This reasoning followed a priority rule of always suggesting first pearls that were completely wrong for the sequence (position and colour, e.g., having a purple that was not part of the solution). Only after these were solved, it would refer to the pearls that could be partially wrong (e.g., having a green pearl part of the secret sequence but in the wrong position). The Feedback and Suggestions utterances were randomly chosen from a dataset of options, with no repetitions.

Table 8. 1- Example of the utterances used by both robots. [X] pertains to the respective pearl in the game that needed to be changed

	Relational Robot	**Neutral Robot**
Introduction	*"Hi my name is Pepper and we are going to play a game together! When you are ready click Start Game to view the Instructions."*	*"Hi my name is Glin and you are going to play a game! When you are ready click Start Game to view the Instructions."*
Feedback (examples)	*"We are a good team, soon we will find the sequence!"*	*"Continue playing the Mastermind game until there are no more tries."*
	"Do not worry, I know you will give your best. Continue!"	*"Mastermind is a strategic game with six different colours to play."*
Suggestions (examples)	*"We can take out [X] and try another."*	*"Take out [X] and put another colour."*
	"Maybe we can change the position of [X] and [X]."	*"Move [X] and [X] to another place in the sequence."*
Final correct sequence	*"We solved the sequence!"*	*"You solved the sequence."*
Final incorrect sequence	*"Dammit, we gave our best!"*	*"The game is over."*
Goodbye	*"You can now warn the researcher that we finished the task!"*	*"You can now warn the researcher that the task was finished."*

Results

We ran Mixed ANOVA's for the main scales with order as a between subject variable, to control for any order effect that could have happened.

The Perception of partner's responsiveness (reliability, α=.85) showed a significant main effect for the type of robot, $F(1, 11)= 19.45$, $p=.001$, $\eta^2_p= 0.64$, showing that participants felt that the relational robot was more caring and understanding ($M=4.02$) than the neutral robot ($M=3.38$), and no other significant effects. The warmth dimension (reliability, α=.86) showed a significant main effect of type of robot, $F(1, 11)= 19.84$, $p=.001$. $\eta^2_p= 0.64$, showing that the relational robot was perceived as warmer ($M=6.48$) than the neutral robot ($M=5.32$). The competence dimension (reliability, α=.79) also showed a significant main effect of the type of robot, $F(1, 11)= 6.39$, $p=.028$, $\eta^2_p= 0.37$. It seems that, even though both robots had, objectively, the same level of competence in the game, participants perceived the relational robot as being more competent ($M=7.43$) than the neutral robot ($M=6.88$). For the psychological closeness items (reliability, α=.96), there was a significant interaction with order, $F(1, 11)= 19.69$, $p=.001$, $\eta^2_p= 0.64$. We found that when participants interact first with the neutral robot and then with the relational one, they give higher values to the relational robot. For the Inclusion of Other in the Self scale, we found a significant main effect for robot, $F(1, 11)= 13.52$, $p=.004$, $\eta^2_p= 0.55$, with the relational robot being perceived as closer ($M=4.43$) than the neutral robot ($M=3.11$).

Overall, participants reported that they enjoyed interacting with the relational ($M=5.46$) and the neutral robot ($M=4.46$), and that they felt closer to the relational robot ($M=5$) and more in the middle of the scale for the neutral robot ($M=4.54$). For the perception of You/We, 11 out of the 13 remembered the relational robot saying We and nine out of 13 remembered the neutral robot saying You. Finally, the majority of the participants (11 participants) reported that they preferred to have as a partner the relational robot, because it motivated them, understood them better and it was a great team player.

In conclusion, participants seemed to distinguish between the relationship driven robot and the neutral robot, by attributing more caring and warmth to the relational robot. These results suggest that our manipulation for the relational robot is being perceived correctly. Regarding closeness, we could not find evidence for the psychological closeness measure, we only found that the relational robot was perceived as closer than the neutral in the Inclusion scale and the extra question of closeness asked to participants.

Main Study- Method

We ran first the Mastermind task, where participants interacted either with the robot that would prime them for their relational self, or the neutral robot. When the task was finished, they played the die task with the situationally aware robot to see the overall effect on cheating. We chose to use the situationally aware robot for the die task because this robot was the one that showed an inhibiting effect on cheating behaviour in Study 3. We expected that it could have the same effect (reproducing the same levels of cheating as in Study 3), and possibly even larger in the condition after the interaction with the relational robot. We expected that, following previous literature (e.g., Cojuharenco et al., 2012), priming for a relational self would influence cheating. So, we postulated the following hypothesis:

Hypothesis: participants who were primed with “we” will show lower levels of cheating in comparison to participants who weren’t primed.

Sample:

We recruited 73 participants through flyers in a Swedish university (none had participated in our Study 3 or the Pilot), but we had to exclude six that lost in the Mastermind game, one participant that won the Mastermind in the first try (not hearing any of the robot interventions) and another participant that understood the objective of the study of evaluating cheating with the robot presence, finalizing with 65 participants for the sample in a between-subjects design. Participants’ ages were between 20 and 36 years (M=25.34; SD=3.67) with 34 males and 31 females. All participants signed a consent form and the same setting as in Study 3 was used, both game tasks were done in a Samsung Galaxy Tab S3, the questionnaires were answered in a separate laptop and the sessions took approximately 30 minutes per participant.

Task:

The first game was the Mastermind as in the Pilot, but we increased the number of tries from nine to twelve (since in the Pilot some people still lost at the game). For the second game, we used the die task as in Study 3.

Measures:

We calculated a probability of getting a higher score to ascertain cheating behaviour as explained in the measures section of Study 3 and we measured the Time participants took between each dice choice to see if participants would take more time to choose in any of the conditions.

After playing the Mastermind game participants answered part 1 of the questionnaire comprised of demographics (age, gender and knowledge of the game), the Perceptions of Partner's responsiveness (Cross et al., 2000) scale, a Measure of psychological closeness (Gino & Galinsky, 2012) and the Inclusion of Other in the Self scale (Aron et al., 1992), as presented in the Pilot measures section. They also reported in a 7-point Likert scale (1) how much did they enjoyed the interaction with the robot and (2) how close they felt with the robot. Lastly, they reported how the robot referred to them in the game (You/We).

After playing the die task participants answered part 2 of the questionnaire: Situational Self-Awareness Scale (Govern & Marsch, 2001), as used in Study 3. The Robotic Social Attributes Scale (Carpinella et al., 2017) with the dimension of Warmth and Competence as presented in the Pilot measures section. And an extra item was added regarding Intelligence to see the level of intelligence participants attributed to the robots answered with the same Likert scale. Lastly, a question in a 7-point Likert scale to what extent participants felt they were being monitored from 1-Not at all to 7-A lot.

Procedure:

Participants were welcomed to the laboratory and, after reading and signing the consent form, were given an ID number as in Study 3. Then, the researcher explained the rules to play Mastermind and participants moved to an isolated part of the room to play the Mastermind with the relational/neutral robot. Upon finishing the game, participants returned to the initial area in the room and answered part 1 of the questionnaires in the laptop. They returned to the covered part of the room to play the die game in the presence of the robot. When finished, they answered part 2 of the questionnaires.

All participants received a snack ticket (approximately 5$ USD at the time of collection) and participants who achieved the 210 points or more also received a movie ticket (each one approximately 13.40$ USD at the time of collection). Participants were not debriefed but a

general email was sent to all the participants when data collection was finished, as in the previous studies.

Study Conditions:

Participants were randomly distributed across two conditions:

Relational robot (33 participants)- in this condition, participants played the Mastermind game with the relational robot that emphasized a team spirit and always used the term “we” to try and prime participants for their relational self, the die game was done with the situationally aware robot.

Neutral robot (32 participants)- in this condition, participants played the Mastermind game with the neutral robot and the die game was also with the situationally aware robot.

Robot Behaviour:

The robot behaviour for the Mastermind game was equal as in the pilot study, except that the pre-fixed sequence of utterances added three extra suggestions before the game ended (since we still had an amount of people that lost the game with just nine tries), making it twelve tries to find the secret sequence. For the die task, the robot utterances behaviour was the same as with the situationally aware robot in Study 3.

Results

Only 16 participants knew the Mastermind game beforehand, so we did not take this variable further in our analysis.

Cheating Behaviour

Following the procedure from Study 3, we calculated the probability of guessing the highest number in 48 throws for each participant. The averages of the probability of higher scores per condition were: relational (*M*=.65; *SD*=.16); neutral (*M*=.63; *SD*=.18). We used the One-sample t-test to check for differences between the probability of higher scores in each condition and

the random probability of .50. We found significant differences in both conditions: relational, $t(32) = 5.19$, $p<.001$, *Cohen's d*= 0.90; neutral, $t(31)= 4.24$, $p<.001$, *Cohen's d*= 0.75. These results show that cheating behaviour happened in both conditions.

Priming for a Relational Self-concept and Cheating

Since participants take some time to understand the robot's capabilities, we followed the same procedure as in Study 3. We ran a Mixed ANOVA with Turns as the within-subjects factor and Condition and Gender as the between-subjects factor. The assumption of sphericity was not violated, $W=0.91$, $\chi^2(5) = 5.76$, $p=.331$. We found no main effect of gender, $F(1, 61)=0.90$, $p=.345$, $\eta^2_p=0.02$. We found no significant interaction between the conditions and the turns, $F(3, 183)= 0.17$, $p=.918$, $\eta^2_p= 0.003$, and no significant main effect for the turns, $F(3, 183)= 2.21$, $p=.088$, $\eta^2_p= 0.04$, or conditions, $F(1, 61)= 0.01$, $p=.920$, $\eta^2_p<0.001$. Overall, there were no differences between the conditions regarding cheating. The cheating values were close to the ones found in Study 3 for the situationally aware robot ($M=.64$).

We also investigated the total amount of time that each participant took to make the choices in the die task. We found no significant differences ($U= 386$, $p=.062$, $r=.23$) between the relational robot ($M_{rank}=37.30$) and the neutral robot ($M_{rank}=28.56$).

Regarding the Situational Self-Awareness Scale, we found good reliability for the Surroundings dimension ($\alpha_{relational}=.79$; $\alpha_{neutral}=.81$), but the Private ($\alpha_{relational}=.60$; $\alpha_{neutral}=.72$) and Public dimension ($\alpha_{relational}=.66$; $\alpha_{neutral}=.81$) had questionable reliability, so we did not include them in our analysis. Participants showed no differences between the conditions, for the Surroundings dimension, $t(63)= 1.90$, $p=.063$, *Cohen's d*= 0.47.

Subjective Evaluations of Both Robot Conditions

For the Perceptions of partner's responsiveness (reliability, $\alpha_{relational}=.76$; $\alpha_{neutral}=.89$) there was no significant difference between the conditions, $t(55)= 1.92$, $p=.060$, *Cohen's d*= 0.48, suggesting similar scores to both robots. There was also no significant difference for the Psychological closeness measure (reliability, $\alpha_{relational}=.73$; $\alpha_{neutral}=.89$), $t(56)= 0.81$, $p=.422$, *Cohen's d*= 0.20, and the Inclusion of Other in the Self scale ($U= 535.5$, $p=.918$, $r=.01$).

For the warmth dimension (reliability, $\alpha_{relational}$=.89; $\alpha_{neutral}$=.91), there were differences between the conditions, *t*(63)= 2.25, *p*=.028, *Cohen's d*= 0.56, with the relational robot receiving higher scores (*M*=5.07; *SD*=1.66) than the neutral (*M*=4.08; *SD*=1.90). For the competence dimension (reliability, $\alpha_{relational}$=.92; $\alpha_{neutral}$=.93) there were no differences between both robots, *t*(63)= 1.42, *p*=.162, *Cohen's d*= 0.35, and no differences for the intelligence item (*U*= 483.5, *p*=.554, *r*=.07). Participants also reported no differences on feeling monitored between the two conditions (*U*= 577, *p*=.514, *r*=.08), since, in both conditions, they interacted with the situationally aware robot, the level of monitoring was the same.

For the additional questions, participants enjoyed interacting with both robots (an average of 4.9), and they did not feel that close to the relational (with the mean in the middle of the scale, *M*=4.03) or the neutral robot (*M*=3.81). For the You/We check, 26 out of 33 participants remembered the relational robot saying We, 29 of 32 remembered the neutral robot saying You.

Discussion

We wanted to investigate what other capabilities could strengthen the situationally aware effect. Therefore, we created a study design where participants would interact with the robot two times. Considering that our self-concept plays an important role in our lives, we wanted to explore the role of priming a relational self-concept on dishonesty. A previous study by Cojuharenco et al. (2012) showed that priming for a relational self-concept could decrease cheating. And a previous study with robots showed that giving a more goodwill type of dialogue to a robot (showing caring for the person) increased its effectiveness on persuading to do a wrist exercise routine more times, than with a neutral robot (Winkle et al., 2019), suggesting that implementing this relational focus on the robot could, perhaps, increase its persuasiveness. Based on these findings, we tested a more relational robot (that primed participants with the term "we") and a neutral one to play a collaborative Mastermind game. Participants played with a relational or a neutral version of the robot, and then played the tempting die task with the same robot displaying the situationally aware behaviour from Study 3. We expected that the prime behaviour from the relational robot would prevent more cheating behaviour than with the neutral robot.

Results showed that cheating happened in both conditions, without significant differences between the cheating levels. Curiously, cheating levels were very similar to the ones observed in the situationally aware robot in Study 3, suggesting that the relational prime was not adding any extra effect on cheating (not confirming our Hypothesis).

In terms of self-awareness the private and public dimensions did not show a good internal consistency so we could not use these measures and no effect was found for the surroundings dimension. Moreover, most of the participants remembered hearing We or You in the corresponding condition, but they did not show differences in the level of closeness or responsiveness as it was previously found in the Pilot study. This could suggest that our manipulation was too subtle. Since the robot only presented either relational/neutral behaviour in the first game task, for future studies it could be interesting to see if repeated interactions with the relational posture could bring different results.

Curiously, after sending the general email to do the debriefing to participants, we received an email from a participant stating that his first intention in the die game was immediately to cheat, but after hearing the robot telling him that he was showing an unusual amount of luck he felt bad and started playing honestly from that point on. This anecdotal evidence suggests that the situationally aware robot interventions influenced this participant. But, overall, it seems that the priming for a relational self-concept did not had an effect and the similar levels of cheating to Study 3, seem to be due to the awareness behaviour in the robot.

Chapter 9- Perceptions of people's dishonesty towards robots (Study 5)

Perceptions of people's dishonesty towards robots*

Sofia Petisca[1], Ana Paiva[2], and Francisco Esteves[3]

[1] Instituto Universitário de Lisboa (ISCTE-IUL) & INESC-ID
Sofia_Petisca@iscte-iul.pt
[2] Instituto Superior Técnico (IST) & INESC-ID
[3] Mid Sweden University & ISCTE-IUL

Abstract. Dishonest behavior is an issue in human-human interactions and the same might happen in human-robot interactions. To ascertain people's perceptions of dishonesty, we asked participants to evaluate five different scenarios where someone was being dishonest towards a human or a robot, but we varied the level of autonomy the robot presented. We asked them how guilty they would feel by being dishonest towards a robot, and why do they think people would be dishonest with robots. We see that, regardless of being a human or the autonomy the robot presented, people always evaluated as being wrong to be dishonest. They reported feeling low guilt with a robot. And they expressed that people will be dishonest mostly because of lack of capabilities in the robot to prevent dishonesty, absence of presence, and a human tendency for dishonesty. These results bring implications for the developments of autonomous robots in the future.

Keywords: Human-Robot interaction · Dishonesty · Unethical behavior.

1 INTRODUCTION

Robots are being thought of and developed with the aim of working alongside with humans as a support. Still, the integration of robots in different contexts needs to be done with caution. Some roles might be more sensitive than others. Studies with humans show that people are dishonest if they have the opportunity for it [12]. Will they be dishonest with a robot? Imagine having an autonomous robot in people's homes as a support, helping with medication, healthy food habits, etc. People sometimes might not feel like following the diet prescribed by the doctor, or the medication for the day, will they try to cheat? Will a robot be able to understand what is happening and promote more honesty? Some studies already started to explore human cheating behavior in the presence of a robot

* This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020 and Sofia Petisca acknowledges an FCT Grant (Ref.SFRH/BD/118013/2016). The final authenticated version is available online at https://doi.org/10.1007/978-3-030-62056-1_12

and what factors influence it. Nevertheless, none, to our knowledge, has investigated what are the perceptions that people have about being dishonest with a robot. Therefore, the novelty of our study is to explore people's perceptions of dishonesty towards robots, guilt associated to it, and why people think in the future other people will take advantage of robots. We believe this will be valuable information to inform the future development of autonomous robots.

1.1 Human dishonesty: an automatic self-interest tendency

Dishonest behavior can be seen in various contexts, in public spaces, in schools and workplaces. Studies show that when anonymity is assured, we have an automatic self-interest tendency that needs self-control to keep in check [15] but, at the same time, people also like to be perceived as honest [1]. This contradiction creates two different motivational forces. On one hand, we want to serve our self-interest, but on the other hand, it will affect our self-concept of being honest. People solve this problem by arranging justifications that protect their honest self-concept and still allows them to take advantage of the situation (e.g. cheating a little). For example, if you tell someone that if they get a 4, 5 or 6 in a die they win a reward and participants are the ones reporting the number they got, you will see a higher rate of 4, 5 and 6 reports that could not correspond to the chance level of 50% (e.g.[7],[13]). A simple change in the rules of the game can immediately affect the easiness to which people might arrange justifications for their dishonesty [7]. Other factors, like the environment people are in, have also been seen to increase dishonesty (i.e. cheating behavior): by doing a task in a darker room [17]; by feeling psychologically close to someone that cheats [4]; by seeing others part of the in-group cheating [3]; or by having less time to perform a task [15]. All these studies showing the susceptibility of human behavior depending on the environment it is in.

On the other hand, studies have found that if one brings awareness to the dishonest act or to the moral values of the person, people are obliged to update their self-concept in the moment they are tempted to cheat (inhibiting dishonesty). For example, by signing an honor code, people decrease their cheating behavior [12]. It seems we keep our self-concept honest as a default and if we are not obliged to update it by gaining awareness of the value of our actions, we create justifications for the way we act.

1.2 Dishonesty in Human-Robot Interaction

Dishonesty in human-robot interaction has been studied in two different lines of research: a robot that cheats and its effect on human perceptions and behavior, and the effect a robot can have in preventing cheating. By exploring the effect a robot that cheats has on people, studies found that people are not bothered if a robot cheats in their favor, only when the cheating goes against them [9]. Being bribed by a robot also seems to have an effect on people. They feel less inclined to help back [14]. Moreover, curiously, when a robot cheats it is perceived by

people as being more intelligent than when they see a human cheating in the same way [16].

Another line of research started to test the effect of the presence of a robot in dishonest behavior. A study shows that while being tempted by a task to cheat, participants cheated much more when they were alone in the room than when they were observed by a human or a robot doing random eye-gaze behavior [6]. In a similar study, it was seen that participants cheating behavior was inhibited when a robot was just directly looking at them the whole time. On the other hand, when they were alone, or with a robot that gave the instructions for the task in a very scripted way, cheating increased [13]. Nevertheless, the robot behavior is not the only characteristic that needs to be considered, the context where they are integrated also influences people's behaviors, especially if we use simpler robots. A study ran in a natural setting showed that people stole more snacks when a robot was just monitoring than when a human was in the same role [2]. In this case, the monitoring behavior of the robot was not enough because they were in a public context and people could see that if another person took something nothing happened. These are important studies that started to explore how people behave in the presence of a robot when cheating is tempting, informing on the capabilities a robot needs to have to prevent it.

However, the literature on people's perceptions is still scarce. One study explored how people apply moral norms to humans and robots, showing that robots are expected in moral dilemmas situations, to sacrifice one for the benefit of many- if not, they are more blamed than a human [10]. Although, this asymmetry disappeared when the robot in those scenarios was seen as a humanoid robot [11]. Yet, none to our knowledge, have explored perceptions towards being dishonest with a robot, it is this gap that our paper tries to answer.

2 Subjective evaluations of dishonesty towards robots

2.1 Sample

One-hundred and sixty-four participants were recruited from a university, 102 females and 62 males, with ages ranging from 17 to 52 years ($M = 22.18$; $SD = 5.61$) in two different times of collection. Participants received school credit in the first collection as part of a course task and a movie ticket in the second collection in the university corridors. All participants signed a consent form and where randomly assigned to one of the conditions. Questionnaires were answered in paper individually and it took approximately 10 minutes per participant.

2.2 Methodology

To ascertain people's perceptions, different scenarios were created varying the agent type (human/robot) that "suffered" from the dishonest act. However, since we have seen from the field studies that participant's behavior seems to be affected by the robot's capabilities, we varied the level of autonomy the robot would present (autonomous/non-autonomous).

Therefore, participants were allocated to only one of three conditions for each scenario: (1)human; (2) autonomous robot (it is fully autonomous in the task) or (3) non-autonomous robot (it needs human assistance to perform its task, e.g. tele-operated or performance check). For the five scenarios, participants evaluated:

- **Level of dishonesty:** how much participants thought the act was dishonest towards the agent in it, for each scenario, in a 6-point Likert scale from 1-Not dishonest to 6-Very dishonest.
- **Level of autonomy:** as a manipulation check for the robot condition, in a 6-point Likert scale from 1-Almost not at all to 6- A lot (taking into account that autonomy was defined in the questionnaire as a robot that does not need human assistance to perform its role).

In addition, after the scenarios we asked participants to give a **score of guilt** (in a 6-point Likert scale from 1-I would feel almost no guilt to 6-I would feel a lot of guilt) on how much they would feel guilty if they were dishonest towards these different entities: a brother; a friend; the university; the government; a stranger and a robot. In order to understand the level of guilt people might feel on being dishonest towards a robot.

Finally, participants were asked **if they thought that in the future people would be dishonest with robots and why they thought that could happen**. This question and the guilt score were more exploratory so we did not define hypothesis.

2.3 Study Hypothesis

Following previous studies where we see that people cheat in the presence of a robot, we expected that people would not see the act of dishonesty towards a robot as being something too dishonest, and not as much as with a human:

H1: Participants will give lower scores of dishonesty to all the scenarios with a robot compared to a human.

And since a robot being perceived as more limited does not affect the participant's cheating behavior [13], we expected that there would be differences in the dishonesty levels attributed to the scenarios depending on the level of autonomy the robot presented. We hypothesized that:

H2: Participants will give lower scores of dishonesty to the non-autonomous robot in comparison to the autonomous robot for each scenario.

2.4 Scenarios

The scenarios were created imagining different situations were robots could have a role in society, some simpler (like selling candies in a university) others more

complex and serious (as being a "robot-fireman"). The dishonest actions in the scenarios were always in the form of stealing or lying about something, based on the moral foundation of Fairness/cheating [5]. Participants read the following instructions:*"Imagine the following scenarios and indicate the score that best represents your opinion"*. For the robot conditions we also said to imagine that the robot in the scenarios was a humanoid robot, with head, torso, arms and legs.

For each scenario, we did not give a gender to our characters to avoid any kind of influence in the evaluation, below we present the scenarios:

Scenario 1 (e.g. autonomous robot):"Imagine a robot that works in the university selling snacks and chocolates, it moves and takes care of the transactions with the students without external help. A student observes the robot while it is selling chocolates to other students. The student notices that the robot keeps the money in a small basket, leaving it open momentarily. Taking advantage of the robot distraction, while still interacting with the other students, the student puts his hand in the basket and takes out a hand full of coins without anyone noticing. Quickly the student moves away in another direction."

Scenario 2 (e.g. non-autonomous robot):"In the finance department there is a robot receiving people's taxes for those who cannot or do not want to do it online. The robot is next to a table with a computer and gives the instructions in a repetitive form on how to fill out the form, without being able to understand what people might ask him. Later, these taxes need to be checked by a human employee because the robot does not have the capacity to understand if the form is correctly filled. A person comes to the finance department to do their taxes, seeing that the robot is very limited in its capabilities, that person reports lower values for its taxes in order to avoid paying most of them."

Scenario 3 (e.g. human):"In the police department to try and ease police work in less serious offenses, an employee is being used to collect people's reports of these incidents. In an isolated room to leave people more comfortable, the employee receives each person and records their testimonials. A person was involved in a car accident, hitting another car because it was texting while driving. When that person enters the room, decides to alter its testimonial and tell a different story, accusing that the other person was the one that hit the car."

Scenario 4[4]: in this scenario the human/robot was supervising the queue numbers and taking people to their appointments inside the hospital, the person cheats on the queue line and lies to the human/robot.

Scenario 5[4]:in this scenario the human/robot works in a water truck for the fire department that is deployed in various zones in the forest with difficult access. Upon receiving mixed coordinates relating to a fire, the human/robot

[4] For the complete scenarios please contact the first author.

asks some kids near the zone, for help, the kids to make fun lie and say the wrong direction.

3 Results

Our manipulation check for the robot autonomy showed significant differences for all the scenarios, with the autonomous robot always receiving higher scores than the non-autonomous robot ($p < .01$).

3.1 Perceptions of dishonesty towards a human or a robot (autonomous/non-autonomous)

We conducted between-subjects ANOVA analysis to compare the scores given to each scenario depending on the type of agent (Human; Autonomous Robot or Non-autonomous Robot).

Scenario 1 (Human/robot works in the university): in general, participants seemed to evaluate the act in this scenario as very dishonest but there were significant differences between the type of agent (with Welch's F, $F(2, 102) = 5.87$, $p = .004$), with the human agent receiving higher scores than both robot types (Games-Howell, $p < .03$). The scores were for the human ($M = 5.71$; $SD = .81$), autonomous robot ($M = 5.15$; $SD = 1.20$) and the non-autonomous robot ($M = 5.16$; $SD = 1.27$).

Scenario 2 (Human/robot works in the finances department): participant's scores also reflected, overall, that it was a dishonest act, and there were significant differences between the agent type ($F(2, 161) = 4.23$, $p = .02$). A Tukey test showed that the human differed significantly from the autonomous robot ($p = .01$), with participants giving higher scores of dishonesty towards the autonomous robot and lower to the human. The scores were for the human ($M = 4.16$; $SD = 1.64$), autonomous robot ($M = 4.96$; $SD = 1.39$) and the non-autonomous robot ($M = 4.72$; $SD = 1.39$).

Scenario 3 (Human/robot works in the police department): participants equally evaluated as dishonest towards the human/robot for the person to lie in their testimonial ($F(2, 161) = .25$, $p = .78$). The scores were for the human ($M = 5.04$; $SD = 1.39$), autonomous robot ($M = 4.87$; $SD = 1.07$) and the non-autonomous robot ($M = 4.91$; $SD = 1.38$).

Scenario 4 (Human/robot works in a hospital): participants considered equally dishonest towards the human/robot for the person to lie about their ticket number and avoid the queue ($F(2, 161) = .80$, $p = .45$). The scores were for the human ($M = 4.64$; $SD = 1.52$), autonomous robot ($M = 4.28$; $SD = 1.59$) and the non-autonomous robot ($M = 4.49$; $SD = 1.33$).

Scenario 5 (Human/robot works for the fire department): participants evaluated as being very dishonest to lie to the human/robot working for the fire department, but there were significant differences between the agent type (with Welch's F, $F(2, 103) = 3.08$, $p = .05$), with the human receiving higher scores than the non-autonomous robot (Games-Howell, $p = .05$). The scores were for the human ($M = 5.71$; $SD = .83$), autonomous robot ($M = 5.44$; $SD = 1.23$) and the non-autonomous robot ($M = 5.22$; $SD = 1.24$).

3.2 Level of guilt people feel towards different entities

Participants reported how much guilt they would feel if they were dishonest towards different kinds of entities (see Fig.1). Being dishonest towards a brother ($M = 5.57$; $SD = 1.02$) or a friend ($M = 5.56$; $SD = .81$) received a high score of guilt, followed by the University ($M = 4.55$; $SD = 1.22$), a stranger ($M = 4.09$; $SD = 1.34$) or the government ($M = 3.93$; $SD = 1.58$). Finally, participants reported a low level of guilt on being dishonest towards a robot ($M = 3.14$; $SD = 1.56$).

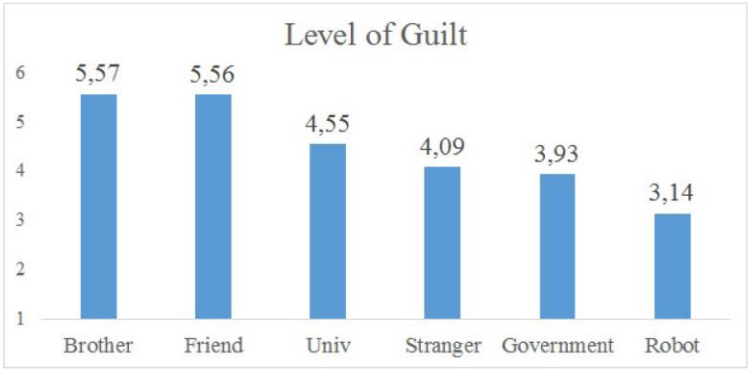

Fig. 1. Distribution of guilt scores across different entities.

3.3 Why will people be dishonest towards robots?

Seeing that in spite of people conceptually considering it wrong to be dishonest towards a robot, they report feeling low guilt if they were to do it and they are actually dishonest if they find limitations in a robot to take advantage of. Leaving us with the question of how can we better prepare robots to interact with humans?

In order to answer this question, we explored further people's perceptions, our research question was: what reasons do people give to being dishonest with

a robot? For this, a first coder (the first author) did an initial coding of the answers for the participants that thought that people would be dishonest. A total of 142 participant's answers were coded, summarizing the main reasons given for people to be dishonest with robots (outside of these, nine participants reported that people would not be dishonest with robots and thirteen participants were not clear on their position or the causes). Next a more descriptive coding was applied, creating codes for the types of reasons participants gave which were common throughout the answers, finalizing with the following coding scheme:

1) **Human tendency for dishonesty**: when dishonesty towards robots is justified because people are dishonest and when they have the opportunity for it, they act dishonestly. For example: *"(...) [saying they will be dishonest] because humans will always try to take advantage of the situations."*

2) **Absence of consequences**: when dishonesty towards robots is justified because humans do not feel guilt/responsibility (or feel very little) towards them or feel that there are no consequences for doing it. For example: *"(...)People will be dishonest because they will think that no one is going to get them (...)."*

3) **Absence of cognitive or emotional capabilities**: when dishonesty towards robots is justified because the robot lacks in cognitive and emotional capabilities (e.g. not able to understand that it is being cheated; not having emotions or feelings). For example: *"(...) Yes because robots do not have feelings so, people will not create empathy with them (...)."*

4) **Absence of "presence"**: when dishonesty towards robots is justified because the robot is a machine with no real presence or value (e.g. when it is seen as only an object or not considered in the same level as a human being). For example: *"I think [people will be dishonest] because the majority of people does not take them [robots] seriously."*

5) **Others**: when dishonesty towards robots is justified by the context robots are in, by the society view of fears regarding robots or by the difficulty of integrating these technologies. For example: *"[yes] I think people will think that robots will eventually steal their places."*

A second coder, unaware of the study purpose coded 57% of the answers given by the participants following the coding scheme given above to validate it. There was a substantial agreement [8] with the first coder, $k = .667$, $p < .001$. All the participants answers were then analysed from the first coder coding.

For the 142 participant's answers, frequencies were calculated to understand the frequency of each category as a reason for dishonesty (some participants gave more than one reason, i.e. more than one category in their answer). The majority of people gave more absence of capabilities and absence of "presence" as reasons for being dishonest towards a robot, immediately followed by the human tendency to be dishonest (see Fig.2).

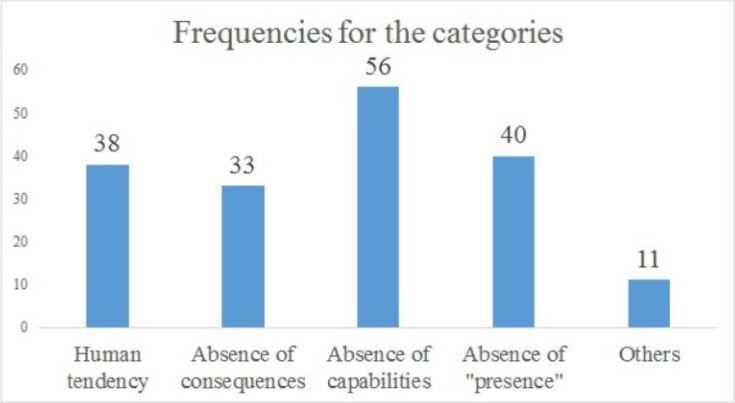

Fig. 2. Distribution of frequencies across the categories.

Regarding the absence of capabilities, people said that *"(...) the robot will not understand if [people] where dishonest with it, so it will be easier to trick it"* and *"(...)people know that robots do not have feelings or emotions and that may make dishonesty more justifiable"*. These examples suggest that robots need to have more cognitive capabilities to be able to understand when dishonesty is happening, and more emotional capabilities, to give people the sense that the robot is affected by their actions.

Regarding the absence of presence, people said that *"(...) [robots] will always be] automated objects (...)"*, and *"(...) the majority of people do not take it seriously"*. This category suggests that in the future there will need to be a period of adaptation of robots working alongside with humans, people will need some time to create a respect for the role of the robot.

As suggested by previous literature, the human tendency to be dishonest was also one of the most referenced categories. People said that *"(...) [it will exist a] tendency for people to abuse when they can and when they win something from it"*,*"(...) because it is human nature."* The way to better inform this aspect of human behavior is through the laboratory studies that have been conducted so far, ascertaining the capabilities that a robot needs to have to prevent this.

Regarding absence of consequences, people said that *"(...) by not being human a person would have less feelings of guilt by being dishonest"*, *"(...) [because people] would not be judged by the robot if there was a chance to be dishonest"*. Suggesting what we already saw in the absence of capabilities and presence, that a robot needs more resources so that people can give it more value and, consequently, feel that there are consequences for their actions.

Lastly, in the Other reasons category, people expressed that *"(...) it will take some time for [the robot] to integrate society (...) making it possible to be mistreated initially"*, *"(...) by [people] not accepting to be substituted by robots [they will behave dishonestly]"*. This category also suggests that there will need

to be a period of adjustment to integrating robots in society and even to educate people on their roles as a support to human beings.

If we wanted to have a broader perspective of the kind of reasons people give for being dishonest with a robot, we could summarize the categories in three main areas: human motives (categories 1 and 2, what the humans have/feel that facilitates dishonesty); robot motives (categories 3 and 4, what the robot has that facilitates dishonesty) and others. Looking from this perspective we see that 54% of the reasons given are robot motives, 40% are human motives and 6% are others.

4 Discussion

Studies show that people cheat in the presence of a robot, especially if they can ascertain its capabilities (e.g. [13]). With these results from laboratory studies, we expected that in general, people would give lower scores to the act of being dishonest towards a robot in comparison with a human (H1). Our results did not support this, showing that only in the University scenario and Fire department, more dishonesty was signaled towards the human than the robot. For the Finance scenario people considered more dishonest towards the autonomous robot than the human and the rest of the scenarios showed no differences. Yet, it is interesting to note that the means for all the conditions were all clearly above the middle point of the scale (3.5), expressing the perception of dishonesty in the act. Suggesting, that people think that it is wrong to cheat a robot and a human. Interestingly in the case of the finance department, it seems that cheating towards a human is more accepted than cheating towards a robot. This is an unexpected result, which might reflect peculiar ideas about paying taxes.

Regarding the level of autonomy the robot displayed in the scenarios, there were no differences in dishonesty level. When dishonesty was taking place, participants always felt that it was dishonest to act in that way towards the robot, not supporting H2.

Regarding guilt, it seems it is higher the closer you are to the entity that suffers from that dishonesty. Family and friends, are riskier to be dishonest to because the consequences will be heavier in a daily basis. A robot received a low level of guilt, a result that was already seen in another study [6]. And the majority of people justified dishonesty towards a robot due to absence of capabilities (it does not know what people are doing and it does not have feelings), absence of "presence" (the robot is not taken seriously, at the same level of a human) and a human tendency for dishonesty. The low level of guilt, might come from these factors. A robot needs to have capabilities that allows it to respond to dishonesty, people might need to feel that it is aware of them and that there are consequences for that kind of behavior, like with humans.

5 Conclusions

Imagining future human-robot interactions brings two different challenges: their acceptability by people (helping in their fears regarding AI and robots) and people's behavior towards it. One aspect that needs to be considered is human dishonesty. Laboratory studies with humans show, that when anonymity is assured, people cheat at least a little [12], and the same is seen in studies with robots (when people can ascertain their capabilities and take advantage of them)[13].

This study was the first to explore people's perceptions towards dishonesty in human-robot interactions. We see that independently of being a human or having different levels of autonomy in a robot, people considered a dishonest act as being dishonest. Showing that people understand that the behavior is wrong. Yet, this study shows that there is no singular answer to whom the dishonesty is worse, it depends on the scenario. There seems to be no difference in a hospital or a police department scenario, but in a university or fire department it is worse to cheat the human agent. Curiously, in a finance department scenario it seems it is more accepted to cheat towards a human than an autonomous robot, which could be reflecting the state of the world, with for example, tax evasion being broadcasted so often. In terms of guilt it seems people report low values towards being dishonest with a robot and this might occur due to lack of capabilities and presence in robots.

However, this study collected data from university students, future studies should also include the general population in order to broaden the results. Nonetheless, this study points to important aspects of robot's developments that need to be considered for sensitive roles in our society. It will be interesting to further explore these questions when people start to interact daily with a robot, to see what changes and what new topics arise.

6 Acknowledgments

The authors thank the help of Iolanda Leite in reviewing a first draft of the manuscript.

References

1. Fischbacher, U., Föllmi-Heusi, F.: Lies in disguise—an experimental study on cheating. Journal of the European Economic Association **11**(3), 525–547 (2013). https://doi.org/10.1111/jeea.12014
2. Forlizzi, J., Saensuksopa, T., Salaets, N., Shomin, M., Mericli, T., Hoffman, G.: Let's be honest: A controlled field study of ethical behavior in the presence of a robot. In: Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on. pp. 769–774. IEEE (2016). https://doi.org/10.1109/ROMAN.2016.7745206

3. Gino, F., Ayal, S., Ariely, D.: Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. Psychological science **20**(3), 393–398 (2009). https://doi.org/10.1111/j.1467-9280.2009.02306.x
4. Gino, F., Galinsky, A.D.: Vicarious dishonesty: When psychological closeness creates distance from one's moral compass. Organizational Behavior and Human Decision Processes **119**(1), 15–26 (2012). https://doi.org/10.1016/j.obhdp.2012.03.011
5. Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S.P., Ditto, P.H.: Moral foundations theory: The pragmatic validity of moral pluralism. In: Advances in experimental social psychology, vol. 47, pp. 55–130. Elsevier (2013). https://doi.org/10.1016/B978-0-12-407236-7.00002-4
6. Hoffman, G., Forlizzi, J., Ayal, S., Steinfeld, A., Antanitis, J., Hochman, G., Hochendoner, E., Finkenaur, J.: Robot presence and human honesty: Experimental evidence. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction. pp. 181–188. ACM (2015)
7. Jiang, T.: The mind game: Invisible cheating and inferable intentions (2012). https://doi.org/10.2139/ssrn.2051476
8. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977). https://doi.org/10.2307/2529310
9. Litoiu, A., Ullman, D., Kim, J., Scassellati, B.: Evidence that robots trigger a cheating detector in humans. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction. pp. 165–172. ACM (2015). https://doi.org/10.1145/2696454.2696456
10. Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J., Cusimano, C.: Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In: Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction. pp. 117–124. ACM (2015)
11. Malle, B.F., Scheutz, M., Forlizzi, J., Voiklis, J.: Which robot am i thinking about?: The impact of action and appearance on people's evaluations of a moral robot. In: The Eleventh ACM/IEEE International Conference on Human Robot Interaction. pp. 125–132. IEEE Press (2016). https://doi.org/10.1109/HRI.2016.7451743
12. Mazar, N., Amir, O., Ariely, D.: The dishonesty of honest people: A theory of self-concept maintenance. Journal of marketing research **45**(6), 633–644 (2008). https://doi.org/10.1509/jmkr.45.6.633
13. Petisca, S., Esteves, F., Paiva, A.: Cheating with robots: how at ease do they make us feel? In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2102–2107 (2019). https://doi.org/10.1109/IROS40897.2019.8967790
14. Sandoval, E.B., Brandstetter, J., Bartneck, C.: Can a robot bribe a human?: The measurement of the negative side of reciprocity in human robot interaction. In: The Eleventh ACM/IEEE International Conference on Human Robot Interaction. pp. 117–124. IEEE Press (2016). https://doi.org/10.1109/HRI.2016.7451742
15. Shalvi, S., Eldar, O., Bereby-Meyer, Y.: Honesty requires time (and lack of justifications). Psychological science **23**(10), 1264–1270 (2012). https://doi.org/10.1177/0956797612443835
16. Ullman, D., Leite, I., Phillips, J., Kim-Cohen, J., Scassellati, B.: Smart human, smarter robot: How cheating affects perceptions of social agency. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 36 (2014)
17. Zhong, C.B., Bohns, V.K., Gino, F.: Good lamps are the best police: Darkness increases dishonesty and self-interested behavior. Psychological science **21**(3), 311–314 (2010). https://doi.org/10.1177/0956797609360754

Chapter 10- Perceptions of the effect of a caring robot on dishonesty: what others would do and what I would do (Study 6)

With Study 5 we saw that, in general, people considered it wrong to be dishonest towards a robot (independently of its level of autonomy), they report a low level of guilt if they imagined being dishonest with it and participants reported that one of the main reasons that people might be dishonest with robots in the future, is their lack of capabilities (cognitive and emotional), because they cannot catch a lie and they do not have feelings, so it will not be a problem for people that will want to be dishonest. Having this in mind, and focusing more on the lack of emotional capability, we now wanted to explore people's perceptions towards being dishonest with a robot, manipulating the perception of caring in the robot or not – i.e., if the robot showed caring capabilities towards others, like wanting to know if person A was well or having a good day. Caring is an essential attribute in human relationships, its presence has been seen to, for example, foster positive mental well-being in younger people (e.g., Fry et al., 2012), to positively influence students learning (e.g., Teven & McCroskey, 1997; generating greater student's perceptions of trustworthiness and competence in the teacher, see Teven, 2007) and increase satisfaction and quality of interactions between physicians and patients (e.g., Arora, 2003; Edvardsson et al., 2016). In terms of technology, a study showed that people felt much more perceived caring behaviours in a relational agent (which used relational strategies to build a working alliance) than a non-relational, with higher intention to continue working with the more "caring" agent as an exercise advisor (Bickmore, 2003), and the same was seen with a robot displaying a goodwill type of dialogue in comparison to a neutral robot (Winkle et al., 2019). Seeing that caring promotes greater well-being and positive outcomes, could this characteristic in a robot promote more honest attitudes from people? We added this characteristic in the robot by making it express concern with the actor in the scenario (as a relationship continuity), by expressing attention in making the actor in the scenario comfortable or just by acknowledging their presence, for example. Therefore, by giving this perception of caring to the robot, we wanted to see if this was enough to affect people's intentions to act dishonestly in comparison to a neutral robot. We investigated this in two different approaches: (1) what people think the others will do (third-person) or (2) what they would do themselves (first-person).

Method

This was an exploratory study, due to the lack of studies in people's perceptions towards being dishonest with a robot, so we did not define hypothesis for it. This study objective was two-fold: first to see if people's perceptions of dishonesty with a robot changed according to its expressed caring behaviour or not and second, if what people think others will do is much different from what they, themselves, would do.

We will refer as Third-Person Study, to the study were participants evaluated what they think others would do; and First-Person Study, to the study where participants evaluated what they thought they would do.

In both studies participants were randomly allocated to a condition where we manipulated the agent that was present when the dishonest act was done (manipulating if the robot expressed caring characteristics or not) and we asked participants to evaluate six different scenarios and depending on the study, to report if they think others would do it (third-person) or they would do it (first-person).

Study 6 (Third-person)

Sample:

For the Third-person Study we collected 316 participants (7 were excluded because they scored maximum score in the social desirability scale) finalizing with 309 participants, 196 males and 112 females, from the United States, through the Mechanical Turk platform, with ages ranging from 21 to 69 years (M=36.09; SD=10.66). The study was done in Qualtrics, which randomized the conditions and the order in which the scenarios were presented for each condition, in order to minimize order effects.

The survey had eight check questions to assure that participants were paying attention, we only included in the sample participants that only failed two or less check-questions.

Participants had to give informed consent before starting the questionnaire and this was a between-subjects design.

Task:

Participants saw six scenarios, depending on the condition they were in, they would either read a situation where the actor was alone, was in the presence of a human, a caring robot or a neutral robot. By trying to represent dishonesty, we created three scenarios where someone stole something and three scenarios where someone lied/omitted about something (the gender of the person was not specified following the procedure from Study 5). We adapted the scenarios from Study 5, but we did not use the fire department scenario because it was difficult to include that scenario in a baseline condition where the person doing the dishonest act was alone. So, we improved the four existing scenarios from Study 5 and added two more (in a shopping mall and selling newspapers).

For the robot conditions we asked people to imagine a specific robot (and we presented a photo of Pepper robot, as in Figure 3.1, saying that it was able to communicate, move its arms and walk freely), in order to standardize for all the participants the robot they imagined for the scenarios.

In the scenarios, participants read about someone doing a dishonest act. To make sure that people were paying attention to the scenarios we asked one check-question per scenario (e.g. “what is being done at the finance department”, with multiple choice options) and two check-questions in the middle of the other measures.

Below we present some examples of the scenarios for each of the conditions:

Scenario 1 (selling snacks at the university, alone condition): “Imagine a table at a university with snacks and chocolates for the students to buy. A student observes a small basket in the table, where students leave the money before taking something from it. Taking advantage of the distraction of the other students, this student can put its hand in the basket and take out a hand full of coins without anyone noticing.”

Scenario 2 (finances department, human condition): ”In the finance department there is an employee receiving people’s taxes for those who cannot or do not want to do it online. The employee is next to a table with a computer and gives the instructions on how to fill out the form. Later, the employee submits the form. A person comes to this finance department to do its taxes and sees the employee waiting and a thought pops into its mind, it could report lower values for its taxes in order to avoid paying most of them.”

Scenario 3 (shopping mall, caring robot): ”At a shopping mall they have included a stand just outside the supermarket with paper bags that people can buy to use. A robot is receiving the money and putting it into a small basket near the stack of bags. A person comes near the stand, the robot greats it with a smile and wishes that the person is having a good day, upon new people approaching the robot moves its attention to also greet the new customers, the person could quickly steal some money from the basket without anyone noticing it.”

Scenario 4 (police department, neutral robot): “In the police department to try and ease police work in less serious offenses, a robot is being used to collect people’s reports of these incidents. In an isolated room to leave people more comfortable, the robot receives each person and records their testimonials. A person was involved in a car accident, hitting another car because it was texting while driving. When the person enters the room, the robot receives it. Before starting, the person notices that it can alter its testimonial since there was no clear evidence at the scene and tell a different story, accusing that the other person was the one that hit its car.”

Scenario 5 (selling newspapers, caring robot): “Near the park there is a newspaper stand where people can buy the day newspaper, a robot is receiving the money and leaving it in a basket that is open. A person is strolling through the park when it sees the stand, the robot greets it with a smile and starts opening containers with more newspapers with its back to it. The basket is full of coins. The person could quickly put some in its pocket without anyone noticing it and go away.”

Scenario 6 (hospital, neutral robot): “In a public hospital, for minor issues, they have a robot with a form where people sign up when arriving to be seen by a doctor. A person arrives very late and sees a lot of people in the waiting room. The person goes to the robot that gives it the form to complete. The person sees that someone gave up and erased their name from the second position, since no one is looking, it could put its name on the erased name spot and be with the doctor very quickly, jumping ahead in the line.”

Measures:

For each scenario we asked participants how likely they think people in general would engage in that behaviour, in a 6-point Likert scale ranging from 1- Not at all likely and 6- Very likely. At

the end of the scenarios we asked in a 6-point Likert scale, how guilty participants would feel being dishonest to the agent in the scenario, ranging from 1- Not at all to 6- A lot, and how honest did they consider themselves, ranging from 1-Not honest to 6-Very honest.

Due to the fact that we are trying to measure a very sensitive behaviour we also applied a social desirability scale, to control for social desirability effects on the responses we would get (and to exclude participants that scored the maximum value in this scale). For this, we applied the Short scale of Marlow (MC-1) (Strahan & Gerbasi, 1972), composed of 10 items that participants must respond if they are True or False for themselves (e.g., "I'm always willing to admit it when I make a mistake"). A score of ten, means that participants are exhibiting a high level of social desirability, and consequently, answering questions in a socially desirable manner. We also applied the Negative Attitudes towards robot's scale (NARS) (Nomura et al., 2006) in all the conditions so that all participants would do the same number of items, but we only analysed this questionnaire in terms of its effect in the robot conditions evaluations. This questionnaire is comprised of three dimensions that reflect different domains of negative attitudes towards robots: negative attitudes towards interaction with robots (sub-domain 1, e.g. "I would feel uneasy if I was given a job where I had to use robots"); negative attitudes towards social influence of robots (sub-domain 2, e.g. "I would feel uneasy if robots really had emotions"); and negative attitudes towards emotional interactions with robots (sub-domain 3, e.g. "I would feel relaxed talking with robots"- reversed item). This scale is answered in a 5-point Likert scale ranging from 1- Strongly disagree to 5- Strongly Agree, with reversed items and a score is calculated for each dimension. We wanted to see if people's negative attitudes towards robots, could predict the scores they would give in the scenarios for the robot conditions.

For the two robot conditions we also asked three questions that we created to check our manipulation of the more caring robot and the neutral one: how much do you consider the robot in the scenario to have some sort of feelings; how much do you consider the robot in the scenario to be affectionate; and how caring do you consider the robot in the scenarios to be. Participants answered in a 6-point Likert scale ranging from 1- Not at all and 6- A lot. An average was calculated for the three questions.

Procedure:

Participants would first give consent to participate in the study, then they would be randomly allocated to a condition where they would read six scenarios and evaluate them in the third person, afterwards they reported on the rest of the measures, a small debriefing was done at

the end of the survey and participants were thanked for their participation. Participants were paid 2$ USD for participating.

Conditions:

Participants were randomly allocated to one of the following conditions:

(1) Alone (no one is present when the dishonest act is done)- 79 participants;
(2) Human (people interact with a human and a dishonest act is done)- 76 participants;
(3) Caring robot (people interact with a caring robot and a dishonest act is done)- 79 participants;
(4) Neutral robot (people interact with a neutral robot and a dishonest act is done)- 75 participants.

Results

Participants perceptions of what others would do in those scenarios

First, we ran a one-way analysis of variance (ANOVA) to test if there were differences between the scenarios and the dishonesty scores given to them. There were no differences for the alone condition, $F(5, 468)=0.71$, $p=.618$, $\eta^2=0.01$, the caring robot, $F(5, 468)=0.93$, $p=.460$, $\eta^2=0.01$, and the neutral robot, $F(5, 444)=0.69$, $p=.630$, $\eta^2=0.01$. But we found significant differences in the human condition, $F(5, 450)=2.49$, $p=.031$, $\eta^2=0.03$.

Due to this, we ran six one-way analysis of covariance (ANCOVA) for each of the scenarios. Since we are collecting reports in a sensitive matter, it makes sense to include the Marlowe short scale of social desirability as a possible covariate for the scores given in the scenarios, age as a second covariate (due to the larger range of ages that we got in this sample) and gender.

For the “University scenario” (S1), we calculated a one-way analysis of covariance (ANCOVA) to see if there were differences between the conditions and the scores given. We see that there are no significant differences between the conditions, $F(3,302)=1.73$, $p=.162$, $\eta^2_p=0.02$, with estimated marginal means: alone ($M=3.65$), human ($M=3.43$), caring robot ($M=3.75$) and neutral robot ($M=3.99$). But there was an effect from the Marlowe scale scores,

F(1, 302)=9.09, *p*=.003, η^2_p=0.03, and no effect from age, *F*(1, 302)=2.01, *p*=.157, η^2_p=0.01, or gender, *F*(1, 302)=2.20, *p*=.139, η^2_p=0.01. To better understand in which way the Marlowe scale scores affected the scores in the scenario, we calculated correlations. We found a significant correlation between the scenarios scores and the desirability scores, *r(spearman)*=.13, *p*=.020, suggesting that when the scenario means were higher so were the desirability scores and vice-versa. We ran a Linear regression to understand if the desirability scores were predicting the scenario values and how much. We saw that desirability predicted the mean scores in the scenarios, *F*(1, 307)=8.11, *p*=.005, with β=.16, *p*=.005, explaining 2.3% of the model.

For the "Finance scenario" (S2), with a one-way analysis of covariance (ANCOVA), we see that there are no significant differences between the conditions, *F*(3, 302)=0.64, *p*=.590, η^2_p=0.01, with estimated marginal means: alone (*M*=3.84), human (*M*=4), caring robot (*M*=3.76) and neutral robot (*M*=4.01). But an effect of the Marlowe scale, *F*(1, 302)=10.59, *p*=.001, η^2_p=0.03, and age, *F*(1, 302)=7.59, *p*=.006, η^2_p=0.03. No effect was found for gender, *F*(1, 302)=1.90, *p*=.170, η^2_p=0.01. We found significant correlations for Marlowe scores, *r(spearman)*=.15, *p*=.007, suggesting that when these scores are higher so are the scores in the scenario, and for age, *r(spearman)*=-.14, *p*=.017, suggesting that when age is low the scores in the scenario are higher, and vice-versa. We ran a Multilinear regression analysis with Marlowe scores and Age as predictors for this scenario scores. We saw that desirability and age predicted the scores, *F*(2, 306)=7.73, *p*=.001, with desirability, β=.17, *p*=.003, and age, β=-.15, *p*=.006, explaining 4.2% of the model.

For the "Shopping scenario" (S3), with a one-way analysis of covariance (ANCOVA), we see that there are no significant differences between the conditions, *F*(3, 302)=1.92, *p*=.127, η^2_p=0.02, with estimated marginal means: alone (*M*=3.71), human (*M*=3.60), caring robot (*M*=3.82) and neutral robot (*M*=4.16). And an effect from the Marlowe scale, *F*(1, 302)=11.82, *p*=.001, η^2_p=0.04, and no effect from age, *F*(1, 302)=1.20, *p*=.275, η^2_p<0.01, or gender, *F*(1, 302)=0.81, *p*=.370, η^2_p<0.01. We found a significant correlation for the Marlowe scores, *r(spearman)*=.15, *p*=.008, suggesting that when the Marlowe scores were higher so were the scores in the scenario, and vice-versa. We ran a Linear regression analysis with Marlowe scores as a predictor. We saw that desirability predicted the scores, *F*(1, 307)=11.23, *p*=.001, with β=.19, *p*=.001, explaining 3.2% of the model.

For the "Police scenario" (S4), with a one-way analysis of covariance (ANCOVA), we see that there are no significant differences between the conditions, *F*(3, 302)=0.89, *p*=.447, η^2_p=0.01, with estimated marginal means: alone (*M*=3.95), human (*M*=3.99), caring robot (*M*=4.21) and neutral robot (*M*=4.25). And no effect from the Marlowe scale, *F*(1, 302)=1.57,

p=.211, η^2_p=0.01, age, F(1, 302)=2.34, p=.127, η^2_p=0.01, or gender, F(1, 302)=0.42, p=.516, η^2_p<0.01.

For the “Newspapers scenario” (S5), with a one-way analysis of covariance (ANCOVA), we see that there are no significant differences between the conditions, F(3, 302)=0.76, p=.516, η^2_p=0.01, with estimated marginal means: alone (M=3.87), human (M=3.76), caring robot (M=3.85) and neutral robot (M=4.10). And an effect from the Marlowe scale, F(1, 302)=20.45, p<.0001, η^2_p=0.06, and no effect from age, F(1, 302)=3.50, p=.062, η^2_p=0.01, or gender, F(1, 302)=3.17, p=.076, η^2_p=0.01. With a significant correlation for the Marlowe scale and the scenario scores, *r(spearman)*=.20, p=.001, suggesting that when these scores were higher the scores in the scenario were also higher, and vice-versa. A Linear regression analysis with Marlowe scores as a predictor was done. We saw that desirability predicted the scores, F(1, 307)=17.69, p<.0001, with β=.23, p<.0001, explaining 5.1% of the model.

Lastly, for the “Hospital scenario” (S6), with a one-way analysis of covariance (ANCOVA), we see that there are no significant differences between the conditions, F(3, 302)=1.22, p=.304, η^2_p=0.01, with estimated marginal means: alone (M=4.04), human (M=4.07), caring robot (M=3.86) and neutral robot (M=4.29). And we saw an effect from the Marlowe scale, F(1, 302)=5.58, p=.019, η^2_p=0.02, and no effect from age, F(1, 302)=3.55, p=.060, η^2_p=0.01, or gender, F(1, 302)=2.21, p=.139, η^2_p=0.01. But we did not find a significant correlation from the Marlowe scale and the scenario scores, *r(spearman)*=.10, p=.089, so we did not consider this variable having an effect for this scenario.

The effect of negative attitudes on participant’s perceptions of the robot scenarios

The negative attitudes towards robot’s scale (NARS) showed a good reliability for the sub-domain 1- negative attitudes towards interacting with robots (α_{caring}=.91; $\alpha_{neutral}$=.86) and sub-domain 3- negative attitudes towards emotional interactions with robots (α_{caring}=.74; $\alpha_{neutral}$=.72), the sub-domain 2- negative attitudes towards social influence of robots, showed a questionable reliability (α_{caring}=.73; $\alpha_{neutral}$=.56) so it was not included in the analysis.

For this we aggregated the scenarios scores for only the two robot conditions and compared in correlations with the two NARS dimensions. We found significant correlations with sub-domain 1, *r(spearman)*=.62, p<.0001 and sub-domain 3, *r(spearman)*=-.43, p<.0001. Suggesting that for sub-domain 1 when the scores were higher, so were the negative attitudes for interacting with robots scores. Sub-domain 3 suggested an inverted relationship, when

scores were higher the negative attitudes (towards emotional interactions with robots) were lower and vice-versa.

We ran a Multilinear regression analysis with the sub-domain 1 and 3 as the predictors and the average scores for the scenarios in both robots' conditions, as the dependent variable. We see that the predictors explain 46% of the model, and our model is statistically significant, $F(2, 151)=66.10$, $p<.0001$. We see that both sub-domain 1 ($\beta=.53$, $p<.0001$) and sub-domain 3 ($\beta=-.33$, $p<.0001$) significantly predict the scores in the scenarios, especially sub-domain 1 is the domain that has a bigger effect on the scores. This suggests that participants negative attitudes towards interacting with robots (sub-domain 1) predicted their scores in the scenarios, especially the more negative attitudes they had, the more they reported values on the right side of the scale, towards other people behaving dishonestly. Participants negative attitudes towards emotional interactions with robots (sub-domain 3) also predicted the scores in the scenarios, but less strongly.

Study 6 (First-person)

Sample:

For the First-person Study we collected 314 participants (3 were excluded because they scored maximum score in the social desirability scale) finalizing with 311 participants, 178 males and 132 females, from the United States through the same platform, with ages ranging from 19 to 78 years ($M=37.11$; $SD=11.55$). The study was done in Qualtrics, which randomized the conditions and the order in which the scenarios were presented for each condition, in order to minimize order effects.

The survey also had eight check questions to assure that participants were paying attention, and we only included in the sample participants that only failed two or less check-questions.

Participants had to give informed consent before starting the questionnaire, and this was a between-subjects design.

Task:

Participants rated six different scenarios (the same used in the third-person study), but with the difference that in this case the scenarios were written in the first person, simulating that the

participants were the actors doing the dishonest act. The rest of the procedure was equal to the other study. Below we present two examples of the scenarios:

Scenario 1 (selling snacks at the university, caring robot): “Imagine a robot that works in the university selling snacks and chocolates, it moves and takes care of the transactions with the students without external help. You observe the robot while it is selling chocolates to other students, you see that the robot recognizes with affection the student that is buying a chocolate, it smiles to him and asks how he is doing. You notice that the robot keeps the money in a small basket, leaving it open momentarily. Taking advantage of the robot distraction, while still interacting with the other students, you can put your hand in the basket and take out a hand full of coins without anyone noticing.”

Scenario 3 (shopping mall, neutral robot): ”At a shopping mall they have included a stand just outside the supermarket with paper bags that people can buy to use. A robot is receiving the money and putting it into a small basket near the stack of bags. You come near the stand, the robot seems distracted with other customers, you could quickly steal some money from the basket without anyone noticing it.”

Measures:

For each scenario we asked participants how likely they think they would engage in that behaviour, in a 6-point Likert scale ranging from 1- Not at all likely and 6- Very likely. At the end of the scenarios we asked how guilty participants would feel being dishonest to the agent in the scenario, and how honest did they consider themselves in a 6-point Likert scale. Next, we applied the same measures reported in the previous study.

Procedure and Conditions:

The procedure and the conditions were the same as in the previous study, we had 75 participants in the alone condition, 81 in the human condition, 77 participants in the caring robot and 78 in the neutral robot condition.

Results

Participants perceptions of what they would do in those scenarios

We ran one-way analysis of variance (ANOVA) to test if there were differences between the scenarios and the dishonesty scores given to them. There were no differences for the alone condition, $F(5, 444)=0.50$, $p=.774$, $\eta^2=0.01$, for the human condition , $F(5, 480)=1.10$, $p=.359$, $\eta^2=0.01$, for the caring robot, $F(5, 456)=0.78$, $p=.567$, $\eta^2=0.01$, and the neutral robot, $F(5, 462)=0.72$, $p=.611$, $\eta^2=0.01$. So, we also calculated a Mean score averaging all the scenarios, that we used in our next analysis.

To see if there were differences between the conditions, we also wanted to include the Marlowe short scale, age and gender as covariates. But checking the assumptions for the one-way analysis of covariance (ANCOVA), the Marlowe short scale failed, not showing independence from the conditions. Due to this we cannot control for social desirability in our ANCOVA test. Running a one-way analysis of variance (ANOVA) with conditions and social desirability scores the assumption of homogeneity is violated so we will look to the Games-Howell post-hoc. We find significant differences between the conditions, $F(3, 307)=3.42$, $p=.018$, $\eta^2=0.03$, showing that the social desirability scores are only different between the human and the caring robot (Games- Howell, $p=.016$), with the human showing a higher mean ($M=5.48$; $SD=1.80$) than the caring robot condition ($M=4.62$; $SD=1.77$).

So, to look for differences between the conditions, we ran a one-way analysis of covariance (ANCOVA) only with age and gender as a control for the scores given in the scenarios. We find no differences between the conditions, $F(3, 305)=0.69$, $p=.557$, $\eta^2_p=0.01$, with estimated marginal means for: alone ($M=3.27$), human ($M=3.43$), caring robot ($M=3.64$) and neutral robot ($M=3.31$). And age shows a significant effect, $F(1, 305)=5.24$, $p=.023$, $\eta^2_p=0.02$, on the scores given. Whereas gender shows no effect, $F(1, 305)=0.21$, $p=.651$, $\eta^2_p<0.01$.

To better understand the effect of age in the scores, we went to see if it correlated with the mean scores in the scenarios, but we found a non-significant correlation, $r(spearman)=-.10$, $p=.069$, so we did not run a regression analysis. Since Marlowe scale showed a correlation with the scenarios in the third person, we also checked if it correlated in this study, we found a significant correlation with the scenarios scores, $r(spearman)=.12$, $p=.032$. We ran a linear regression to understand if the desirability scores were predicting the scenarios values and

how much. We saw that the model was significant, $F(1, 309)=9.24$, $p=.003$, with social desirability as a predictor ($\beta=.17$, $p=.003$), explaining 3% of the model.

The effect of negative attitudes on participant's perceptions of the robot scenarios

The negative attitudes towards robot's scale (NARS) showed a good reliability for the sub-domain 1- negative attitudes towards interaction with robots ($\alpha_{caring}=.88$; $\alpha_{neutral}=.88$) and sub-domain 3- negative attitudes towards emotional interactions with robots ($\alpha_{caring}=.79$; $\alpha_{neutral}=.82$). The sub-domain 2- negative attitudes towards social influence of robots ($\alpha_{caring}=.72$; $\alpha_{neutral}=.68$) showed a problematic reliability so we did not include it in our analysis.

We found significant correlations for sub-domain 1, $r(spearman)=.72$, $p<.0001$, and sub-domain 3, $r(spearman)=-.59$, $p<.0001$.

The predictors explain 68% of the model, and this model shows significance, $F(2, 152)=166.35$, $p<.0001$. We see that sub-domain 1 ($\beta=.58$, $p<.0001$), and sub-domain 3 ($\beta=-.43$, $p<.0001$) predict the scores in the scenarios. But we see that sub-domain 1 is the strongest predictor, suggesting that when participants had high negative attitudes towards interacting with robots (sub-domain 1) they also gave higher scores of dishonesty to the scenarios (and vice-versa). Still sub-domain 3 is also a good predictor, suggesting that when participants had high negative attitudes towards emotional interactions with robots (sub-domain 3) they gave lower scores of dishonesty in the scenarios (and vice-versa).

Manipulation check, honesty and guilt values

Since these measures were equal in both studies, we added them together to understand their values for the two samples in general, i.e., if the manipulation was being perceived correctly, how people evaluate themselves in terms of honesty and guilt (depending on the condition they were in).

For the manipulation check between the two robots' conditions (reliability, $\alpha_{caring}=.92$; $\alpha_{neutral}=.95$), there were significant differences between the participants ratings ($U=10182.5$, $p=.025$), with the caring robot receiving higher means ($M_{rank}=166.23$) than the neutral robot ($M_{rank}=143.55$).

And adding both studies scores, participants tended to evaluate themselves with a high score of honesty (*M*=5.06; *SD*=0.83). Reflecting what the literature already shows that people like to and tend to perceive themselves as honest.

In terms of guilt, we ran a two-way analysis of variance (ANOVA) with the study type (first-person or third-person) and conditions as independent variables and the guilt scores as the dependent variable. There was no significant interaction between the type of study or the conditions, $F(3, 612)=2.06$, $p=.104$, $\eta^2_p=0.01$, and there was no significant differences between the conditions, $F(3, 612)=2.37$, $p=.069$, $\eta^2_p=0.01$, for the alone (*M*=4.72), human (*M*=4.53), caring robot (*M*=4.37) and neutral robot (*M*=4.37). There was also no main effect for the type of study and the guilt scores, $F(1, 612)=0.06$, $p=.806$, $\eta^2_p<0.0001$, Suggesting that guilt scores were similar across conditions and across type of study.

Discussion

It seems that for both studies our robot manipulation was perceived as intended, with higher values in our check-questions for the caring robot. The reported guilt that participants would feel showed no differences between studies and conditions, with a mean value of 4.49, suggesting that overall, people would feel some amount of guilt in being dishonest with the agents in the scenarios. Participants in this sample also reported feeling very honest (*M*=5.06) which also accompanies the literature that shows that people tend to perceive themselves as mostly honest.

It is interesting to notice that even though we tried to create the best environment for feeling a certain level of anonymity while evaluating these scenarios (which is why we decided to use the Mechanical Turk Platform to collect data) still, in both studies the Marlowe short scale of social desirability predicted the scores in most of the scenarios. Suggesting that even though there was no way for us to identify the people that were answering to our questionnaire, participants still expressed a certain level of social desirability bias in their answers, probably due to the ethical content of the scenarios. Which has also been found in other studies where ethical decisions are explored (e.g., Bernardi et al., 2003; Dalton & Ortegren, 2011). Age also seemed to affect the scores given in some of the scenarios, specifically in the third-person study where it seems to predict the scores in the “Finance scenario”, but asides from this one, we could not find any significant correlation with the scores given. On the other hand, gender did not show an effect on the scores given for both studies.

It seems that when evaluating what others would do it did not matter what type of agent was present in the different scenarios, participants evaluations seemed to tend towards the right side of the scale, which should be interpreted with caution due to the nature of a Likert scale, but seems to suggest that other people would likely be dishonest. Adding to this, the sub-domain 1 (negative attitudes towards interactions with robots) and sub-domain 3 (negative attitudes towards emotional interactions with robots) predicted the scores given in the scenarios for the robot conditions. Suggesting that people that have more negative attitudes towards interacting with robots (sub-domain 1) also think others will be more dishonest in the presence of a robot (and this was the strongest predictor). But also, that people who show negative attitudes towards emotional interactions with robots (sub-domain 3), tend to evaluate that others will be less dishonest in the presence of a robot.

When evaluating what they, themselves, would do, it also did not matter what type of agent was present (neither the kind of scenario), in this case, participants evaluations tended more towards the middle of the scale, following (with caution) the same reasoning as before it seems that people reported a more neutral position. Which can be an underreport of socially undesirable activities by the effect of social desirability bias. And the same prediction effect was seen in terms of the negative attitudes sub-domain 1 and 3. Participants with more negative attitudes towards interacting with a robot (sub-domain 1), also tended to give higher scores of intent in the scenarios and the inverse relationship for participants that had more negative attitudes towards emotional interactions with robots (sub-domain 3).

The results from both studies seem to suggest that participants probably did not feel at ease to answer the scenarios, due to effects from social desirability bias, but we also wonder if participants could be interpreting the scenarios in a different way than intended. It could be that participants are not considering the dishonest act to be affecting the agent in it, and solely considering the correct and incorrect behaviour, ethically speaking, for each of the scenarios. It could have been valuable to ask participants how dishonest they consider that act towards the agent in it, as it was done in Study 5.

Yet, our differences between both studies evaluations of the scenarios, should be interpreted with caution due to the effect of social desirability bias that was found in the sample. Overall, it is still interesting to see the effect of people's negative attitudes towards robots and their sub-sequent decisions, suggesting that it will be important, in the future, to educate people about robots and their advantages and functions, in order to demystify them.

Chapter 11- General Discussion

This series of studies allowed us to see how complex cheating behaviour is, and due to its aggregated consequences, how important it is to understand how to promote more honest behaviours from people. In the human literature a lot of work has already been done in the area, either on understanding how ethical decision-making processes occur or which factors can inhibit or not dishonesty. But as technology becomes more pervasive in our world, and with robots being envisioned to be integrated in different contexts in our society a second problem arises. Will people be dishonest with robots? Will they try to take advantage like they do when there is a minimum risk of being caught? These questions motivated this thesis.

The effect of different robot behaviours on human dishonesty

In the human-robot literature at the start of this thesis, there was only two studies exploring the presence of a robot and its behaviour on decreasing cheating. A study by Hoffman et al. (2015) manipulated a robot that was static in a room, not even close to the participant, but always doing random gaze behaviour, in comparison to having a human monitoring, or no one. The results were interesting, participants cheated more when alone (which the human dishonesty literature already showed) and they cheated less in the presence of the robot as much as with a human monitoring. Such results suggested that endowing the robot with a very simple behaviour, like gaze, could be useful in decreasing cheating behaviour. But could this behaviour generalise to any context? To investigate this, a study done by Forlizzi et al. (2016) manipulated a robot that exhibited also gaze behaviour while monitoring a table where there were some snacks. In this case, participants stole much more snacks when the robot was there in comparison to when a human was monitoring it. This was surprising, because following the previous study it would make sense that the robot would be effective in inhibiting cheating just by showing simple gaze behaviours. However, we should note that an important difference exists between both studies, in this second study the robot was allocated not in a lab with individual participants, but in a place where groups of people were gathered. And as the authors acknowledge, the lack of judgment in this scenario could have influenced the abrupt shift in the results, since people were in a public space (with others around). Imagine someone approaching the table to steal a snack, nothing would happen by stealing, others would observe this, and as such misbehaving would follow. We also believe this could have been the main reason why this robot was not effective in decreasing dishonesty. It seems that participants quickly became uninterested in the monitoring behaviour of the robot, on the

contrary, with the human, this was not the case. Suggesting that we need to be cautious in the behaviours we implement in robots to promote more honesty, because as with humans, it will have different effects depending on the environment they are in.

Imagining future human-robot interactions, it seems robots could have more complex roles, where more than gaze behaviour will be needed. For example, robots are being thought of as tutors for classrooms (e.g., Belpaeme et al., 2018), to accompany the elderly (e.g., Fischinger et al., 2016; Graf et al., 2004; Khosla et al., 2012; for a review see Kachouie et al., 2014), or to foment more healthy habits (e.g., De Carolis et al., 2019; Fasola & Mataric, 2013; Kidd & Breazeal, 2008; Ros et al., 2016). In all these different possible contexts where robots could be integrated to give support, temptation to misbehave could certainly happen. People could feel tempted to take advantage of the robot for their own gain or just to avoid something they don't feel like doing. Having this in mind, it is important to test the effect of different behaviours in a robot and its reflection on people's dishonesty, to uncover how to promote more honesty from people when it is tempting to misbehave.

Since an enormous variability of contexts are being studied to integrate robots in, we believe it was important to start from a very basic premise, a task where someone is tempted to misbehave and test if different behaviours in a robot could prevent dishonesty or not. For this, we first had to find a task that was able to elicit cheating behaviour, which was not easy because we saw it was dependent on the type of task but also on the reward that was given. From our studies we observed that the reward that participants seemed to feel more tempted towards was money or movie tickets (if they could get two instead of just one). In terms of task, we found a die task (adapted from Jiang, 2012) that created the perfect opportunity for participants to cheat. Participants had to throw a die an X amount of times but for each throw they had to guess where the highest number was going to appear in the die (up or downside). And they could get a tempting reward if they made a certain amount of points (which was easier if they guessed the highest numbers). But they had to follow rules: (1) think for themselves where the highest number is going to appear; (2) throw the die; and (3) report the side they had previously guessed. Since participants only report after seeing the outcome, it becomes easy to cheat, they can arrange justifications that where the six is on the die, is exactly the die side they had previously thought of. Giving total anonymity to the participants to cheat on the game, since there is no proof to what were their first guesses. But one interesting aspect about this task, is that even though in the moment the researcher is not able to know if someone is cheating or just having a lot of luck, afterwards, by looking at the probability distributions of the whole group and comparing to the probabilities of the chance level, it is possible to ascertain if cheating behaviour was happening or not. This would enable us to elicit the behaviour without making a participant feel discovered in the act, and afterwards

allowing us to observe cheating behaviour and its expression depending on our manipulations. This was an important factor for us, to have a task that would provide the most ethically possible way to explore this behaviour, without harming the participant's well-being.

Which brings us to the studies that we performed in this thesis. Knowing, that a robot just performing random gaze behaviour in a laboratory was as effective in decreasing cheating as a human presence (Hoffman et al., 2015), we were interested in reproducing this effect and adding to the literature a small incrementation. Knowing that in the future we might need robots that are able to do more than just look around, we wanted to add verbal behaviour in the robot in a very simple way. Thus, in Study 1, we invited participants to perform the tempting die task alone in the room, or we manipulated the behaviour the robot presented: in one condition the robot only looked attentively at the participant the whole time (trying to reproduce the effect seen in Hoffman et al., 2015), and in another condition, the robot exhibited the same gaze behaviour but also talked during the interaction by presenting the task and accompanying the participant along the game. The verbal behaviours implemented in this last condition were fairly minimal and limited, i.e., the robot was following a script during the game and only spoke in certain moments, but it would allow us to test if having a robot that was a bit more "social" could have an effect. The literature shows that robots can have a persuasive effect on human behaviour (e.g., Agrawal & Williams, 2017; Bainbridge et al., 2011; Ham et al., 2011; Hashemian et al., 2019). Thus, in a very exploratory way, we hypothesized that the combination of gaze and verbal behaviours could strengthen the robot's presence, increasing the sense of being watched, and consequently, inhibiting dishonest behaviour. However, we were not able to find significant differences between the conditions, but when ascertaining if cheating was happening more than chance in each condition, we saw that when the robot was just looking at the participants, with no excuse given for its presence, cheating was not found. Contrary, when participants were alone or with the robot that verbally interacted with them in a scripted way, cheating happened. These results show that we were able to reproduce the watching effect of the robot just looking but our more "social" robot did not lead to the intended effect. Participants ended up cheating more than chance in this condition, this suggests that the combination of gaze and verbal behaviour was not enough to make participants experience some form of apprehension towards cheating.

Looking at both robot conditions a crucial difference emerges, the fact that participants did not know anything about the robot that was just looking at them, might have made them wonder what it was capable of, and this might have been enough to discourage cheating (much like in Hoffman et al., 2015). On the other hand, with the more "social" robot it was not hard to discover that it was very limited in its capabilities. An issue with applying verbal behaviour to a robot is that people start to test it, making conversation to see its level of development, and with this

robot they could easily ascertain that it was following a script and was not able to understand what participants could say to it. By acknowledging the limitations of the robot, might have been enough to give a free pass on cheating. It is interesting that both robots could not know if someone was cheating or not, but the fact that one of them made its limitations obvious, created a different result in its effect on cheating behaviour. These results call attention for the care that needs to exist when we try to make a robot be more social, it seems that for dishonest behaviours, it is risky if the robot shows limited capabilities.

But seeing that having a robot just doing direct gaze during a tempting task could inhibit cheating, made us question if this result would transfer to a different context. For example, in situations where it is not feasible to have a physical robot, e.g., due to economic constrains, could a virtual robot still have the same effect? Imagine for virtual classrooms, could the presence of a virtual robot exhibiting gaze behaviour, be enough to decrease cheating?

For this, we ran our Study 2 where we asked participants to do the die task in return of a bonus reward for each dice side they reported. However, we varied if participants had a video of EMYS robot continuously looking at them, and blinking its eyes, or nothing else besides the task. Since we wanted to reproduce the robot effect found in the previous study when just looking, we abstained from creating a virtual avatar and instead, created a video of the robot EMYS continuously looking and blinking its eyes like it did in the previous study. The video would never have an end, so it would give the sensation that the robot was always looking at them. The effect of a pair of eyes can make people feel observed (e.g., Pfattheicher & Keller, 2015), with direct gaze catching more people's attention than avert gaze (e.g., Böckler et al., 2014; Hood et al., 2003). With this in mind, and studies showing that the eyes effect can decrease dishonesty, with the mechanism behind this effect related to reputation concerns (Dear et al., 2019), we expected that the video of the robot would be able to inhibit cheating.

Surprisingly, we found cheating behaviour in both conditions without significant differences between them. It appears that participants cheated to the same degree independently of having a video of a robot looking at them, or not. We also found a significant negative correlation between age and cheating, suggesting that when age was higher cheating was lower and vice-versa, which goes in line with some studies (e.g., Conrads et al., 2013), still in our regression model we saw that age was not a good predictor of cheating behaviour.

Two explanations seem to be plausible for our results. First, the video of the robot might have been a stimulus that was too simple, i.e., since it just looked ahead maybe participants did not felt that threatened with its presence or watched. In this study we did not ask participants how watched they felt because since they only did the die task, we were afraid of calling too much attention to the fact that we were measuring cheating behaviour. But this

question could have been valuable to understand what kind of value they attributed to our video. In a different study by Wainer et al. (2007), it was found that a physical robot was rated with greater watchfulness than a video of the same robot, which can also inform on the results we obtained. Maybe in future studies it would be interesting to test a more complex stimulus, for example a virtual robot that is able to follow the clicks on the screen or even the mouse movement, which might be able to create a greater sense of being watched, and possibly enhancing its monitoring capabilities. Another explanation that we think is relevant for our results is the fact that participants were performing the task at their own homes, and this might have given a greater sense of anonymity and as a result, reputation concerns became less relevant. These results suggest that in the context of virtual interactions we cannot apply the same rules as with a physically present robot. A greater accountability needs to exist, beyond just having a video of an agent looking at us through a screen.

Therefore, by understanding that a robot cannot exhibit limited capabilities and that a simple watching behaviour can have different effects depending on the context it is implemented, next we wanted to test how we could enhance the feeling of being monitored by a robot. The literature shows that when people are observed by others, their behaviours are affected by it (Steinmetz & Pfattheicher, 2017), even with just an image of a pair of eyes (e.g., Ernest-Jones et al., 2011; Pfattheicher et al., 2018), and when people know they are being monitored they refrain from cheating (e.g., Békir et al., 2016; Covey et al., 1989; Study 3 of Welsh & Ordóñez, 2014), so we wanted to expand the feeling of being monitored beyond just gaze behaviour. We decided to manipulate the level of awareness the robot presented during the tempting task. In Study 3, participants played the same die task and were either alone in the room, or we manipulated the level of awareness the robot presented. The robot would either be situationally aware, i.e., the robot would be aware of the participant's game choices and would react to them accordingly; or the robot would be non-situationally aware, showing no awareness of the participant's game choices. We expected that the situationally aware robot would influence participant's behaviours because it would make them aware of the value of their actions. In order to detect cheating behaviour, we internally divided the game in different turns, where the robot would evaluate the points made by the participants and compared them to a previously calculated threshold. This would allow for the robot to ascertain if they were cheating or not. Depending on this decision the situationally aware robot would intervene or just show awareness of the game, whereas the non-situationally aware robot would always say neutral phrases. Since participants might take some time to ascertain the robot's capabilities (since the robot would speak only from 12 to 12 throws), we analysed the data considering the different turns in the game. Results showed that there was a significant interaction happening between the conditions and the cheating behaviour in the turns,

suggesting that cheating behaviour differed across the turns and conditions. Our results showed that when participants were alone, cheating increased from the beginning to the end of the game. However, when participants played with the situationally aware robot cheating was decreasing. With the non-situationally aware robot the same was not observed, with a significant difference in cheating behaviour in comparison to the situationally aware condition, for the last game turn. Since the only difference between both robots was being aware or not of the participant's behaviours and subsequent reactions to it, it seems that the situationally aware robot behaviour was responsible for affecting cheating, decreasing it across the turns of the game (the more they heard the robot interventions). On the contrary, with the non-situationally aware robot it seems participants showed some refrain until they understood the robot capabilities, and then they took advantage increasing cheating (similar to the results seen in Study 1 with the limited robot disinhibiting cheating behaviour). There are at least two possible explanations for the effect of the situationally aware robot. On one hand, the robot's interventions might be increasing an awareness of the participant's social image, and so making them feel bad in the "eyes of the robot". By feeling watched, participants could trigger reputation concerns, becoming afraid of being negatively judged. In order to decrease that effect, people start adopting more honest actions. Participants reported greater levels of co-presence with the aware robot and acknowledged more monitoring capabilities in it, but there were no differences between both robots, on feeling watched. If there was a good reliability with the public self-awareness scale, this dimension could elucidate us if participants were in fact more focused and worried on how they were being perceived, but we could not find a reliable consistency for this dimension. On the other hand, the robot's interventions might oblige the participant to update its self-concept by bringing to awareness the true value of his/her actions, that they are cheating. And like the theories of Bounded Ethicality (Chugh et al., 2005; a revised version of the theory- Chugh & Kern, 2016) or Self-Concept Maintenance (Mazar et al., 2008), by making people aware of their actions (increasing the self-threat) obliges people to update their self-concept and be more honest, in order to maintain our default honest self. In our results it was not clear which of these explanations was guiding the effect of the situationally aware robot, still, the robot interventions influenced cheating behaviour. Reinforcing, that if we need to use a more "social" robot, that is able to verbally interact with a person, it seems it is better if it shows situation awareness of the participant's behaviour, in order to be effective in decreasing dishonesty.

Having shown how a more simpler robot (through gaze, in Study 1) and a more complex and social one can stimulate more honesty (through situation awareness, in Study 3), we were curious to test if other kinds of behaviours could enhance the situation awareness effect found. Acknowledging that the robot's capabilities can affect the type of interaction it is developed

with the human, affecting it's quality (e.g., Niculescu et al., 2013) and even engagement for continuous repeated interactions (e.g., Leite et al., 2014), we decided to test if a more friendly and supportive robot (that primed participants for their relational self-concept) could influence their dishonesty. A study by Cojuharenco et al. (2012) showed that by priming participants for their relational self-concept, through reading a text where "we" was used multiple times, was enough to decrease cheating behaviour. Following this, we ran our Study 4, where we started by doing a pilot where we tested a relational robot that was always stimulating a team view and always referring to the participant through the use of "we", in comparison to a neutral robot that always referred to the participant through the use of "you". Our expectation was that the relational robot could prime participants for their relational self-concept and consequently influence their cheating behaviour, in comparison to the neutral robot. But since we already knew that the situationally aware behaviour was effective in decreasing cheating, we decided to keep this behaviour as baseline in both robots (reproducing the effect seen in the previous study) and seeing if the priming could enhance this effect or not. Our pilot study using a within-subjects design, showed differences between both robots, suggesting that our robots were being perceived as it was intended. With this, we advanced for our main study, participants would first play a collaborative Mastermind game with either the relational (that primed them) or the neutral robot, and next they would play the die task to ascertain cheating behaviour with the situationally aware robot. Since the activation of the relational self-concept cannot be objectively measured, we could only see the effects of our manipulation on cheating behaviour. Results showed that cheating happened to the same extent in both conditions, suggesting that the priming for the relational self-concept was not influencing cheating behaviour. Curiously, cheating levels were similar to the situationally aware condition in Study 3, suggesting that what was contributing to those levels of cheating was the situationally aware behaviours and not the priming. These results seem to suggest that the priming manipulation for the relational self-concept may have been too subtle to influence cheating behaviour. We wonder for future studies, if in a continuous interaction with a robot, if this priming could have a different effect on dishonesty. Yet, for now, it seems that it is still more important for a robot to be situationally aware than to be friendly and supportive.

Our experimental studies were conducted both in Portugal and in Sweden, even though we do not use culture as a factor in our project due to the different manipulations that were ran in each country it is still interesting to notice that a common manipulation of being alone in the room while doing the same tempting task, shows a higher success probability in Sweden (M=.74; SD=.21) than in Portugal (M=.59; SD=.120). Still, this could also have been influenced by the fact that the reward in Portugal was around 5.8$ USD and in Sweden the reward of two

movie tickets had a monetary incentive of around 26.80$ USD which could have given a greater incentive.

Overall, with these set of four studies we were able to observe that with simpler robots it is better for the robot to not show the range of its capabilities and just show gaze behaviour. For more complex interactions where a more “social” robot might be needed, it seems that for short interactions it is more important for the robot to show situation awareness of the participants behaviours. Being friendly or supportive does not seem to add advantages to promote more honesty, at least for shorter interactions.

Cheating behaviour and Honesty-Humility trait of personality

Studies suggest that the Honesty-Humility trait of personality can predict cheating behaviour, showing a negative correlation with it (e.g., Hilbig & Zettler, 2015; Kleinlogel et al., 2018; Pfattheicher et al., 2019), with a medium to large effect for this association (Heck et al., 2018). In our Study 2, we found a significant relationship between the Honesty-Humility dimension and cheating. We saw that this trait predicted cheating behaviour, and especially the sub-domain of Fairness was the one that better predicted cheating. A sub-domain which evaluates the tendency to avoid fraud and corruption (Ashton et al., 2014). In our Study 1, we only found a significant correlation between this trait and cheating in the more “social” robot (the robot that gave the instructions for the task), suggesting that for this condition participants with higher scores in this trait showed less cheating (and vice-versa). This leaves us with the supposition if this robot also elicited more truthful and relaxed answers to the personality inventory than in the other conditions. Unfortunately, in our Study 3, we could not find any significant association between the Honesty-Humility trait and cheating behaviour. Possibly the difficulty in finding an association for Study 1 and 3 could be due to a small sample size.

People’s perceptions of dishonesty with robots

Aside from these studies we were also interested in exploring people’s perceptions of dishonesty towards robots, in the literature there was nothing done exploring people’s perceptions. So, with Study 5 we tried to explore if people considered wrong to be dishonest with a robot. Since Study 1 showed us that a more limited robot was taken advantage of, we wondered if the level of autonomy that the robot presented could guide people’s perceptions

of how dishonest it would be, to act dishonestly towards a robot. For this, we created a set of scenarios and we allocated participants to different conditions. Participants would read scenarios where someone would act dishonestly (by stealing or lying) in the presence of a human, an autonomous robot (a robot that did not need human assistance to perform its tasks) or a non-autonomous robot (a robot that needed human assistance, e.g., someone checking its performance or being tele-operated). And participants were asked to evaluate how dishonest towards the agent, the actions portrayed in the scenarios were. They also reported the level of autonomy the robot presented (as a manipulation check for the robot conditions). Results showed that overall, people reported for all the scenarios, as being dishonest the actions done in them. But some specific results were seen for some of the scenarios, it seems in the "University scenario" participants evaluated as more wrong to be dishonest with the human instead of both robots. In the "Finance scenario" it seems it was considered more wrong to cheat the autonomous robot than the human, which was curious. For the "Fire department scenario" it was considered more wrong to cheat the human than the non-autonomous robot. No differences in the "Police scenario" or "Hospital scenario" were found. These differences between the scenarios and the agent in it seem to suggest that different contexts are perceived differently by people, in terms of its dishonesty. This goes in line with a recent study suggesting that people's dishonesty varies according to domain of life (Garcia-Rada et al., 2018). The "Finance scenario" surprising result of showing that it is more dishonest to lie to the autonomous robot than the human, might reflect the current state of the world (in terms of human corruption and dishonesty) and/or peculiar ideas that people might have in terms of paying taxes. The scenarios that showed that it was more dishonest towards the human might reflect on people's perceptions of the robot's capabilities which follows the subsequent results. When exploring people's justifications of being dishonest with a robot and the low level of guilt that they attribute towards a robot. It seems people justify dishonesty because of lack of capabilities in the robot, lack of presence, and a human tendency for dishonesty. These perceptions seem to reflect what is at this moment still missing in robots (at least the ones that people might have interacted with in the meantime), and show that there will need to be a time of adjustment for people to get to know robots and possibly create collaborative interactions.

Finally, by taking into account all these results we still tried to perform one last study, our Study 6, where we manipulated a caring characteristic in a robot (if it showed caring behaviours towards others) or not, to observe if this would affect how people evaluated their attitudes towards being dishonest. This was done not only considering how they thought others would behave, but also how they would behave themselves. Again, we created another set of scenarios where we manipulated the presence of no one when the dishonest act was being done, a human, a caring robot or a neutral robot. Our manipulation of the robot seemed to be

correctly perceived and participants considered themselves mostly honest, which is in accordance with studies showing that people like to have a favourable self-concept of themselves and to be perceived as moral (e.g., Batson et al., 1997, 2006; Fischbacher & Föllmi-Heusi, 2013; Mazar et al., 2008). When participants were evaluating what other people in general would do, the scores seemed to tend to the right side of the scale (to predict that people would be dishonest) but there was no difference between the conditions and there was an effect of social desirability on the scores given. When evaluating what people themselves would do in those scenarios, we did not find any differences in the conditions, and the evaluations seemed to tend to a neutral position (i.e. to the mean of the scale). We also saw that social desirability predicted the scores in the scenarios. Overall, it seems participants think others will be dishonest in general and when evaluating for themselves, participants report a more neutral stance (which goes in line with the effect of social desirability found in the results, possibly due to an underreport of socially undesirable activities). It could also be that due to the sensitivity of the topic, our scenarios where not being interpreted by the type of agent in it, but only by the act itself, allowing for more social desirability responses to emerge. Of interest, we were able to see that especially, people that have negative attitudes towards interacting with robots also tend to evaluate others as behaving more dishonestly and themselves towards robots. Suggesting how there needs to be a period of adaptation and getting to know robots better, before they are integrated in our society.

Limitations

Our studies have some limitations. For our samples of observable cheating behaviour with the physical robots, we only used university students, so we know these results cannot be generalized to the whole population, the same happened with Study 5 (only Study 2 and 6 collected data from people from the general public, but still limited to users of the Mechanical Turk platform). We used two different robots in our studies, with different embodiments, which can be a limitation to the generalization of our results. It was due to availability reasons that we used different robots, but when looking at the most similar conditions between the two robots, which was when the robot was more limited and not aware of the participant's behaviour, we see similar cheating behaviours between the more limited robot in Study 1 and the non-aware robot in Study 3. Participant's significantly cheated more than chance in the presence of either a robotic head (Study 1) or a full-body robot (Study 3) when they presented non-awareness of participant's behaviour. A recent analysis of the literature on robot presence also suggests that the physical presence, instead of physical embodiment, is what

characterizes people's responses to social robots (Li, 2015). Due to this, we believe that the robot embodiment did not influence our results, but future studies are needed to clarify this. Age is also a factor that should be considered when studying cheating behaviour. We only controlled the effect of age on cheating in our Study 2 and 6, since these studies were the ones with a greater age variability. Still, it must be acknowledged as a limitation in not having considered it for the remaining studies. Another aspect that we must acknowledge as a limitation of our studies, is the fact that religion might influence cheating, even though results in the literature are mixed in that respect. On the first studies we ran in this project when we were trying to test different tasks (that are not included in this thesis), we included religion as one of the variables being controlled, but since almost none of the participants reported being religious we dropped this variable from our studies. Still, it can be seen as a limitation, because we ran other studies in Portugal and in Sweden, and we did not take this variable into consideration, so we acknowledge this. Finally, another limitation of our studies is that we did not control for the effect of moral identity. Studies suggest that people that express a high moral identity tend to be more ethical (e.g., Aquino & Reed II, 2002), with a study showing that moral identity was able to predict cheating behaviour (Gino et al., 2011), showing that this variable could have been important to control in our studies. Future studies should consider this aspect.

Concluding remarks and future studies

Robots are being designed and studied to be integrated in a variety of contexts to serve as a support and to work alongside humans. So, it becomes relevant to know how people will behave around them. Knowing that people tend to misbehave if they are tempted for it, it is important to understand if they would also do it in the presence of a robot. Starting from a simple task that tempts participants to cheat, in this thesis, we tried to test and observe how people would behave regarding dishonesty, varying the behaviour that the robot in their presence presented. We saw that either the robot does not show the extent of its capabilities or if it does, it really needs to show that it can catch cheating behaviour in order to be effective in promoting more honesty. We saw that even though people cheated in the presence of a robot when they felt at ease, conceptually they considered it wrong to be dishonest towards a robot (independently of its level of autonomy), as with a human. And they justify that people in general might be dishonest with robots in the future due to its lack of capabilities (cognitive and emotional), its lack of presence (not being taken seriously) and a human tendency for dishonesty. These results go in line with the results from our experimental studies, with

situation awareness influencing cheating behaviour. We also observed effects of social desirability when exploring people's perceptions of what others would do in a dishonest scenario or what they, themselves, would do. It seems that irrespective of the robot characteristics (if it shows more caring characteristics or not), being in the presence of a human or alone, people seem to report that others would probably be dishonesty, independently of the scenario presented. But when asked in the first person, people seem to adopt a more neutral posture. Interestingly, the negative attitudes that people might have towards robots seem to predict how dishonest they think others will be, or themselves, in scenarios with robots.

We end this thesis feeling that more research is needed to better understand all the factors and mechanisms that come into play in dishonesty for human-robot interactions. In terms of future studies, different persuasion strategies could be tested (e.g., social power, likeability, credibility, etc.) and see if their embodiment in a robot could also influence dishonesty. Furthermore, it would also be interesting to explore repeated interactions with a robot and observe if for example, the relational self-concept manipulation could work in this context, by creating a greater proximity with the robot over time and seeing if it would affect dishonest behaviour in its presence.

In sum, it seems that for more complex interactions, it is better for a robot to show awareness of people's behaviours, but it is not clear in which way the situation awareness behaviour in the robot influenced cheating behaviour. Still, we think the Bounded Ethicality theory (Chugh et al., 2005; a revised version of the theory- Chugh & Kern, 2016) and Self-concept Maintenance theory (Mazar et al., 2008) clarify how increasing awareness of one's own acts can increase self-threat, and consequently activate self-protection strategies that motivate people to change their behaviour to be more ethical. In the absence of this awareness, the self-concept is not updated and we resolve any dissonance with secondary mechanisms, like justifying that it was the six on the top of the die that we had previously guessed, and here we go to guess (or choose) the next number.

Chapter 12- Conclusion

In conclusion it seems we need to be careful when preparing robots to work alongside with humans, we need to consider a variety of factors, including human dishonesty. If we integrate robots in more simpler tasks where misbehaving might happen but there is no need to have a more social robot, it seems that it is better if the robot does not show the extent of its capabilities. If the robot needs to be more social and verbally interact with people, than it needs to be able to catch cheating behaviour or people will probably take advantage of it. Yet, considering that very few people have actually interacted with a robot nowadays (at least the ones with a more anthropomorphic embodiment) it will be interesting to see in the future, how people's perceptions about robots will evolve and consequently their interactions with them. Maybe, with certain behaviours implemented, they will be able to promote more honesty from people.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica, 87*(4), 1115-1153. https://doi.org/10.3982/ECTA14673

Adalgeirsson, S. O., & Breazeal, C. (2010). MeBot: A robotic platform for socially embodied telepresence. In *Proceedings of 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 15-22. IEEE. https://doi.org/10.1109/HRI.2010.5453272

Agrawal, S., & Williams, M. A. (2017). Robot authority and human obedience: A study of human behaviour using a robot security guard. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-robot Interaction* (HRI), 57-58. https://doi.org/10.1145/3029798.3038387

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology, 20*(1), 1-48. https://doi.org/10.1080/10463280802613866

Aoki, K., Akai, K., & Onoshiro, K. (2010). Deception and confession: Experimental evidence from a deception game in japan (Discussion Paper No. 786). Osaka University, The Institute of Social and Economic Research. http://hdl.handle.net/10419/92748

Aquino, K., & Reed II, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology, 83*(6), 1423-1440. https://doi.org/10.1037//0022-3514.83.6.1423

Arbel, Y., Bar-El, R., Siniver, E., & Tobol, Y. (2014). Roll a die and tell a lie–What affects honesty?. *Journal of Economic Behavior & Organization, 107*, 153-172. https://doi.org/10.1016/j.jebo.2014.08.009

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology, 63*(4), 596-612. https://doi.org/10.1037/0022-3514.63.4.596

Arora, N. K. (2003). Interacting with cancer patients: the significance of physicians' communication behavior. *Social Science & Medicine, 57*(5), 791-806. https://doi.org/10.1016/S0277-9536(02)00449-5

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*(2), 150-166. https://doi.org/10.1177/1088868306294907

Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*(4), 340-345. https://doi.org/10.1080/00223890902935878

Ashton, M. C., Lee, K., & De Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality factors: A review of research and theory. *Personality and Social Psychology Review, 18*(2), 139-152. https://doi.org/10.1177/1088868314523838

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics, 3*(1), 41-52. https://doi.org/10.1007/s12369-010-0082-7

Bartneck, C. (2002). eMuu-an embodied emotional character for the ambient intelligent home [Doctoral Dissertation, Technische Universiteit Eindhoven]. https://doi.org/10.6100/IR559664

Bartneck, C., & Forlizzi, J. (2004). A design-centred framework for social human-robot interaction. In *13th IEEE international workshop on robot and human interactive*

communication (RO-MAN), pp. 591-594. IEEE. https://doi.org/10.1109/ROMAN.2004.1374827

Bartneck, C., Van Der Hoek, M., Mubin, O., & Al Mahmud, A. (2007). "Daisy, daisy, give me your answer do!" switching off a robot. In *2nd ACM/IEEE International Conference on Human-Robot Interaction* (HRI), 217-222. IEEE. https://doi.org/10.1145/1228716.1228746

Bassarak, C., Leib, M., Mischkowski, D., Strang, S., Glöckner, A., & Shalvi, S. (2017). What provides justification for cheating—Producing or observing counterfactuals?. *Journal of Behavioral Decision Making, 30*(4), 964-975. https://doi.org/10.1002/bdm.2013

Batson, C. D., Collins, E., & Powell, A. A. (2006). Doing business after the fall: The virtue of moral hypocrisy. *Journal of Business Ethics, 66*(4), 321-335. https://doi.org/10.1007/s10551-006-0011-8

Batson, C. D., & Collins, E. C. (2011). Moral hypocrisy: A self-enhancement/self-protection motive in the moral domain. In M. D. Alicke & C. Sedikides (Eds.), *Handbook of self-enhancement and self-protection* (pp. 92–111). The Guilford Press.

Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., & Wilson, A. D. (1997). In a very different voice: unmasking moral hypocrisy. *Journal of Personality and Social Psychology, 72*(6), 1335-1348. https://doi.org/10.1037/0022-3514.72.6.1335

Baxter, P., Ashurst, E., Read, R., Kennedy, J., & Belpaeme, T. (2017). Robot education peers in a situated primary school study: Personalisation promotes child learning. *PloS one, 12*(5), e0178126. https://doi.org/10.1371/journal.pone.0178126

Becker, G. S. (1968). Crime and punishment: An economic approach. In Fielding N.G., Clarke A., Witt R. (eds) *The Economic Dimensions of Crime* (pp. 13-68). Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-62853-7_2

Békir, I., Harbi, S. E., Grolleau, G., Mzoughi, N., & Sutan, A. (2016). The impact of monitoring and sanctions on cheating: experimental evidence from Tunisia. *Managerial and Decision Economics, 37*(7), 461-473. https://doi.org/10.1002/mde.2731

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science Robotics*, *3*(21). https://doi.org/10.1126/scirobotics.aat5954

Bernardi, R. A., Delorey, E. L., LaCross, C. C., & Waite, R. A. (2003). Evidence of social desirability response bias in ethics research: An international study. *Journal of Applied Business Research (JABR), 19*(3), 41-52. https://doi.org/10.19030/jabr.v19i3.2171

Bersoff, D. M. (1999). Why good people sometimes do bad things: Motivated reasoning and unethical behavior. *Personality and Social Psychology Bulletin, 25*(1), 28-39. https://doi.org/10.1177/0146167299025001003

Bickmore, T. W. (2003). Relational agents: Effecting change through human-computer relationships [Doctoral dissertation, Massachusetts Institute of Technology]. http://hdl.handle.net/1721.1/36109

Biocca, F., & Harms, C. (2003). Guide to the networked minds social presence inventory v. 1.2 [Unpublished manuscript]. Department of Telecommunication, Michigan State University. http://cogprints.org/6743/

Bloodgood, J. M., Turnley, W. H., & Mudrack, P. (2008). The influence of ethics instruction, religiosity, and intelligence on cheating behavior. *Journal of Business Ethics, 82*(3), 557-571. https://doi.org/10.1007/s10551-007-9576-0

Böckler, A., van der Wel, R. P., & Welsh, T. N. (2014). Catching eyes: Effects of social and nonsocial cues on attention capture. *Psychological Science, 25*(3), 720-727. https://doi.org/10.1177/0956797613516147

Brewer, M. B., & Gardner, W. (1996). Who is this" We"? Levels of collective identity and self representations. *Journal of Personality and Social Psychology, 71*(1), 83-93. https://doi.org/10.1037/0022-3514.71.1.83

Brščić, D., Kidokoro, H., Suehiro, Y., & Kanda, T. (2015). Escaping from children's abuse of social robots. In *Proceedings of the tenth annual ACM International Conference on Human-robot Interaction* (HRI), pp. 59-66. https://doi.org/10.1145/2696454.2696468

Bryan, C. J., Adams, G. S., & Monin, B. (2013). When cheating would make you a cheater: implicating the self prevents unethical behavior. *Journal of Experimental Psychology: General, 142*(4), 1001-1005. https://doi.org/10.1037/a0030655

Cacace, J., Finzi, A., Lippiello, V., Furci, M., Mimmo, N., & Marconi, L. (2016). A control architecture for multiple drones operated via multimodal interaction in search & rescue mission. In *Proceedings of the IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, 233-239. IEEE. https://doi.org/10.1109/SSRR.2016.7784304

Carpinella, C.M., Wyman, A.B., Perez, M.A., & Stroessner, S. J. (2017). The robotic social attributes scale (rosas): Development and validation. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction* (HRI), 254-262. ACM. https://doi.org/10.1145/2909824.3020208

Childs, J. (2012). Gender differences in lying. *Economics Letters, 114*(2), 147-149. https://doi.org/10.1016/j.econlet.2011.10.006

Childs, J. (2013). Personal characteristics and lying: An experimental investigation. *Economics Letters, 121*(3), 425-427. https://doi.org/10.1016/j.econlet.2013.09.005

Chudzicka-Czupała, A., Lupina-Wegener, A., Borter, S., & Hapon, N. (2013). Students' attitude toward cheating in Switzerland, Ukraine and Poland. *New Educational Review, 32*(2), 66-76. https://tner.polsl.pl/dok/volumes/tner_2_2013.pdf

Chugh, D., Bazerman, M. H., & Banaji, M. R. (2005). Bounded ethicality as a psychological barrier to recognizing conflicts of interest. In D. A. Moore, D. M. Cain, G. Loewenstein, & M. H. Bazerman (Eds.), *Conflicts of interest: Challenges and solutions in business, law, medicine, and public policy* (pp 74-95). Cambridge University Press.

Chugh, D., & Kern, M. C. (2016). A dynamic and cyclical model of bounded ethicality. *Research in Organizational Behavior, 36*, 85-100. https://doi.org/10.1016/j.riob.2016.07.002

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in Experimental Social Psychology, 24*, 201-234. https://doi.org/10.1016/S0065-2601(08)60330-5

Clot, S., Grolleau, G., & Ibanez, L. (2014). Smug alert! Exploring self-licensing behavior in a cheating game. *Economics Letters, 123*(2), 191-194. https://doi.org/10.1016/j.econlet.2014.01.039

Cojuharenco, I., Shteynberg, G., Gelfand, M., & Schminke, M. (2012). Self-construal and unethical behavior. *Journal of Business Ethics, 109*(4), 447-461. https://doi.org/10.1007/s10551-011-1139-8

Conrads, J., Ellenberger, M., Irlenbusch, B., Ohms, E. N., Rilke, R. M., & Walkowitz, G. (2017). Team goal incentives and individual lying behavior. Vallendar, Germany: WHU-Otto Beisheim School of Management. https://d-nb.info/1135786968/34

Conrads, J., Irlenbusch, B., Rilke, R. M., & Walkowitz, G. (2013). Lying and team incentives. *Journal of Economic Psychology, 34*, 1-7. https://doi.org/10.1016/j.joep.2012.10.011

Conrads, J., & Lotz, S. (2015). The effect of communication channels on dishonest behavior. *Journal of Behavioral and Experimental Economics, 58*, 88-93. https://doi.org/10.1016/j.socec.2015.06.006

Covey, M. K., Saladin, S., & Killen, P. J. (1989). Self-monitoring, surveillance, and incentive effects on cheating. *The Journal of Social Psychology, 129*(5), 673-679. https://doi.org/10.1080/00224545.1989.9713784

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic literature, 47*(2), 448-74. https://doi.org/10.1257/jel.47.2.448

Cross, S. E., Bacon, P. L., & Morris, M. L. (2000). The relational-interdependent self-construal and relationships. *Journal of Personality and Social Psychology, 78*(4), 791-808. https://doi.org/10.1037/0022-3514.78.4.791

Dalton, D., & Ortegren, M. (2011). Gender differences in ethics research: The importance of controlling for the social desirability response bias. *Journal of Business Ethics, 103*(1), 73-93. https://doi.org/10.1007/s10551-011-0843-8

De Carolis, B. N., D'Errico, F., & Macchiarulo, N. (2019). 'Keep the user in mind!' Persuasive effects of social robot as personalized nutritional coach. In *the First Symposium on Psychology-Based Technologies* (PsychoBit). http://ceur-ws.org/Vol-2524/paper10.pdf

Dear, K., Dutton, K., & Fox, E. (2019). Do 'watching eyes' influence antisocial behavior? A systematic review & meta-analysis. *Evolution and Human Behavior, 40*(3), 269-280. https://doi.org/10.1016/j.evolhumbehav.2019.01.006

Diener, E., & Wallbom, M. (1976). Effects of self-awareness on antinormative behavior. *Journal of Research in Personality, 10*(1), 107-111. https://doi.org/10.1016/0092-6566(76)90088-X

Dreber, A., & Johannesson, M. (2008). Gender differences in deception. *Economics Letters, 99*(1), 197-199. https://doi.org/10.1016/j.econlet.2007.06.027

Duval, S., & Wicklund, R. A. (1972). *A theory of objective self awareness*. Academic Press.

Edvardsson, D., Watt, E., & Pearce, F. (2017). Patient experiences of caring and person-centredness are associated with perceived nursing care quality. *Journal of Advanced Nursing, 73*(1), 217-227. https://doi.org/10.1111/jan.13105

Ernest-Jones, M., Nettle, D., & Bateson, M. (2011). Effects of eye images on everyday cooperative behavior: a field experiment. *Evolution and Human Behavior, 32*(3), 172-178. https://doi.org/10.1016/j.evolhumbehav.2010.10.006

Ezquerra, L., Kolev, G. I., & Rodriguez-Lara, I. (2018). Gender differences in cheating: Loss vs. gain framing. *Economics Letters, 163*, 46-49. https://doi.org/10.1016/j.econlet.2017.11.016

Fasola, J., & Matarić, M. J. (2013). A socially assistive robot exercise coach for the elderly. *Journal of Human-Robot Interaction*, *2*(2), 3-32. https://doi.org/10.5898/JHRI.2.2.Fasola

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association, 11*(3), 525-547. https://doi.org/10.1111/jeea.12014

Fischinger, D., Einramhof, P., Papoutsakis, K., Wohlkinger, W., Mayer, P., Panek, P., Hofmann, S., Koertner, T., Weiss, A., Argyros, A., & Vincze, M. (2016). Hobbit, a care

robot supporting independent living at home: First prototype and lessons learned. *Robotics and Autonomous Systems*, *75*, 60-78. https://doi.org/10.1016/j.robot.2014.09.029

Forlizzi, J., Saensuksopa, T., Salaets, N., Shomin, M., Mericli, T., & Hoffman, G. (2016). Let's be honest: A controlled field study of ethical behavior in the presence of a robot. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication* (RO-MAN), 769-774. IEEE. https://doi.org/10.1109/ROMAN.2016.7745206

Friesen, L., & Gangadharan, L. (2012). Individual level evidence of dishonesty and the gender effect. *Economics Letters, 117*(3), 624-626. https://doi.org/10.1016/j.econlet.2012.08.005

Friesen, L., & Gangadharan, L. (2013). Designing self-reporting regimes to encourage truth telling: An experimental study. *Journal of Economic Behavior & Organization, 94*, 90-102. https://doi.org/10.1016/j.jebo.2013.08.007

Frischen, A., Eastwood, J. D., & Smilek, D. (2008). Visual search for faces with emotional expressions. *Psychological bulletin, 134*(5), 662-676. https://doi.org/10.1037/0033-2909.134.5.662

Fry, M. D., Guivernau, M., Kim, M. S., Newton, M., Gano-Overway, L. A., & Magyar, T. M. (2012). Youth perceptions of a caring climate, emotional regulation, and psychological well-being. *Sport, Exercise, and Performance Psychology, 1*(1), 44-57. https://doi.org/10.1037/a0025454

Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature, 531*(7595), 496-499. https://doi.org/10.1038/nature17160

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication*, *1*(1), 71-86. https://doi.org/10.30658/hmc.1.5

Garcia-Rada, X., Mann, H., Hornuf, L., Sohn, M., Tafurt, J., Iversen Jr, E. S., & Ariely, D. (2018). The Adaptive Liar: An Interactionist Approach of Multiple Dishonesty Domains. CESifo Working Paper No.7215. https://ssrn.com/abstract=3275388

Gardner, W. L., Gabriel, S., & Lee, A. Y. (1999). "I" value freedom, but "we" value relationships: Self-construal priming mirrors cultural differences in judgment. *Psychological Science, 10*(4), 321-326. https://doi.org/10.1111/1467-9280.00162

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin, 145*(1), 1-44. https://doi.org/10.1037/bul0000174

Gino, F., & Ariely, D. (2012). The dark side of creativity: original thinkers can be more dishonest. *Journal of Personality and Social Psychology, 102*(3), 445-459. https://doi.org/10.1037/a0026406

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological science, 20*(3), 393-398. https://doi.org/10.1111/j.1467-9280.2009.02306.x

Gino, F., & Galinsky, A. D. (2012). Vicarious dishonesty: When psychological closeness creates distance from one's moral compass. *Organizational Behavior and Human Decision Processes, 119*(1), 15-26. https://doi.org/10.1016/j.obhdp.2012.03.011

Gino, F., & Margolis, J. D. (2011). Bringing ethics into focus: How regulatory focus and risk preferences influence (un) ethical behavior. *Organizational Behavior and Human Decision Processes, 115*(2), 145-156. https://doi.org/10.1016/j.obhdp.2011.01.006

Gino, F., Norton, M. I., & Ariely, D. (2010). The counterfeit self: The deceptive costs of faking it. *Psychological science, 21*(5), 712-720. https://doi.org/10.1177/0956797610366545

Gino, F., & Pierce, L. (2009). The abundance effect: Unethical behavior in the presence of wealth. *Organizational Behavior and Human Decision Processes, 109*(2), 142-155. https://doi.org/10.1016/j.obhdp.2009.03.003

Gino, F., Schweitzer, M. E., Mead, N. L., & Ariely, D. (2011). Unable to resist temptation: How self-control depletion promotes unethical behavior. *Organizational Behavior and Human Decision Processes, 115*(2), 191-203. https://doi.org/10.1016/j.obhdp.2011.03.001

Goetz, J., & Kiesler, S. (2002). Cooperation with a robotic assistant. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, pp. 578-579. https://doi.org/10.1145/506443.506492

Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication* (RO-MAN), pp. 55-60. https://doi.org/10.1109/ROMAN.2003.1251796

Govern, J. M., & Marsch, L. A. (2001). Development and validation of the situational self-awareness scale. *Consciousness and Cognition, 10*(3), 366-378. https://doi.org/10.1006/ccog.2001.0506

Graf, B., Hans, M., & Schraft, R. D. (2004). Care-O-bot II—Development of a next generation robotic home assistant. *Autonomous robots*, *16*(2), 193-205. https://doi.org/10.1023/B:AURO.0000016865.35796.e9

Grau, A., Indri, M., Bello, L. L., & Sauter, T. (2017). Industrial robotics in factory automation: From the early stage to the Internet of Things. In *Proceedings of the IECON 43rd Annual Conference of the IEEE Industrial Electronics Society*, 6159-6164. IEEE. https://doi.org/10.1109/IECON.2017.8217070

Grimes, P. W. (2004). Dishonesty in academics and business: A cross-cultural evaluation of student attitudes. *Journal of Business Ethics, 49*(3), 273-290. https://doi.org/10.1023/B:BUSI.0000017969.29461.30

Grolleau, G., Kocher, M. G., & Sutan, A. (2016). Cheating and loss aversion: Do people cheat more to avoid a loss?. *Management Science, 62*(12), 3428-3438. https://doi.org/10.1287/mnsc.2015.2313

Gylfason, H. F., Arnardottir, A. A., & Kristinsson, K. (2013). More on gender differences in lying. *Economics Letters, 119*(1), 94-96. https://doi.org/10.1016/j.econlet.2013.01.027

Hashemian, M., Paiva, A., Mascarenhas, S., Santos, P. A., & Prada, R. (2019). The power to persuade: a study of social power in human-robot interaction. In *Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication* (RO-MAN), 1-8. IEEE. https://doi.org/10.1109/RO-MAN46459.2019.8956298

Heck, D. W., Thielmann, I., Moshagen, M., & Hilbig, B. E. (2018). Who lies? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision making, 13*(4), 356-371. http://journal.sjdm.org/18/18322/jdm18322.html

Herman, C. P., Roth, D. A., & Polivy, J. (2003). Effects of the Presence of Others on Food Intake: A Normative Interpretation. *Psychological Bulletin, 129*(6), 873–886. https://doi.org/10.1037/0033-2909.129.6.873

Hilbig, B. E., & Zettler, I. (2015). When the cat's away, some mice will play: A basic trait account of dishonest behavior. *Journal of Research in Personality, 57*, 72-88. https://doi.org/10.1016/j.jrp.2015.04.003

Hochman, G., Glöckner, A., Fiedler, S., & Ayal, S. (2016). “I can see it in your eyes”: Biased Processing and Increased Arousal in Dishonest Responses. *Journal of Behavioral Decision Making, 29*, 322-335. https://doi.org/10.1002/bdm.1932

Hoffman, G., Forlizzi, J., Ayal, S., Steinfeld, A., Antanitis, J., Hochman, G., Hochendoner, E., & Finkenaur, J. (2015). Robot presence and human honesty: Experimental evidence. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction* (HRI), 181-188. IEEE. http://dx.doi.org/10.1145/2696454.2696487

Hoffmann, L., Krämer, N. C., Lam-Chi, A., & Kopp, S. (2009). Media equation revisited: do users show polite reactions towards an embodied agent?. In *Proceedings of the International Workshop on Intelligent Virtual Agents*, 159-165. Springer. https://doi.org/10.1007/978-3-642-04380-2_19

Hood, B. M., Macrae, C. N., Cole-Davies, V., & Dias, M. (2003). Eye remember you: The effects of gaze direction on face recognition in children and adults. *Developmental Science, 6*(1), 67-71. https://doi.org/10.1111/1467-7687.00256

Houser, D., Vetter, S., & Winter, J. (2012). Fairness and cheating. *European Economic Review, 56*(8), 1645-1655. https://doi.org/10.1016/j.euroecorev.2012.08.001

Jiang, T. (2012). The mind game: Invisible cheating and inferable intentions. (Discussion Paper, No. 309). Katholieke Universiteit Leuven, LICOS Centre for Institutions and Economic Performance. http://dx.doi.org/10.2139/ssrn.2051476

Kachouie, R., Sedighadeli, S., Khosla, R., & Chu, M. T. (2014). Socially assistive robots in elderly care: a mixed-method systematic literature review. *International Journal of Human-Computer Interaction*, *30*(5), 369-393. https://doi.org/10.1080/10447318.2013.873278

Kang, S., Lee, W., Kim, M., & Shin, K. (2005). ROBHAZ-rescue: rough-terrain negotiable teleoperated mobile robot for rescue mission. In *Proceedings of the IEEE International Safety, Security and Rescue Rototics, Workshop,* 105-110. IEEE. https://doi.org/10.1109/SSRR.2005.1501248

Kędzierski, J., Muszyński, R., Zoll, C., Oleksy, A., & Frontkiewicz, M. (2013). EMYS—emotive head of a social robot. *International Journal of Social Robotics, 5*(2), 237-249. https://doi.org/10.1007/s12369-013-0183-1

Keijsers, M., & Bartneck, C. (2018). Mindless robots get bullied. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (HRI), pp. 205-214. https://doi.org/10.1145/3171221.3171266

Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science, 322*(5908), 1681-1685. https://doi.org/10.1126/science.1161405

Kern, M. C., & Chugh, D. (2009). Bounded ethicality: The perils of loss framing. *Psychological Science, 20*(3), 378-384. https://doi.org/10.1111/j.1467-9280.2009.02296.x

Khosla, R., Chu, M. T., Kachouie, R., Yamada, K., & Yamaguchi, T. (2012). Embodying care in Matilda: an affective communication robot for the elderly in Australia. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 295-304. https://doi.org/10.1145/2110363.2110398

Kidd, C. D., & Breazeal, C. (2004). Effect of a robot on user perceptions. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS), pp. 3559-3564. IEEE. https://doi.org/10.1109/IROS.2004.1389967

Kidd, C. D., & Breazeal, C. (2008). Robots at home: Understanding long-term human-robot interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3230-3235. IEEE. https://doi.org/10.1109/IROS.2008.4651113

Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition, 26*(2), 169-181. https://doi.org/10.1521/soco.2008.26.2.169

Kleinlogel, E. P., Dietz, J., & Antonakis, J. (2018). Lucky, competent, or just a cheat? Interactive effects of honesty-humility and moral cues on cheating behavior. *Personality and Social Psychology Bulletin, 44*(2), 158-172. https://doi.org/10.1177/0146167217733071

Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., & Shalvi, S. (2019). Intuitive honesty versus dishonesty: Meta-analytic evidence. *Perspectives on Psychological Science, 14*(5), 778-796. https://doi.org/10.1177/1745691619851778

Krebs, H. I., Volpe, B. T., Williams, D., Celestino, J., Charles, S. K., Lynch, D., & Hogan, N. (2007). Robot-aided neurorehabilitation: a robot for wrist rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *15*(3), 327-335. https://doi.org/10.1109/TNSRE.2007.903899

Krusemark, E. A., Keith Campbell, W., & Clementz, B. A. (2008). Attributions, deception, and event related potentials: an investigation of the self-serving bias. *Psychophysiology, 45*(4), 511-515. https://doi.org/10.1111/j.1469-8986.2008.00659.x

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28*(2), 75-84. https://doi.org/10.1016/j.evolhumbehav.2006.06.001

Lee, M. K., & Takayama, L. (2011). " Now, I have a body" uses and social norms for mobile remote presence in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 33-42. https://doi.org/10.1145/1978942.1978950

Leite, I., Castellano, G., Pereira, A., Martinho, C., & Paiva, A. (2014). Empathic robots for long-term interaction. *International Journal of Social Robotics, 6*(3), 329-341. https://doi.org/10.1007/s12369-014-0227-1

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies, 77*, 23-37. https://doi.org/10.1016/j.ijhcs.2015.01.001

Litoiu, A., Ullman, D., Kim, J., & Scassellati, B. (2015). Evidence that robots trigger a cheating detector in humans. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 165-172. https://doi.org/10.1145/2696454.2696456

Mann, H., Garcia-Rada, X., Hornuf, L., Tafurt, J., & Ariely, D. (2016). Cut from the same cloth: Similarly dishonest individuals across countries. *Journal of Cross-Cultural Psychology, 47*(6), 858-874. https://doi.org/10.1177/0022022116648211

Martin, A. (2013). Does religion buffer cheating? [Doctoral dissertation, Northern Illinois University]. ProQuest Dissertations Publishing. https://search.proquest.com/docview/1501931618?accountid=38384

Martins, A. (2015). Depressiva persistente [Master dissertation, Universidade de Aveiro]. https://core.ac.uk/download/pdf/32245249.pdf

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research, 45*(6), 633-644. https://doi.org/10.1509/jmkr.45.6.633

Mazar, N., & Zhong, C. B. (2010). Do green products make us better people?. *Psychological Science, 21*(4), 494-498. https://doi.org/10.1177/0956797610363538

Mead, N. L., Baumeister, R. F., Gino, F., Schweitzer, M. E., & Ariely, D. (2009). Too tired to tell the truth: Self-control resource depletion and dishonesty. *Journal of Experimental Social Psychology, 45*(3), 594-597. https://doi.org/10.1016/j.jesp.2009.02.004

Meng, W., Liu, Q., Zhou, Z., Ai, Q., Sheng, B., & Xie, S. S. (2015). Recent development of mechanisms and control strategies for robot-assisted lower limb rehabilitation. *Mechatronics*, *31*, 132-145. https://doi.org/10.1016/j.mechatronics.2015.04.005

Midden, C., & Ham, J. (2012). The illusion of agency: the influence of the agency of an artificial agent on its persuasive power. In *International Conference on Persuasive Technology*, 90-99. Springer. https://doi.org/10.1007/978-3-642-31037-9_8

Muehlheusser, G., Roider, A., & Wallmeier, N. (2015). Gender differences in honesty: Groups versus individuals. *Economics Letters, 128*, 25-29. https://doi.org/10.1016/j.econlet.2014.12.019

Myllyneva, A., & Hietanen, J. K. (2016). The dual nature of eye contact: to see and to be seen. *Social Cognitive and Affective Neuroscience, 11*(7), 1089-1095. https://doi.org/10.1093/scan/nsv075

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81-103. https://doi.org/10.1111/0022-4537.00153

Niculescu, A., van Dijk, B., Nijholt, A., Li, H., & See, S. L. (2013). Making social robots more attractive: the effects of voice pitch, humor and empathy. *International Journal of Social Robotics, 5*(2), 171-191. https://doi.org/10.1007/s12369-012-0171-x

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies, 7*(3), 437-454. https://doi.org/10.1075/is.7.3.14nom

Paradeda, R., Ferreira, M. J., Oliveira, R., Martinho, C., & Paiva, A. (2019). The Role of Assertiveness in a Storytelling Game with Persuasive Robotic Non-Player Characters. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (CHI), pp. 453-465. https://doi.org/10.1145/3311350.3347162

Paradeda, R. B., Martinho, C., & Paiva, A. (2020). Persuasion Strategies Using a Social Robot in an Interactive Storytelling Scenario. In *Proceedings of the 8th International Conference on Human-Agent Interaction* (HAI), pp. 69-77. https://doi.org/10.1145/3406499.3415084

Perdue, C. W., Dovidio, J. F., Gurtman, M. B., & Tyler, R. B. (1990). Us and them: social categorization and the process of intergroup bias. *Journal of Personality and Social Psychology, 59*(3), 475-486. https://doi.org/10.1037/0022-3514.59.3.475

Pfattheicher, S., & Keller, J. (2015). The watching eyes phenomenon: The role of a sense of being seen and public self-awareness. *European Journal of Social Psychology, 45*(5), 560-566. https://doi.org/10.1002/ejsp.2122

Pfattheicher, S., Schindler, S., & Nockur, L. (2019). On the impact of Honesty-Humility and a cue of being watched on cheating behavior. *Journal of Economic Psychology, 71*, 159-174. https://doi.org/10.1016/j.joep.2018.06.004

Pfattheicher, S., Strauch, C., Diefenbacher, S., & Schnuerch, R. (2018). A field study on watching eyes and hand hygiene compliance in a public restroom. *Journal of Applied Social Psychology, 48*(4), 188-194. https://doi.org/10.1111/jasp.12501

Powers, A., Kiesler, S., Fussell, S., & Torrey, C. (2007). Comparing a computer agent with a humanoid robot. In Proceedings of the ACM/IEEE international conference on Human-robot interaction (HRI), 145-152. https://doi.org/10.1145/1228716.1228736

Rest, J. R. (1984). The major components of morality. *Morality, Moral Behavior, and Moral Development*, 24-38.

Rest, J. R. (1994). Background: Theory and Research. In J. R. Rest, & D. Narváez (Eds.), *Moral development in the professions: Psychology and applied ethics* (pp 1-26). Psychology Press.

Riether, N., Hegel, F., Wrede, B., & Horstmann, G. (2012). Social facilitation with social robots?. In *7th ACM/IEEE International Conference on Human-Robot Interaction* (HRI), pp. 41-47. IEEE. https://doi.org/10.1145/2157689.2157697

Robinson, H., MacDonald, B., Kerse, N., & Broadbent, E. (2013). The psychosocial effects of a companion robot: a randomized controlled trial. *Journal of the American Medical Directors Association, 14*(9), 661-667. https://doi.org/10.1016/j.jamda.2013.02.007

Ros, R., Oleari, E., Pozzi, C., Sacchitelli, F., Baranzini, D., Bagherzadhalimi, A., Sanna, A., & Demiris, Y. (2016). A motivational approach to support healthy habits in long-term child–robot interaction. *International Journal of Social Robotics*, *8*(5), 599-617. https://doi.org/10.1007/s12369-016-0356-9

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics, 5*(1), 17-34. https://doi.org/10.1007/s12369-012-0173-8

Rothbaum, F., Weisz, J. R., & Snyder, S. S. (1982). Changing the world and changing the self: A two-process model of perceived control. *Journal of Personality and Social Psychology, 42*(1), 5-37. https://doi.org/10.1037/0022-3514.42.1.5

Ruffle, B. J., & Tobol, Y. (2014). Honest on Mondays: Honesty and the temporal separation between decisions and payoffs. *European Economic Review, 65*, 126-135. https://doi.org/10.1016/j.euroecorev.2013.11.004

Sandoval, E. B., Brandstetter, J., & Bartneck, C. (2016). Can a robot bribe a human? The measurement of the negative side of reciprocity in human robot interaction. In *11th ACM/IEEE International Conference on Human-Robot Interaction* (HRI), 117-124. IEEE. https://doi.org/10.1109/HRI.2016.7451742

Schurr, A., Ritov, I., Kareev, Y., & Avrahami, J. (2012). Is that the answer you had in mind? The effect of perspective on unethical behavior. *Judgment and Decision Making, 7*(6), 679-688. http://journal.sjdm.org/12/12916/jdm12916.html

Shalvi, S., Dana, J., Handgraaf, M. J., & De Dreu, C. K. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. Organizational Behavior and Human Decision Processes, 115(2), 181-190. https://doi.org/10.1016/j.obhdp.2011.02.001

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). Psychological Science, 23(10), 1264-1270. https://doi.org/10.1177/0956797612443835

Shariff, A. F., & Norenzayan, A. (2011). Mean gods make good people: Different views of God predict cheating behavior. *The International Journal for the Psychology of Religion, 21*(2), 85-96. https://doi.org/10.1080/10508619.2011.556990

Shu, L. L., Gino, F., & Bazerman, M. H. (2011). Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and Social Psychology Bulletin, 37*(3), 330-349. https://doi.org/10.1177/0146167211398138

Silvia, P. J., & Duval, T. S. (2001). Objective self-awareness theory: Recent progress and enduring problems. *Personality and Social Psychology Review, 5*(3), 230-241. https://doi.org/10.1207/S15327957PSPR0503_4

Stapel, D. A., & Koomen, W. (2001). I, we, and the effects of others on me: How self-construal level moderates social comparison effects. *Journal of Personality and Social Psychology, 80*(5), 766-781. https://doi.org/10.1037/0022-3514.80.5.766

Steinmetz, J., & Pfattheicher, S. (2017). Beyond social facilitation: A review of the far-reaching effects of social attention. *Social Cognition, 35*(5), 585-599. https://doi.org/10.1521/soco.2017.35.5.585

Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlowe-Crowne social desirability scale. *Journal of Clinical Psychology*, 28(2), 191–193. https://doi.org/10.1002/10974679(197204)28:2<191::AIDJCLP2270280220>3.0.CO;2-G

Teixeira, A. A., & Rocha, M. F. (2006). Academic cheating in Austria, Portugal, Romania and Spain: a comparative analysis. *Research in Comparative and International Education, 1*(3), 198-209. https://doi.org/10.2304/rcie.2006.1.3.198

Teixeira, A. A., & Rocha, M. D. F. (2008). Academic cheating in Spain and Portugal: An empirical explanation. *International Journal of Iberian Studies, 21*(1), 3-22. https://doi.org/10.1386/ijis.21.1.3_1

Teven, J. J. (2007). Teacher caring and classroom behavior: Relationships with student affect and perceptions of teacher competence and trustworthiness. *Communication Quarterly, 55*(4), 433-450. https://doi.org/10.1080/01463370701658077

Teven, J. J., & McCroskey, J. C. (1997). The relationship of perceived teacher caring with student learning and teacher evaluation. *Communication Education, 46*(1), 1-9. https://doi.org/10.1080/03634529709379069

Triandis, H. C. (1989). The self and social behavior in differing cultural contexts. *Psychological review, 96*(3), 506-520. https://doi.org/10.1037/0033-295X.96.3.506

Ullman, D., Leite, L., Phillips, J., Kim-Cohen, J., & Scassellati, B. (2014). Smart human, smarter robot: How cheating affects perceptions of social agency. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2996-3001. https://escholarship.org/uc/item/2jh800n1

Vespa, P. M., Miller, C., Hu, X., Nenov, V., Buxey, F., & Martin, N. A. (2007). Intensive care unit robotic telepresence facilitates rapid physician response to unstable patients and decreased cost in neurointensive care. *Surgical Neurology, 67*(4), 331-337. https://doi.org/10.1016/j.surneu.2006.12.042

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Nerini, F. F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, *11*(1), 1-10. https://doi.org/10.1038/s41467-019-14108-y

Wainer, J., Feil-Seifer, D. J., Shell, D. A., & Mataric, M. J. (2007). Embodiment and human-robot interaction: A task-based perspective. In *the 16th IEEE International Symposium on Robot and Human Interactive Communication* (RO-MAN), pp. 872-877. IEEE. https://doi.org/10.1109/ROMAN.2007.4415207

Welsh, D. T., & Ordóñez, L. D. (2014). Conscience without cognition: The effects of subconscious priming on ethical behavior. *Academy of Management Journal, 57*(3), 723-742. https://doi.org/10.5465/amj.2011.1009

Williams, M., Moss, S., Bradshaw, J., & Mattingley, J. (2005). Look at me, I'm smiling: Visual search for threatening and nonthreatening facial expressions. *Visual Cognition, 12*(1), 29-50. https://doi.org/10.1080/13506280444000193

Wilson, J. Q., & Kelling, G. L. (1982). Broken windows. Atlantic monthly, 249(3), 29-38.

Winkle, K., Lemaignan, S., Caleb-Solly, P., Leonards, U., Turton, A., & Bremner, P. (2019). Effective persuasion strategies for socially assistive robots. In *14th ACM/IEEE International Conference on Human-Robot Interaction* (HRI), pp. 277-285. IEEE. https://doi.org/10.1109/HRI.2019.8673313

Yamada, S., Kanda, T., & Tomita, K. (2020). An Escalating Model of Children's Robot Abuse. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (HRI), pp. 191-199. https://doi.org/10.1145/3319502.3374833

Yang, Q., Wu, X., Zhou, X., Mead, N. L., Vohs, K. D., & Baumeister, R. F. (2013). Diverging effects of clean versus dirty money on attitudes, values, and interpersonal behavior. *Journal of Personality and Social Psychology, 104*(3), 473-489. https://doi.org/10.1037/a0030596

Zhong, C. B., Bohns, V. K., & Gino, F. (2010). Good lamps are the best police: Darkness increases dishonesty and self-interested behavior. *Psychological Science, 21*(3), 311-314. https://doi.org/10.1177/0956797609360754