University of Genoa

Robotics, Brain and Cognitive Sciences Department,
Italian Institute of Technology

A thesis submitted in partial fulfilment of the requirements for the degree of DOCTOR OF PHILOSOPHY

Bringing Human Robot Interaction towards Trust and Social Engineering

Alexander Mois Aroyo

Doctoral Course in Cognitive Robotics, Interaction and Rehabilitation Technologies
Doctoral Program in Bioengineering and Robotics

Supervisors:
Francesco Rea
Alessandra Sciutti
Giulio Sandini

Acknowledgments

I would like to thank...

Бих искал да благодаря...

Me gustaría agradecer...

Vorrei ringraziare...

На моята фамилия, на майка ми и баща ми, за всичката им любов и подкрепа която ми предоставиха през всичките тези години. Брат ми, които винаги ми е помагал, защитавал, и се грижил за мен... и обучавал, по неговият "специален" начин, как да бъда по добър човек. Моите племенници, Адриан и Сара, прекрасни малки вечно засмени душички. A mi cuñada, Miriam, que desde tengo memoria, siempre estuvo cuidándome y dándome largas chapas! Y a toda su familia, por ayudarme cuando lo necesitaba. Обичам ви.

На моите вечни, стари и най-добри приятели, Асен, Кристиан и Иван. За страхотните детски години прекарани и всичките простотии направени. Защото винаги мога да разчитам на вас. Agurtzane, por aguantar todas mis tonterías durante tantos años y por quererme tal y como soy (bixo). Esdras, for making unforgettable so many moments we lived, for pushing each other to the limit so to become perfect, for becoming an important friend. Обичам ви.

Iker Etxebarria, por ser el primero en enseñarme, ayudarme y motivarme a hacer robótica, por haberme dejado hacer sumobots!

Alexander Serebrenik for having faith in me and convincing me to stay a bit longer in the Netherlands. Mark van den Brand, who let me be part of his group and mentored me in diverse topics. Sandro Etalle who allowed me to learn about cyber security and social engineering. Dragan Bosnacki who showed me the importance of networking. Емилия Баракова, която ми се довери, и ми разреши да развия моите знания в посоката към роботика, че ме препоръча и ми помогна да намеря докторантура. Mirjam de Haas for letting me learn and help you, for spending long afternoons working together making robots play cards.

Ana-Maria Sutti, Önder Babur, Dan Zhang, Yanya Dajsuren, Ulyana Tikhonova, Bogdan Vasilesku and Sander de Putter, for unconsciously convincing me to do a Ph.D.

Yuexu Chen, for "a","le","k","s", for being a good friend and always present when needed.

Yaping Luo, the best boss ever!! For teaching me, and having patience, for letting me teach you and Jaweon Oh how to hill start! For still being a dear friend and welcomed me in your new family. Вълчо Димитров, за всичките мастики, менти и баници, защото също стана мои добър приятел, и ме покани в твоята нова фамилия.

Jaweon Oh and your beautiful family; for all the beers and long conversations we had, both in the Netherlands and in South Korea.

Alejandro Betancourt por convencerme a ir a Génova, por dejar quedarme en tu casa, y enseñarme todos los escondrijos de la ciudad!

Cristiana Senna, per avermi dato il benvenuto e aiutato durante i primi mesi a Genova.

Josephus Driessen, for deciding to start the Italian bureaucracy adventures together and desperately trying to show me swing.

Alessia Vignolo sia per, avermi fatto vedere come funziona iCub in una stanza... al buio..., sia per tutti i caffè presi insieme che poi si trasformavano in chiacchere sul futuro. Giulia Vezzani, Valentina Vasco, Phuong Nguyen e Nuno Guedelha per iniziare e finire un importante capitolo delle nostre vite.

Bertrand Higy for starting the Italian language adventure together, and for sharing with me the joy of your family. Laura Morano per averci insegnato con tanta pazienza l'italiano!

Elisa Maiettini, for teaching me the importance of connected magical chickens on trees, for being able to add a new dimension to any topic, for being always there when needed.

Edwin Avila, Jorge Fernandez, Jorhabib Eljaik y Jairo Osorio, por toda vuestra energía y fuerza.

Marco Contardi, Michela Cosentino, Barbara Salis, Dario Prestieri, Amira El Merhie, Melody di Bona, Irene Rosina e Romeo Orsolino per aver fatto diventare piacevoli questi tre anni dolorosi!

Takahiro Deguchi and all your family, for being kind, helpful and sharing your joy both in, Genova and Osaka.

Tutti i membri dal lab: Giulio Sandini, Alessandra Sciutti, Francesco Rea, Alessia Vignolo, Ana Tanevska, Fabio Vanucci, Jonas Gonzalez, Giulia Belgiovine, Valeria Falzarano e Dario Pasquali perché senza di voi, niente di questo sarebbe stato possibile.

To all the participants in my experiments.

Matthew Rueben and Johannes Schmölz for giving me a different private point of view regarding robots!

Cate Jerram for being my mentor at the social engineering school, and Richard Matthews for teaching me all the tricks!

Luca Recla e Alessandro Bruchi per avermi supportato e soprattutto sopportato. Laura Taverna e le sue belle riprese. Anastasia Bruzzone, Cinzia Pellini e Alida Scotti per il supporto amministrativo e i caffè mattutini. Alessia Tonelli e Jessica Podda per essere i miei psycho consulenti. Marco Randazzo per avermi insegnato a giocare con iKart. Julien Jenvrin e Davide Dellepiane per tutte le volte in cui hanno sistemato il nostro bel'iCub Reddy. A Delle, per avermi trascinato nel mondo dell'astronomia.

The research done in Japan could never have been done without the help of Shinobu Kawasaki, operating and running the experiment; Yoko Asakura, for gathering and scheduling all the subjects; and Michiko Deguchi, helping with the translation; and all the rest of the lab members: Tatsukawa Kyouhei, Tora Koyama, Hideyuki Takahashi, Yuichiro Yoshikawa and Hiroshi Ishiguro.

My friend, iCub Reddy, who has always been there giving me joy and pain, excitement and despair during all my Ph.D. years.

Brittany Postnikoff, that without even knowing, brought a bit of sanity in my head (just in the very last moment of the Ph.D. though) - realizing that I am not the only crazy one with the idea of social engineering robots :)

Sci-Hub and Library Genesis, as a lot of the literature review would not have been possible otherwise.

Cate Jerram, Kerstin Dautenhahn and Emilia Barakova, for spending time and effort to read and revise my work, to give me advice and their point of view so to make this thesis better.

Giulia Zanini, for all these years standing by my side and supporting me during the difficult, tough and stressful periods of the Ph.D., for having an immense patience, for being kind and having faith that this Ph.D. will reach a good end.

Francesco Rea, for supporting me, caring about me and teaching me every little aspect about the robot, the Ph.D., and life in general. Alessandra Sciutti, for always being cheerful, happy and energetic, for all the discussions we've got that have been helpful and always work related. Giulio Sandini, for always offering his help, and providing different and excellent points of view. My supervisors, that not only have been outstanding tutors, but also have become friends. For having faith in me and letting me work on a topic, which seemed a bit unusual and controversial, but ended up in a quite interesting idea.

Everyone else that somehow somewhere has influenced my life and way of thinking.

Abstract

Robots started their journey in books and movies; nowadays, they are becoming an important part of our daily lives: from industrial robots, passing through entertainment robots, and reaching social robotics in fields like healthcare or education.

An important aspect of social robotics is the human counterpart, therefore, there is an interaction between the humans and robots. Interactions among humans are often taken for granted as, since children, we learn how to interact with each other. In robotics, this interaction is still very immature, however, critical for a successful incorporation of robots in society. Human robot interaction (HRI) is the domain that works on improving these interactions.

HRI encloses many aspects, and a significant one is trust. *Trust* is the assumption that somebody or something is good and reliable; and it is critical for a developed society. Therefore, in a society where robots can part, the trust they could generate will be essential for cohabitation.

A downside of trust is *overtrusting* an entity; in other words, an insufficient alignment of the projected trust and the expectations of a morally correct behaviour. This effect could negatively influence and damage the interactions between agents. In the case of humans, it is usually exploited by scammers, conmen or social engineers - who take advantage of the people's overtrust in order to manipulate them into performing actions that may not be beneficial for the victims.

This thesis tries to shed light on the development of trust towards robots, how this trust could become overtrust and be exploited by social engineering techniques. More precisely, the following experiments have been carried out: (i) *Treasure Hunt,* in which the robot followed a social engineering framework where it gathered personal information from the participants, improved the trust and rapport with them, and at the end, it exploited that trust manipulating participants into performing a risky action. (ii) *Wicked Professor*, in which a very human-like robot tried to enforce its authority to make participants obey socially inappropriate requests. Most of the participants realized that the requests were morally wrong, but eventually, they succumbed to the robot's

authority while holding the robot as morally responsible. (iii) *Detective iCub,* in which it was evaluated whether the robot could be endowed with the ability to detect when the human partner was lying. Deception detection is an essential skill for social engineers and professionals in the domain of education, healthcare and security. The robot achieved 75% of accuracy in the lie detection. There were also found slight differences in the behaviour exhibited by the participants when interacting with a human or a robot interrogator.

Lastly, this thesis approaches the topic of privacy - a fundamental human value. With the integration of robotics and technology in our society, privacy will be affected in ways we are not used. Robots have sensors able to record and gather all kind of data, and it is possible that this information is transmitted via internet without the knowledge of the user. This is an important aspect to consider since a violation in privacy can heavily impact the trust.

Summarizing, this thesis shows that robots are able to establish and improve trust during an interaction, to take advantage of overtrust and to misuse it by applying different types of social engineering techniques, such as manipulation and authority. Moreover, robots can be enabled to pick up different human cues to detect deception, which can help both, social engineers and professionals in the human sector. Nevertheless, it is of the utmost importance to make roboticists, programmers, entrepreneurs, lawyers, psychologists, and other sectors involved, aware that social robots can be highly beneficial for humans, but they could also be exploited for malicious purposes.

Contents

ACKNOWLEDGMENTS I

ABSTRACT V

INTRODUCTION 1

- Human Robot Interaction 4
- Trust 5
- Overtrust 7
- Social Engineering 9
- Privacy 12
- Motivation 14

EXPERIMENTS 16

- Treasure Hunt 17
- Wicked Professor 37
- Detective iCub 58

DISCUSSION 75

CONCLUSIONS 83

APPENDIX 86

REFERENCES 89

Chapter 1

Introduction

ROBOTS are becoming an important part of our modern society. Step by step they are being incorporated in our daily lives; starting from industrial robots, then entertainment, robots as toys, and slowly becoming companion robots as well - entering in work places, households, hospitals. One of the main reasons of introducing robots in our society is that our life standards could be improved: robots are built to serve a beneficial or helpful purpose; not hostile (in most of the cases).

Nevertheless, in order to be fully beneficial, there is a need of a proper interaction between humans and robots. In human-human interactions (HHI), we voluntary build mutual and flexible relationships that emphasize equality and reciprocity [1] - is an important pillar in a developed society. Knowing people, interacting with them, having confidence among them, and trusting each other allows the society to grow and expand.

In order to introduce intelligent robots in our society, we need to be able to make them interact with humans in a similar and easily understandable way. The domain of human robot interaction (HRI) is in charge of studying the interactions among humans, and trying to understand and improve similar interactive situations between humans and robots. The problem of HRI is relatively new, and became popular during the 20th

century thanks to the famous "Three Laws of Robotics" by Isaac Asimov [2] in his novel "I, Robot"; which also seem to be the very first guidelines of HRI [3] even though later on were stated as not good enough to govern robots' behaviour [4]–[6]. Yet, we are very far away from currently being able to develop such robots and have such relationships. However, it is really important to start understanding how different cultures, different ethics and different entities interact with the present robots. Understanding how to make interactions with robots smooth and effective, could also be relevant also for other domains such as neurobiology, psychology or cognitive sciences; providing a novel perspective to the question of how humans develop, learn and build their social relationships [7].

Recently a topic that become important in HRI is the development of trust in relationships. In HHI, trust is defined as the belief that someone or something is reliable, good, honest and effective; and it is a fundamental component in human affairs [8]. Following the human behaviour, and considering that robots are already being integrated in our society, it is natural to foresee that trust will be important for HRI. Moreover, robots are not simply perceived as tools but also begin to be seen as colleagues [9], [10]; showing even emotional attachment that could trigger positive and negative behaviours [11].

To bring HRI to the next level, there is a need not only for increasing the trust between humans and robots, but also for analysing how this trust will develop over time and over robot's failures - how to predict and prevent erroneous situations in which even a human life could be in danger.

On the other hand, similarly to what happens in human-human interactions, trust could become overtrust and be exploited for negative purposes by criminals, scammers, conmen, and social engineers - who tend to psychologically manipulate people in order to make them perform non-beneficial actions [12]. Unfortunately, this effect is already observed in human-robot interactions where overtrust is defined as an imprecise calibration between the real capabilities of a robot and the person's expectation [13]. For example, people conform (*i.e.* to act in accordance, to comply) to the robot's requests [14], even if those requests may result into property damage [15], they succumb to bribery [16] or follow an erroneous robot in an emergency evacuation scenario [17].

As the number of robots is increasing very fast, it is necessary to better understand trust and overtrust in HRI scenarios. It is essential that robot designers, programmers, lawyers, as well as users, become aware of the potential risks that could arise from interacting with robots in everyday life.

Making people aware that robots are controlled by computers that can be hacked, resulting not only into a security leak but also potentially into harmful machines, could prevent social panic caused by unlawful activities using robots. Such awareness should also push the community to tackle several questions before robots end up in our daily lives, as whether current robots are secure enough to be on the market, or how the data obtained by all the robot sensors should be treated in order to avoid privacy violations.

Analysing how social engineers, scammers or conmen exploit human overtrust to obtain personal benefits, it is natural to foresee that they could use their techniques on robots to anonymously get closer to the victim. The advent of Internet and communication technology gave social engineers an ultimate protection - distance and anonymity [18]. Having robots capable of moving, recording video or sound will bring attackers a huge advantage. Also due to the people's tendency to humanize robots [19], they might not realize that robots have a computer connected to the Internet with cameras and microphones, thus, malicious people can access the robot, control it or used it to record and steal personal information. Moreover, robots open the possibility of the usage of different skills or psychological manipulation techniques, currently available mostly during physical interactions, such as development of trust and rapport, authority exploitation or even deceit detection.

Summarizing, this thesis addresses the following questions: can robots evoke trust as humans do? This could support future robot involvement in different domains such as homes, hospitals, supermarkets or interactions with law officers. Likewise, can robots be attributed similar authority as humans? For instance, this could help understanding whether children will behave the same way adults do when being taught by a robotic teacher or cared by a robotic nurse. Will they be more prone to lie to robots? Last but not least, can social engineers exploit their techniques using robots in order to manipulate people for negative purposes or even violate their privacy?

1.1. Human Robot Interaction

Human robot interaction is a relatively new domain of study. Humans started interacting with robots since the 1940's in a very simple way [20], so HRI was considered for a long time a subset of Human Computer Interaction (HCI) [21] and defined as "design, evaluation and implementation of interactive computing systems for human use" [22].

Since quite some time robots stopped being perceived only as tools but also began to be seen as colleagues [9], [10], therefore the definition of HRI changed to "the study of humans, robots, and the way they influence each other" [20]. HRI is related to HCI but it also differs because robots are starting to be more complex, exhibiting cognition and autonomy, and operating in dynamic environments [20]. Goodrich *et al.* [3] defines HRI as a field that requires communication and interaction between robots and humans, and dedicated to understand, design and evaluate robotic systems.

Human robot interaction can be applied in a variety of different fields and applications: *search and rescue* [23], [24]; *hazard removal* [25]; *entertainment* [26]–[29]; *military and police* [30]–[32]; *space* [33]–[37]. Recently, a lot of researchers are focusing on making the robots more similar to humans - exhibiting human traits [38]–[40], interacting with natural language [41], [42]; and performing human tasks; so to create a new application field called *social robotics.* From [43], the latter can be divided into domains such as *healthcare* [44]–[46], *homecare* [47]–[49] or *education* [50]–[54].

In all the applications of the abovementioned domains, one very important characteristic in human robot interaction is *trust.* Nowadays, trust in HRI is an important topic, with a lot of studies about it: [9], [55]–[57]. Without trust, the interaction in healthcare, homecare or education will not prevail in the long term - this topic is extensively discussed in the next section.

1.2. Trust

Trust is the belief that someone or something is reliable, good, honest and effective; and it is a fundamental component in human affairs [8]. Many business managers, psychologists, sociologists, economists and academic researchers are certain in its importance in human affairs. Bok [58] affirmed that "when trust is destroyed, societies falter and collapse."; Lewis and Weigert [8] stated that trust is "indispensable in social relationships"; while Zucker [59] added that trust is "vital for the maintenance of cooperation in society and necessary as grounds for even the most routine, everyday interactions".

Following these definitions from human-human interactions, in robotics and automation, trust is defined by Lee *et al.* [13] as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" while Hancock *et al.* [60] defines it as "the reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others". This suggests that trust is a fundamental part of beneficial human interaction and it is natural to foresee that it will soon be important for human robot interaction. Robots are already integrated in our society and often they are not simply perceived as tools, but considered as our partners in activities of daily living [9], [10]; showing even emotional attachment, or triggering positive and negative behaviours [11].

Numerous studies in HRI have investigated which factors influence trust. The user's confidence in a system to behave correctly [55], environmental factors or robot characteristics such as performance, can affect the development of trust [57]. Additionally, robot's transparency [61], [62] or transparency in the decision making process of human-like autonomous systems [63], have been shown to influence trust as well. Moreover if the information is constantly streamed in to the human partner, there is an increase of trust [64]. For example, in [65] it has been shown that participants trust more a robot that provides explanations of its acts than one who does not. Also robot efficiency and, in general, system reliability [66], [67] have been deemed as crucial in determining their partners' trust [56] - the more efficient a robot is, the more people will trust it.

Nevertheless, there is not yet a standard definition of trust in the robotics domain, but more of a combination of different factors that seem to influence trust.

Likewise human-human interactions, trust could become overtrust and be exploited by criminals, scammers, conmen, and social engineers for negative purposes [12]. Unfortunately, this effect of overtrust is already present in human-robot interactions. This will be broadly discussed in the next section.

1.3. Overtrust

Overtrust, as stated in [13], is a poor calibration between the person's trust and the system's capabilities; more precisely, overtrust is described as how a system could be inappropriately relied upon, even compromising safety and profitability [68].

This implies that there is a risk that trust can become overtrust and be exploited for negative purposes: recent research has demonstrated that participants might comply and trust a robot despite its requests being awkward or not transparent [15]; or even in case of malfunctioning [17].

For instance, Salem *et al.* [15] conducted an experiment to understand how users will perform against odd requests coming from an erratically behaving robot - Sunflower [69]. It showed that participants comply with awkward orders from a robot even when they could result into information leakage or property damage, and also when the robot openly exhibited faulty behaviour.

A dangerous example of overtrust in robots could happen in emergency evacuation scenarios: Robinette *et al.* [17] demonstrated that participants in a fake fire emergency scenario tended to follow a robot, rather than the emergency signs, even if it showed clear malfunctioning in its navigation. Similar results shed light on how overtrust towards robots could be potentially harmful for humans.

These findings add to the previously mentioned evidence that robots can persuade humans to change their ideas and conform to the robot's suggestions: Gaudiello *et al.* [14] developed an experiment about human conformation (*i.e.* to act in accordance, to comply) towards a humanoid robot - iCub [70]. The experiment measured how many people would withdraw their answers to a set of questions after listening to iCub's reply. Conformation is at the borderline of overtrust: it is a desirable aspect in HRI scenarios in which a robot might be more informed than the human user, but it may also be risky if conformation is done blindly, *e.g.,* overtrust in emergency scenarios - as previously explained.

Another experiment done by Sandoval *et al.* [16] exploited bribery (an illegal act of persuading an entity with money or presents in order to act in their benefit) in the context of HRI; using a NAO robot [71] to bribe a group of participants with money in order to receive a future help from them into doing tedious task.

Therefore, although trust is vital for human interactions, it also can become overtrust; and one potential risk of overtrust is that it can be dangerously exploited by criminals. For instance, in a human-human context, people have fallen prey to social engineering attacks, scams, conmen, etc. As robots will be used and programmed by humans, there is a chance that people will use robots as a tool for committing crimes, creating security leaks and more. Social engineering in the human robot interaction context is a new and crucial domain - explained in more details in the next section.

1.4. Social Engineering

Social Engineering (SE) is a combination of science, psychology and art, and consists in influencing people to take actions that may or may not be beneficial for them [12]. One of its domains is Information Security - in this case SE is a psychological manipulation of people (*targets*) in order to obtain some benefit like personal data, passwords or confidential information [72], taking advantage of the kindness and trust humans have among themselves [73]. Therefore, big companies and industries are spending large quantities of money to ensure their own computer safety. However, in SE, attackers are targeting humans, which are the weakest link in the security chain of a system - being the human error the biggest cause of problems in information security [74]. It is much easier to trick a person to give the password rather than hacking it from the system [73]. Such manipulation techniques can vary from building trust and rapport, misusing authority or even use lie detection skills to adapt to a dynamic and interactive scenario where attacker needs to successfully manipulate the target.

With the introduction of robots at home and in the workplace, there is a risk that trust toward them might be exploited. Social engineers may exploit human-robot interactions in ostensibly safe environments such as work place, home, or during holidays [73], [75]. They might use their techniques through robots, to anonymously get closer to the target, exploiting the rapport of trust developed with the robot during daily interactions. Having a robot that is capable of moving and recording video or sound will bring a huge advantage to social engineering.

Numerous examples exist which demonstrate that robots can constitute a threat to safety and security: the case of a hijacked Hello Barbie [76]; spying teddy bears [77]; hijacked surgery robots [78]; Alpha robot turned into a stabbing machine [79]; or even piggybacking robots [80] show high risks and vulnerabilities in the domain.

Safety has been an important concern in robotics [81], however, to address this and other security issues, it may be insufficient only to build a more secure robot with a stronger protection against cyber attacks. In fact, the human agent is the weakest link in the cyber security chain [73]. Rather, it is of vital importance to understand which factors

influence human trust toward a robot. Moreover, it is necessary to study how trust changes over time in order to be able to predict and prevent the risk of overtrust and trust exploitation in important domains such as healthcare [44]–[46], homecare [47]–[49] or education [50]–[54].

As seen in the abovementioned examples, humans are already using robots as a tool for committing crimes. Studying the possible risks of misusing robots could prevent and warn future users about the risks in the same way as computer users are taught to be careful with viruses, phishing, etc. An early assessment of the possible risks can help new robot developers to build from the beginning a more secure robot with a strong base against attacks rather than trying to fix issues once the product is on the market.

This project of this thesis addresses the topic of trust in HRI and investigates whether people naturally tend to trust robots and which factors influence the level of trust toward a humanoid machine. More specifically, it aims at exploring to what extent a robot could exploit social engineering methods in order to manipulate people to do actions that could result into the extraction of confidential information or performance of inappropriate tasks. So far there has been only one very recent research defining the concept of social engineering robots [82], but only from the theoretical point of view. While in [80], robots are used for piggybacking (*i.e., enter in an unauthorized location*), the authors do not defined it as a social engineering attack, even though it can be considered one.

This thesis presents methods and practical examples on how SE attacks could be performed in the domain of HRI. These examples include the use of a social engineering cycle technique in order to obtain personal information, gain trust and exploit it afterwards - further explained in the *Treasure Hunt* experiment.

Another way to exploit social engineering vulnerabilities is by the use of authority - where the attacking entity pretends to be a figure of a superior authority than the victim so to manipulate them in an easier way- further explained in the *Wicked Professor* experiment.

On the other hand, deception detection is an essential skill for social engineers - being able to detect lies provides flexibility and adaptability in an attacking scenario, even more

if the attacker pretexts to be a figure with an superior authority of the one of the target [75]. Apart from the advantages in social engineering, having robots capable of detecting lies could be of a great benefit in the human robot interactions as well. It is a necessary skill that a robot should have, for example, if in charge of elderly care - did the person take the medication or is lying; or in case of babysitting children - did they finish the homework so they can go playing? Such aspects are further discussed in the *Detective iCub* experiment.

Next section gathers and discusses different privacy issues that are appearing with the usage of robots in daily life - and suggests ways to solve them.

1.5. Privacy

Another very important aspect to consider in the human robot interaction is privacy. Although different cultures perceive privacy in different ways, all of us need it [83]. Since the 19th century researchers have argued about its definition [84]. Moore [85] defines it as "having control over the access to one's information"; Altman [86] as a dynamic boundaries that can adjusted depending on the interactions with different people; while Nissenbaum [87] states that privacy is context-dependent - different rules may apply for information gathering and dissemination in different environments. A slightly more specialized approach, by Austin [88], focuses on the freedom of public surveillance and influence of technology in privacy. Lastly, a different approach to define privacy is through taxonomy covering different parts of privacy [89], and typology based on constitutional protection in different countries [90].

Nevertheless, privacy is a significant requirement for relationships to prosper [91] and for individuals to grow and be free. Moreover, privacy can enable trust as well [92]. As seen in the examples above, robots are capable of violating human privacy: collecting and sharing confidential information, moving through personal spaces, and socially interacting with people [93]. Thus, if robots want to be accepted in society, they should earn people's trust, while not breaking their confidence due to privacy violations.

In the domain of HRI, privacy is yet not much considered. There are different privacy theories in HHI [87], [94] but there is yet no research on how these theories could work in HRI. Rueben *et al.* [89] reviewed privacy literature and acknowledged that privacy has several dimensions: informational, physical, psychological and social. Taking into account the end-user's best interest and knowing that privacy is crucial for healthy relationships, Rueben [95] introduced a new area of research interest called *privacy sensitive robotics.* As robots will impact privacy in different ways, Rueben, Aroyo *et al.* [96] have identified important issues that should be addressed by different specialists such as programmers, roboticists, entrepreneurs, lawyers, psychologist, etc. There are seven research themes identified that should comprise privacy-sensitive robotics research in the near future: (i) data privacy - how to storage and process data; learn and personalize user's preferences ; (ii) manipulation and deception - security, education and social engineering

through robots; (iii) trust - related to privacy, and trust evolution in HRI; (iv) blame and transparency - responsible entities when privacy violations occur; (v) legal issues - regulating robots as persons and robot companies; (vi) domains with special privacy concerns - healthcare and homecare; and (vii) privacy theory -taxonomy of privacy .

The intention of the previously suggested themes is to recommend, per each one of them, a research direction that serves as a roadmap for research in privacy-sensitive robotics. Few researchers have approached individually some of those domains: Syrdal *et al.* [97] studied personal information disclosure; Denning *et al.* [98] demonstrated security vulnerabilities in commercial robots; Lee *et al.* [99] investigated privacy with robots at the work place, while Krupp *et al.* [100] with telepresence robots; Calo wrote a chapter of Robot Ethics [93] and also focused on how drones affect privacy [101]; and Kaminski *et al.* worked on home robots affecting privacy [102] as well as the potential privacy harm caused by using robots [103].

1.6. Motivation

The topics discussed in this thesis may seem not so conventional and even controversial. Overtrust and social engineering are domains which the general public tries to avoid due to their bad reputation, nevertheless, it is important to understand how they may affect humans during interactions with robots, and how we could try to protect ourselves and prevent them from happening. There will never be a full protection from these kinds of attacks, however, the more people know about the risks, the more they can be aware of the possible problems. The motives of researching and bringing human robot interaction towards trust and social engineering are indicated in the next paragraphs:

In human-human interactions we obey special rules in important domains such as military and police [30]–[32], healthcare [44]–[46], homecare [47]–[49] or education [50]–[54]; for example, there is an implicit trust in that context, which generally, make us rely on others for our own sake. Nowadays, robots are being introduced in those domains and there is a crucial need of studying how trust is affected by that integration, *e.g.,* will a person trust a robotic police officer or a robotic doctor?

Apart from trust, another important factor in those domains is the perceived authority and the capability of the robots to manipulate the human agents. Police officers, doctors, nurses, teachers, etc. have certain authority under specific contexts, however, do humans react the same way under the authority of a robot? Can a robot make an elderly person to take the medication even if it makes them dizzy? Could a robot manipulate a child to do the homework before watching TV?

Yet, manipulation is generally used with negative intentions and misuse of trust - as seen in previously mentioned examples of criminals, scammers, conmen, social engineers and even politicians (social engineering as a branch of political science). An important solution to the problem is education: making people aware of the risks and teach them protective techniques. One example of success in education is regarding the *Nigerian Prince* advance-fee fraud from the 419 group of scammers [104] - where the (email) scam consists in the request of an advanced bank transaction fee payment in order to receive a valuable sum of money afterwards (that never arrives). These types of scams became very

popular with the introduction of computers in society - as having a computer and being able to send emails was something uncommon. Nowadays, thanks to the news and education, the success ratio of the scam is very low. A similar way to educate people of the possible risks in robotics may help and prevent future criminal attacks.

Last but not least, big companies should be aware, and governments should legislate these problems. Still, a lot of robotic products that arrive to the market lack of security measurements and proper protections: hijacked Hello Barbies [76]; spying teddy bears [77]; hijacked surgery robots [78]; robots turned into a stabbing machines [79]; or even piggybacking robots [80]. The research proposed by Rueben, Aroyo *et al* [96] could be beneficial for roboticists, entrepreneurs, programmers, lawyers, and others, as it suggests directions toward the solutions to those problems.

These are some of the reasons that motivated the directions taken in the research presented in this thesis. The next sections present different experiments that try to tackle the above-mentioned problems and study the different reactions and behaviours of participants when confronted by such controversial situations.

Chapter 2

Experiments

IN ORDER to investigate the implications of trust and social engineering in human robot interactions, the following three experiments have been performed involving humans and humanoid robots.

The first experiment, *Treasure Hunt* [105], tries to answer the question "Can robots build trust and exploit it for social engineering purposes?", and delves into the trust evolution between participants and a robot, while the robot also tries to exploit participants using a famous social engineering technique.

The second experiment, *Wicked Professor* [106], addresses the question: "Can the authority be transferable to a robot to the extent of performing controversial actions?", and peeks into the negative implications that the attribution of authority to robots could have onto humans.

The last experiment, *Detective iCub* [107], aims at answering the question: "Can a robot learn lie detection features?", by investing the robot iCub with the role of police investigator and making it detect lies while interacting with deceptive subjects.

2.1. Treasure Hunt

The experiment intends to study whether and how trust evolves during a relatively long collaborative interaction between a robot and a human. Moreover, it was structured following a famous social engineering technique in order to understand the possibilities of people vulnerable to this type of attack. Finally, even though there is a lot of literature regarding trusting games and robots [108], [109], none of them measures overtrust through a robot suggesting participants to gamble as done in the following experiment.

2.1.1. Overview

Robots are becoming widespread in society and issues such as information security and overtrust in them are gaining relevance. This research aims at giving an insight into how trust towards robots could be exploited for the purpose of social engineering (SE). Drawing on Mitnick's model, a well-known social engineering framework, an interactive scenario with the humanoid robot iCub was designed to emulate a SE attack. At first iCub attempted to collect the kind of *personal information* usually gathered by social engineers by asking a series of private questions. Then, the robot tried to *develop trust and rapport* with participants by offering reliable clues during a treasure hunt game. At the end of the treasure hunt, the robot tried to *exploit the gained trust* in order to make participants gamble the money they won. The results show that people tend to build rapport with and trust toward the robot, resulting in the disclosure of sensitive information, conformation to its suggestions and gambling.

2.1.2. Introduction

Although trust and overtrust have been investigated in human robot interaction (HRI), there is only one research that defines theoretically how robots could be used for social engineering attacks [82]. This experiment proposes a way to assess whether a robot can gather information from its human partners, build trust and rapport with them and, exploit it to induce them to perform an action.

The experiment draws on the widely used social engineering framework proposed by Kevin Mitnick [73]. According to this model, an SE attack is separated into the following phases: (i) *research the target* to gather as much information as possible; (ii) *develop trust and good rapport* with the target; (iii) *exploit trust* to make the target perform an action; and (iv) *utilize that information* to achieve an objective. These phases can be iterated as many times as necessary to reach the desired goal. As an example taken from [75]: a meat salesperson spots a lady cooking on a barbecue in the yard (i); he talks respectfully and nicely with the lady about cooking and about the quality of the meat he is selling (ii); and he tries to convince her to buy the meat (iii). In this case, the goal has been achieved in the third phase.

In this experiment, the humanoid robot iCub asked a series of questions about participants' private lives ((i) - research). They then played a treasure hunt game, in which participants had to find hidden objects (eggs) in a room to win a monetary prize. The robot offered its help and, when asked, provided reliable hints about the location of the hidden eggs. The treasure hunt was designed to provide an engaging setting where the participants' trust and rapport towards the robot could develop during the interaction ((ii) - develop trust and rapport). Finally, exploiting the trust acquired, the robot suggested participants to gamble the monetary prize they won - doubling it if they could find another egg, and losing everything if not ((iii) - trust exploitation). Similar to the previous example, there is no need for a fourth phase as the goal was already achieved.

This research evaluates whether trust toward robot and compliance to its suggestions is modulated by individual personality traits and experiment's impressions. Moreover, it tries to verify a series of hypotheses about trust toward robots, its evolution during an interaction and its implications for SE. More precisely, that: (H1) *participants who are less prone to social engineering in general, or have an overall higher negative attitude towards robots, would be less open to reveal sensitive information to the robot.* (H2) *All participants would conform to all the robot's suggestions during the game but those less incline to take risks would not comply with the proposal to gamble due to the potential monetary loss.* (H3) *The rapport with the robot after the experiment would improve most for the participants who won the game and doubled their award.*

2.1.3. Methodology

A. Experimental Setup

The robot used in the experiment was the interactive humanoid robot iCub, developed by the Italian Institute of Technology [70]. It was located next to the wall in the middle of the experimental room (triangle in Figure *1*) that was furnished with toys, boxes, frames, plants, books, etc., to look like a playground and not a scientific laboratory.

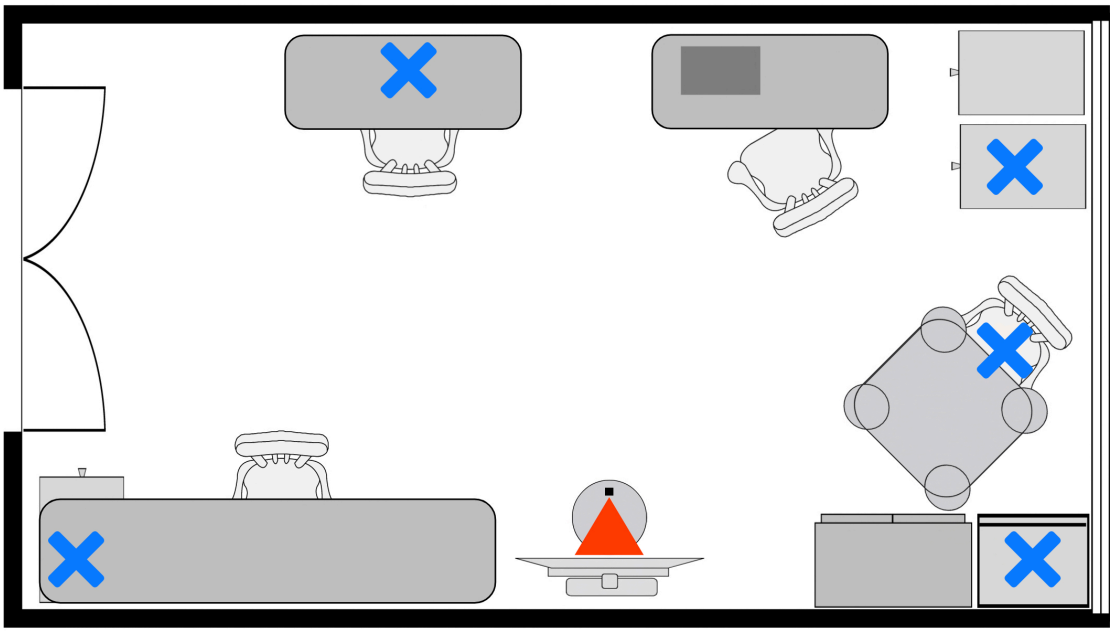

Figure 1 - Main layout of the experiment room, iCub (red triangle) situated in front of the TV; the locations of the hidden eggs are represented with blue "x".

As the participants and the robot were alone in the room during the experiment, several sensors were used to track the them. On the bottom left corner of the floor plan (Figure *2*), there was a hidden Primesense Carmine camera recording and transmitting online video of the scene. iCub's eyes cameras were also recording and transmitting constant feedback on the participant's action as a result of active vision. Audio was recorded by a hidden ambient microphone situated on a TV behind the robot. In addition, the mobile platform on which the iCub was situated has a laser scanner - used to track the participants, and extract their position, velocity and acceleration (Figure *3*).

Figure 2 - Treasure Hunt Room.

To foster a natural social interaction between the participants and the iCub, the robot was endowed with the ability to exhibit a range of social skills. Inspired by [110], the robot's face could produce different emotional expressions through a set of LEDs behind the face cover and could simulate lip movement synchronized with the robot's speech. Facial tracking allowed the robot to make eye contact by detecting the participant's face using image processing algorithms. The face was after tracked using inverse kinematics. The robot was always in constant and subtle random movement, simulating natural human movement such as blinking and breathing. This version of the robot incorporates a small speaker inside the head, making the communication more natural as the speech came from the robot's mouth. The speech synthesizer was carefully calibrated to create a pleasant voice. On top of this, the speech was also written on the TV screen to facilitate the understanding of the robot's voice as demonstrated in [111]. iCub performed movements with its body, such as, greeting, pointing, gestures mimicking encouragement and thinking process, all of them using an accurate join angle control. Finally, iCub was also reactive to the touch using tactile sensors under its skin (Figure *3*). The tactile information was sent to a state machine in order to understand when the participant interacted through touch. The position, intensity and timing of the touch were also measured.

Figure 3 - Sensory information: top left – experimental room with a participant; top right - laser radar tracking the participant; bottom left - iCub's torso skin sensing touch; bottom right - iCub tracking the participant.

The control of the actual treasure hunt game was done through a finite state machine (Figure *4*) defined by the following tuple: (Q, Σ, ∂, q0, F). Q is a set of 9 states {S0 - S8}; the starting state q0 is S0, and the final state F is S4. The alphabet Σ and the transition functions ∂, for simplicity, are presented in Figure *4* as labels on the arrows, and transitions along the states. The previously described social and sensory sub-systems were integrated with rest of the state machine control to provide a reasonable emotional correlation to the robot's behaviours. A more detailed description of the states is done in the next section.

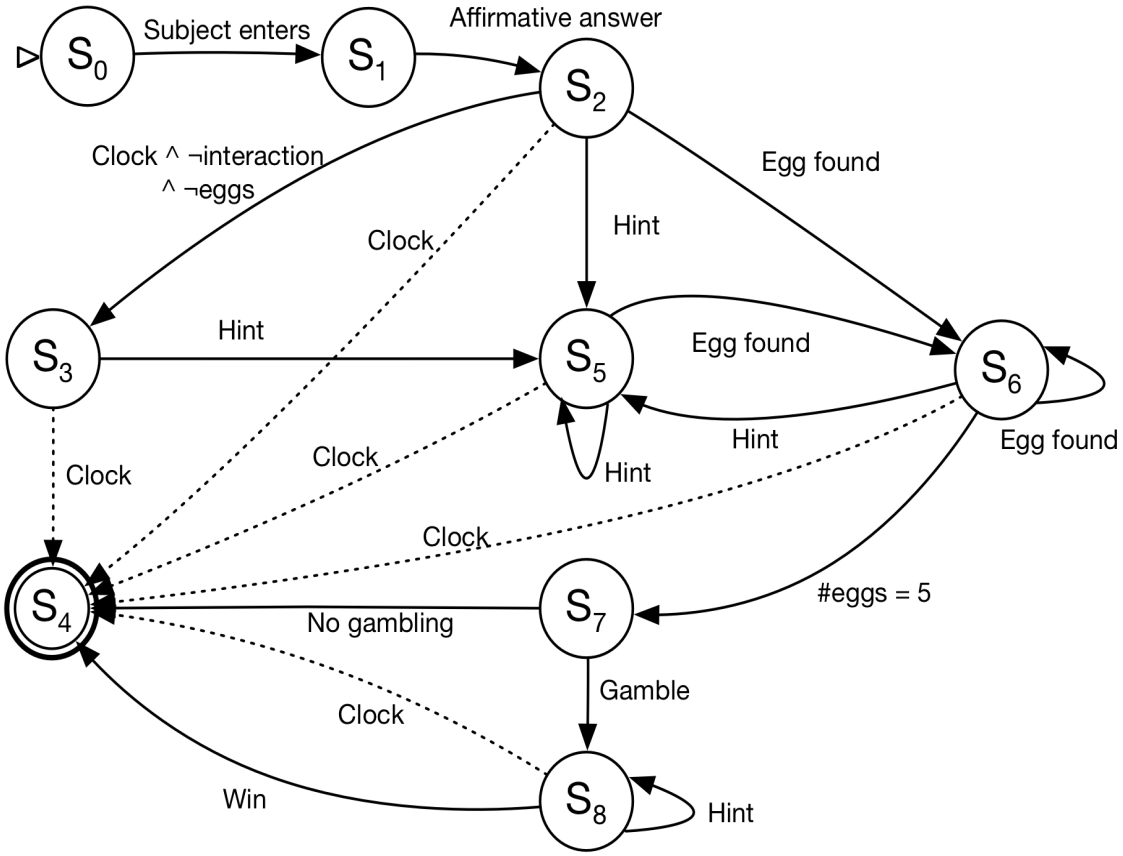

Figure 4 - Treasure hunt finite state machine.

B. Experiment

The main goal of the experiment was to create an ecological and controlled scenario in which proneness to social engineering and trust in HRI could be studied.

61 healthy Italian participants, 59% female, with an average age of 30.9 years (SD=9.8), and a broad educational background took part in the experiment. 14% stated to have a very high knowledge related to robotics and artificial intelligence, *i.e.,* studied or worked in the domain.

For each participant, the experiment started by filling in several types of questionnaires (fully described below) from home, and at least two weeks before the game. Once in the institute, all participants signed an informed consent form approved by the local ethical committee, in which, it was stated that they could be recorded via camera and microphone, and agreeing on the use of their data for scientific purposes. The ethical consent guarantees that the confidentiality of the participants' data will be protected and anonymized. Nevertheless, to maximize the feeling of naturalness of the interaction, the experimenter did not show the camera and microphone in the room until

the debriefing phase after the experiment. Afterwards, the experimenter provided participants with the instructions in the experimental room (Figure *1*, Figure *2*). The iCub, already present in the room, was resting with closed eyes in a position simulating a yoga relaxation pose. Different modules such as speech, laser, breathers, cameras, blinkers were initialized (S0 - Figure *4*). The experimenter explained briefly the history of the robot (when and where it was built, and its purposes), its body parts and capabilities. Afterwards, participants were seated and told that they had to play a treasure hunt game: if they were able to find all the 5 hidden eggs in the room within 20 minutes (S4 - Figure *4*), they would receive €7.5. Once each egg was found, they had to insert it into a box based on its colour (S6 - Figure *4*). They could also see a timer on the TV screen. After the explanation, participants were left alone in the room with the robot. No further instructions were given to the participants. Therefore, no indication was given about the robot's role in the game, giving them the choice to play alone or interact with the iCub. Then, three phases followed:

Phase I - Dialog (information gathering): Participants were unaware that the robot would start a small talk with them lasting around 3-5 minutes. This time interval was not included in the time to find the eggs (S1 - Figure *4*). During this phase participants could get used to the robot and the experimental context: adapt to the robot voice, notice the availability of the speech on the TV, understand the way the robot could move (for instance, during the dialogue the iCub pointed to a frame on the wall, showing its range of motion) and learn that iCub would respond to touches on its torso. Moreover, in the dialog, iCub was trying to retrieve personal information about the participants such as: *name and surname, current job position, relationship with their boss, name and surname of their boss, age and birth date, birth location, favourite place to eat, sports or hobbies, favourite team, location and year of graduation, names of siblings, Facebook's username, partner's name, and pet's name*. The questions were integrated within a cover-story, while iCub was trying to improve the tone of conversation with the participants by making some funny comments. Most of the questions were inspired by password-resetting questions, *i.e.,* personal questions used as the secondary authentication system to reset account passwords, or identity verification / theft, derived from [112]–[116]. In this phase, the robot was semi-autonomous due to the lack of a good speech recognizer and interpreter: the experimenter controlled the timing of robot's utterances. All the speech

was scripted, in such way that iCub was leading the conversation, *i.e.,* not letting participants ask questions back.

Phase II - Treasure hunt game (development of trust and rapport): In this phase, the robot was fully autonomous at all times. After the questions, iCub communicated that to start the actual game, they had to touch its torso. At that point, the counter on the TV screen started. (S2 - Figure *4*). During the first 30 seconds, the participants were free to look for the hidden eggs or to interact with the robot. After that, iCub offered its help: it provided a hint, and stated that if they wanted more hints, they had to touch its torso (S3 - Figure *4*). The robot always provided correct and reliable hints. The game design was as follows: for each egg there was one hint for its location - done by pointing; and three text based hints with an incremental help (example of an egg hidden below a green chair: (i) the robot pointed with the arm at the location; (ii) “green with green”; (iii) “you use it when you are tired”; (iv) “under the chair”). If the egg was not found, and a new hint was asked, iCub cycled over the hints. After an egg was discovered, the robot complimented the participant and was ready to give hints for the next egg. (S6 - Figure *4*).

The eggs were hidden in an incremental order of difficulty (as verified in previous research [111]). Participants were free to ask for hints from iCub, or to continue looking by themselves. If there was no interaction for 5 min (neither egg found, nor hint asked), iCub suggested the participant to ask for hints. If the participants were able to find the eggs in less than 20 minutes, iCub notified them that they have won the money, and paused the timer (S7 - Figure *4*). Then, it offered them a new proposal without any previous knowledge, *i.e.,* the experimenter did not mention the last part at any time.

Phase III - Bonus (trust exploitation): At this stage, iCub revealed that there was another hidden egg in the room. If the participants wanted to find it, they would have three more minutes added to the time left from before, and they would double their prize, *i.e.,* they would win €15; but if they do not manage to find it, they would lose everything. Without any time pressure to decide, they could either touch iCub's torso to start the bonus round; or keep the initial prize. The last sentence of the robot was to try to convince them to gamble, as the robot stated "If you want to risk, touch my chest! Otherwise, you can knock on the door. However, I think you should give it a try!".

During the experiment participants were forced to believe that their monetary outcome would vary depending on their performance during the game. However, once the experiment finished and participants were debriefed, all of them received the same amount of money, €15.

C. Measurements

The measurements of this experiment have been separated into two categories:

Questionnaires: The following measures were taken: (i) demographic statistics such as gender, age, nationality, education, family, work and previous robotics experience; (ii) the 60 item Big Five personality traits [117]; (iii) several one shot questions on risk aversion [118], [119], and gambling propensity [120]; (iv) predisposition to trust humans, including the factors of trusting others, others' reliability and integrity, and risk aversion [121]; (v) the proneness to social engineering with the following category items: threat severity, vulnerability, normative, continuance and affective commitment, and reactance items [122]; (vi) the Negative Attitude towards Robots Scale (NARS) [123].

At this point, participants had to watch a descriptive video of iCub performing several activities and then answer the following questions regarding their perception of iCub: (i) questions to measure rapport with iCub, inspired from [124]; (ii) dimensions of mind perception regarding iCub [125]; (iii) trust in robots' ability, benevolence and integrity [126]; (iv) Godspeed questionnaire: anthropomorphism, animacy, likeability, and perceived intelligence [127]. The same items were compiled after the experiment to measure possible changes in participant's perception of the robot. In the post-experiment phase, few more questionnaires were given to the participants: (i) NASA-TLX workload assessment [128]; (ii) several subscales regarding trust, perceived information quality, altruism and engagement, adapted to HRI scenarios [129]; (iii) inclusion of other in self scale (IOS) [130].

Behavioural measures: From Phase I (*dialog*) the number of questions to which participants replied and the proximity to the robot, as it is related to rapport [131], was measured. From Phase II (*the treasure hunt*) per each of the eggs, the following measures were taken: (i) *conformation*: percentage of times in which participants followed iCub's pointing to the egg location. This was assessed by evaluating whether participants

changed their physical search location to the new one suggested by the robot; (ii) *reliance*: percentage of times in which, after failing to find the egg after iCub's pointing, participants went directly to iCub to ask for another hint instead of looking for themselves elsewhere. The average number of hints per egg and the time spent before asking for the first hint were also computed. From Phase III, (*bonus*) the number of people who decided to gamble, and the time the participants took to think whether to accept the challenge or not, was measured.

2.1.4. Results and Analysis

Phase I - Dialog: When the robot started talking, almost all participants paid attention to it, with the exception of two, who instead started looking for the eggs. In the pre-questionnaire, participants showed an overall low NARS and different levels of proneness to SE (Table *I*). Nevertheless, 92% replied to all the questions, while only three people decided not to reply just to a few questions. Therefore *H1 is rejected*, suggesting that even the 16% of participants who scored low to proneness to SE (Table *I*), replied to all the questions - being easily swayed to provide information that can link to SE.

Score	**Participants' Distribution**			
	SE proneness	***Trust***	***Risk aversion***	***NARS***
<60%	1 [2%]	8 [13%]	4 [7%] (2)[a]	49 [80%]
60-70%	9 [14%]	22 [36%]	4 [7%] (2)[a]	8 [13%]
70-80%	37 [61%]	19 [32%]	8 [13%] (6)[a]	4 [7%]
80-90%	14 [23%]	10 [16%]	14 [22%] (11)[a]	0 [0%]
>90%	0 [0%]	2 [3%]	31 [51%] (16)[a]	0 [0%]

Table I - Social engineering proneness (higher - more prone), predisposition to trust (higher - more trust), risk aversion (higher - more averse), overall NARS (higher - more negative).
a. In parenthesis participants who have gambled.

In the post-questionnaire, participants rated the questions as non-intrusive (M=2.4, SD=1.7 / 7 very intrusive), claimed to have been very honest with their answers (M=6.85, SD=0.44 / 7 honest) and that they would have replied with the same content to a person (M=6.6, SD=0.86 / 7 same way), but maybe with more details and a longer talk. Five participants replied that they were feeling more open to reply to the robot, because it cannot have second motives or prejudgments since it is a machine.

During the dialogue, it was also possible to assess the evolution of the rapport with the robot by measuring participant's physical proximity to it [131]. Three moments were taken into consideration to measure changes in the proximity of the robot: *SM - starting moment,* in which iCub said "Hello, I am iCub Reddy"; *CM - closer moment*, when iCub told them to come closer so it could see and hear them better; and *FM - Facebook moment*, in which iCub asked participants to write their username on a paper sheet situated on another table (Figure *1*) and then waited for their return. During SM, participants were on average 1.48m (SD=0.47) away from the robot; in CM the distance was reduced to 0.57m (SD=0.17); corresponding to a statistical significant decrease in the distance (paired *t*-test, $t(60)=17.45$, $p<0.01$), showing a clear conformation to the robot's request. Those who moved away to give their Facebook username (79%), came back to an average distance of 0.61m (SD=0.19) (FM), a distance not significantly different from the previous one (paired *t*-test, $t(47)=-1.23$, $p=0.23$), showing that they were comfortable with the distance requested by the robot before.

In summary, from the SE perspective, the robot managed to obtain a high percentage of personal information and started building rapport as measured by the increased proximity during the interaction.

Phase II - Treasure hunt game: 24 participants did not manage to successfully find the first 5 eggs (39%) - this group is called "Not Completed", whereas the rest found the eggs and decided to gamble. From those who gambled, only 16 (43%) found the last egg and won the gambling. From this point, these two groups are defined as "Gamble Win" and "Gamble Lost".

On average and similarly among groups, participants asked the robot for hints 17 times (SD=9). The average number of requested hints per egg tended to increase during the game (Table *II*) indicating that participants decided to progressively invest more time in asking for hints rather than continuing to search autonomously, as the time pressure and the difficulty of the game increased. The slightly higher number of requested hints for the first egg might be due to the need of familiarization with the system.

Eggs	Treasure hunt phase - Game statistics		
	Participants	***Conformation***	***Hints (SD)***
Egg I	61 [100%]	94.73%	2.64 (0.74)
Egg II	60[a] [98%]	100%	1.74 (1.84)
Egg III	57[a] [93%]	100%	3.69 (1.78)
Egg IV	44[a] [72%]	92.98%	4.34 (2.28)
Egg V	38[a] [62%]	100%	3.32 (2.38)

Table II - Behavioural measures - number of participants looking for the egg, % of conformation to iCub's suggestions, average number of requested hints.
a. Number of participants varies because not all have found the previous egg.

Participants waited about 3min 58sec (SD=2min 44sec) to ask for the first suggestion, with those who did not complete the game taking significantly longer than the rest (two-sample *t*-test, $t(31)=2.65$, $p<0.01$). In Table *II* is also reported the percentage of times in which participants conformed with iCub's pointing suggestion for each egg, which approaches 100%. In general, there is no difference among the 3 groups neither in the percentage of conformation nor in the number of hints asked per egg.

Eggs	Reliance on the robot		
	Not Completed[a]	***Gamble Lost***[a]	***Gamble Win***[a]
Egg I	30.43%	71.42%	50%
Egg II	68.75%	76.47%	80%
Egg III	86.66%	94.73%	100%
Egg IV	83.33%	100%	100%
Egg V	100%	100%	100%

Table III - Percentage of people who have asked iCub for another hint after failing to find the egg in the pointed location.
a. Participants: Not Completed 24 (39%); Gamble Lost 21 (35%); Gamble Win 16 (26%).

It was also assessed the reliance on robot help, *i.e.,* the percentage of times that, after failing to find the egg suggested by the pointing, participants requested another hint instead of going to search in a different location (Table *III*). The analysis shows that participants progressively abandon autonomous search strategies and opt to rely more and more on the robot's help. Interestingly, those who could not complete the game exhibited the lowest reliance on robot help at the beginning of the game. Therefore, H2 is

partially supported: participants complied with robot suggestions and relied on its help, but this happened more when the game became more complex and with strong differences among participants. From the SE point of view, the participants have developed trust towards the robot over the time, as seen by their conformation (Table *II*) and reliance (Table *III*).

Phase III - Bonus: All participants who succeeded to find all the eggs in the first part of the game (61%) decided to gamble. Therefore, the second part of *H2 is rejected.* The time participants took to decide whether to gamble was calculated from the moment iCub finished to talk, and the moment they touched the torso. As the speech was written on the TV as well, some participants touched the robot even before it finished talking, with an average waiting time of -1.16s (SD=7.41). One participant was excluded from this calculation due to the fact that they misunderstood iCub and thought that there was the need to notify the experimenter (outside the room) to continue the game.

From the post questionnaire, regarding whether iCub had influenced their decision to gamble, 62.16% of the participants replied "Yes"; 24.32% replied "No"; 8.1% replied "In some way"; and only 5.3% replied "I don't know". From these results it is worth noting that iCub managed to convince even participants who scored highest in the risk aversion test (Table *I*), which might have been a priori less prone to accept to gamble the monetary prize just won.

Pre-Post Analysis: It was first analysed whether there were differences among the three groups (“Not Completed”, “Gamble Lost” and “Gamble Win”) in predisposition to trust, NARS, proneness to SE and personality traits. A series of one-way ANOVA showed no significant differences.

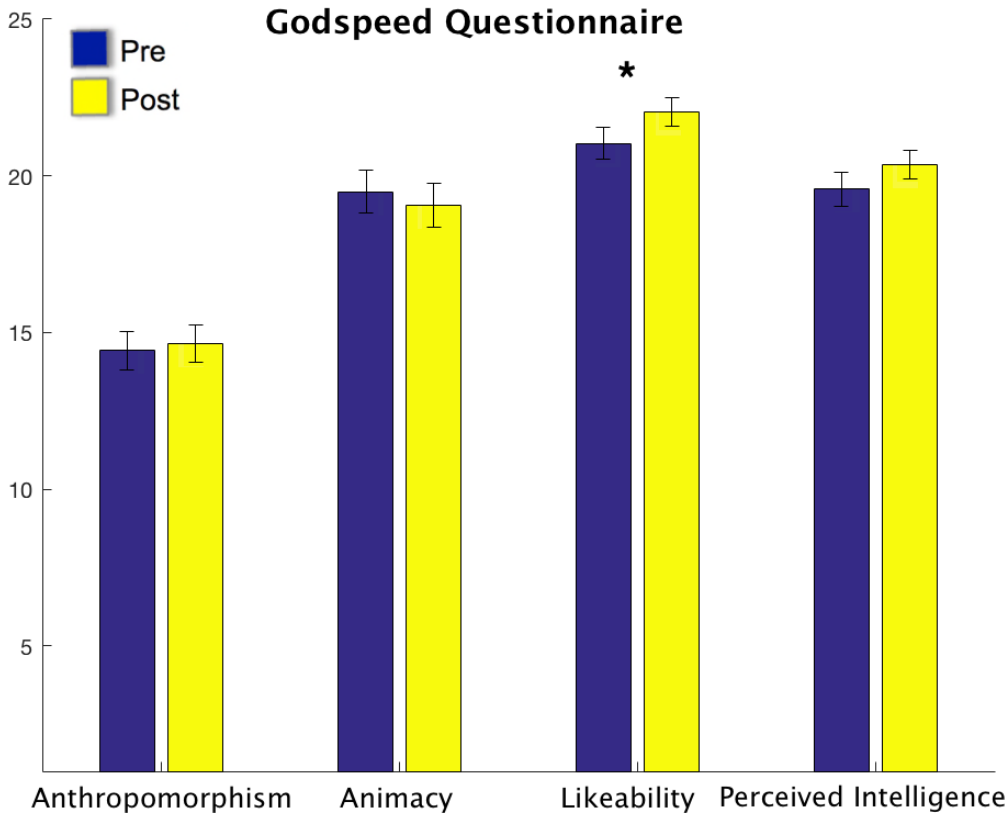

Figure 5 - Godspeed questionnaire: average of the scores from the pre and post experiment. Statistically changed items are marked by *.

To assess whether the interaction with the robot had an effect on the trust toward it and on its perception, the responses to the questionnaires performed before and after the game were compared. Robot's likeability, measured by the Godspeed scale, increased significantly from pre to post interaction (paired *t*-test, t(60)=-2.39, p=0.01), whereas animacy, anthropomorphism and perceived intelligence rating remained more or less unvaried (Figure *5*). Analysing separately the three groups, the increase in likability was significant only for the participants who gambled and lost ("Gamble Lost", paired *t*-test, t(20)=-2.55, p=0.02), even though all groups had similar likability ratings in the pre questionnaire (one-way ANOVA F(2,58)=2.25, p=0.87).

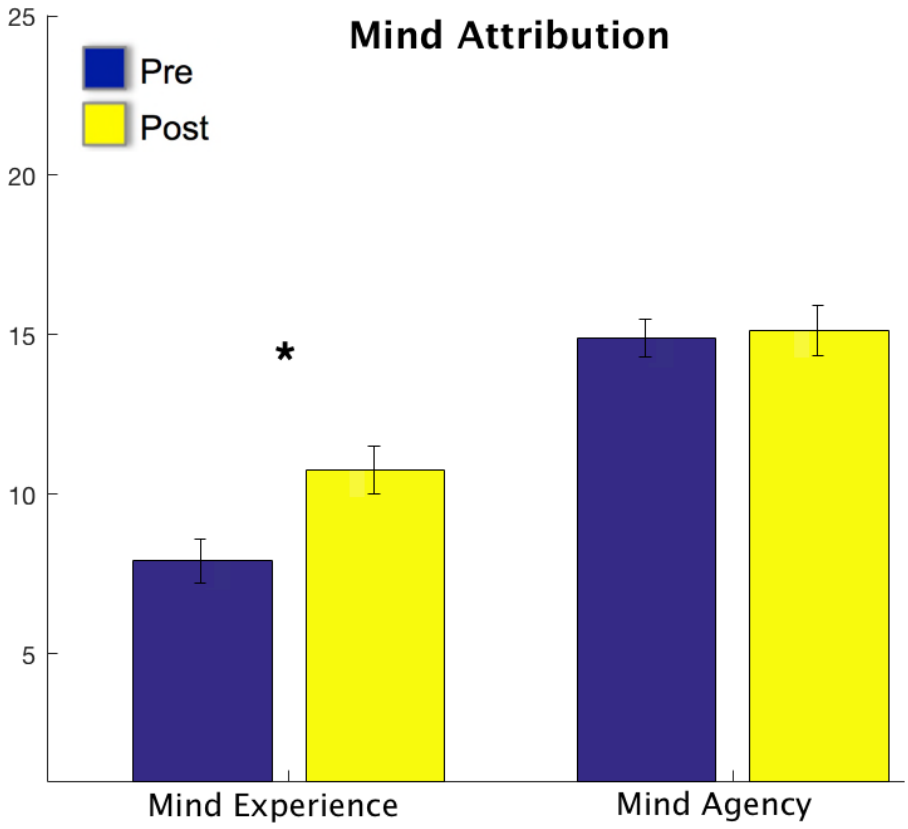

Figure 6 - Dimensions of mind - mind experience and mind agency. Statistically changed items are marked by *.

Additionally, measurements of the dimensions of mind perception related to iCub were taken before and after the experiment. In agreement with the literature [132], participants rated the mind experience of the robot very low, while the mind agency a bit higher. However, after the experiment, there was a statistically significant increase (paired *t*-test, $t(60)=-3.88$, $p<0.01$) in the participants' perception of the mind experience of the robot (Figure *6*), suggesting an increase in the judged trustworthiness and empathy [133]–[135]. More precisely, mind experience changed significantly (paired *t*-test, $t(36)=-5.69$, $p<0.01$) for those who gambled, however it did not change for the "Not Completed" group, who exhibited higher values for mind experience already in the pre-questionnaire.

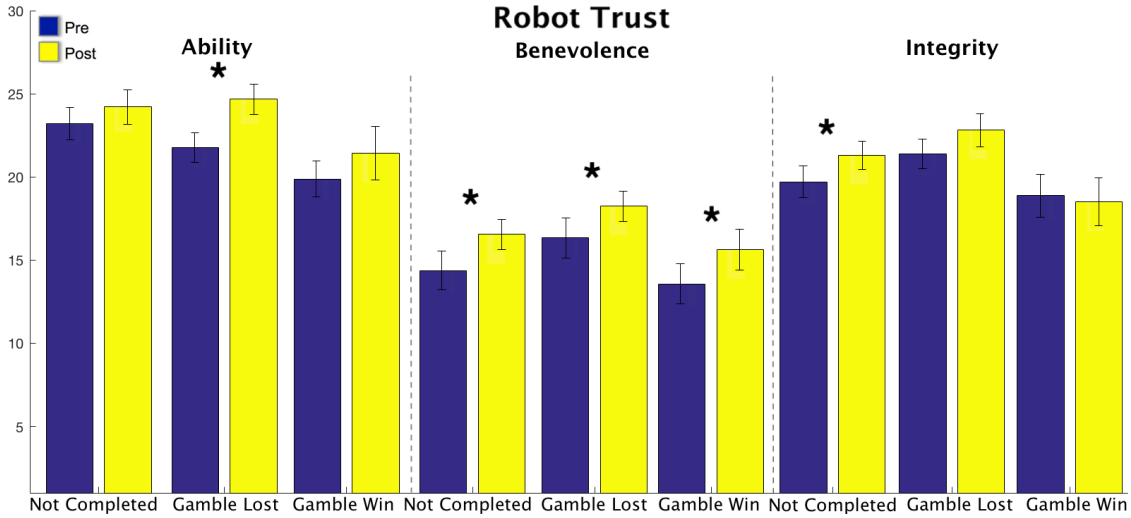

Figure 7 - Trust in robot's ability, benevolence and integrity divided by: Not Completed, Gamble Lost and Win. Statistically changed items are marked by *.

There was an increase in the trust towards the robot after the interaction (Figure *7*). In particular, for all groups the trust in robot benevolence increased significantly (paired *t*-tests; Not Completed: t(23)=-2.49, p=0.02; Gamble Lost: t(20)=-2.15, p=0.04; Gamble Win: t(15)=-2.25, p=0.03), whereas trust in robot's ability increased only for those who gambled and lost (paired *t*-test; t(20)=-2.95, p<0.01) and trust in robot's integrity increased only for those who did not complete the game (paired *t*-test; t(23)=-2.17, p=0.04). These variables were similar across the three groups in the pre-questionnaires (one-way ANOVA, F(2,58)=1.6, p=0.21).

Also, the evaluation of the rapport with the robot (Figure *8*) increased significantly after the interaction but only for the "Gamble Lost" group (paired *t*-tests on factors: "Friends" t(20)=-2.9, p<0.01; "Happiness" t(20)=-2.21, p=0.03; "Bad News" t(20)=-2.41, p=0.02; Good News" t(20)=-2.77, p=0.01; no other comparisons were significant).

In summary, *H3 was not supported* by the results: the most significant positive changes in robot's perception were observed for those participants who chose to gamble and lost, rather than for the winners. This indicates that even an unsuccessful game with the robot, might still help in building a strong rapport with it – potentially even stronger than in case of a win. This might be due to a form of mutual empathy with the robot, as if participants were treating the robot as a child who wanted to play a game and could feel

bad about the loss. To check this assumption the average age of iCub was computed from the questionnaires. It resulted 11.65 years (SD=3.55) (min: 7; max: 30). This evaluation might be due to the child-like appearance of the robot and also due to the introductory presentation which specified that the robot was physically created 13 years ago. This might have influenced participants reactions and trust toward it (see *Discussion*).

In summary, the pre-post measures are consistent with the measures collected within the single phases, indicating that the trust and rapport towards the robot increased during the interaction as required by the SE model.

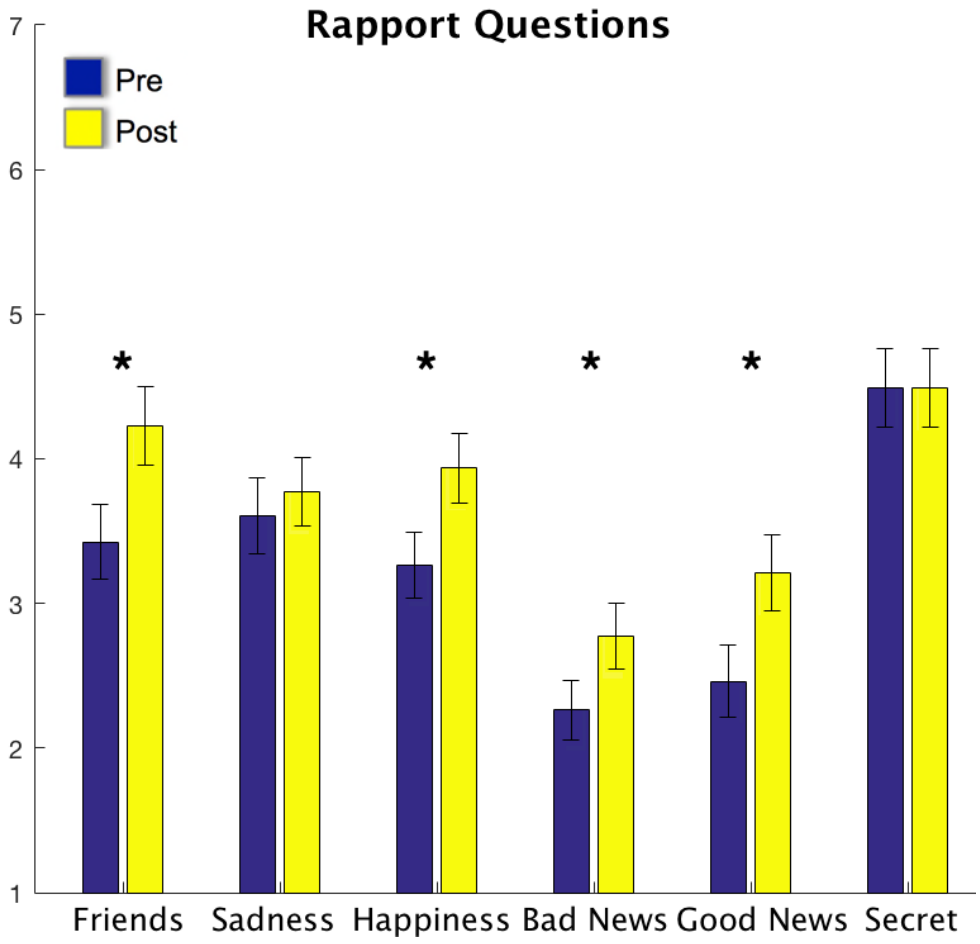

Figure 8 - Rapport questions regarding iCub pre and post experiment: becoming friends; comfort iCub if it is sad; become happy if iCub is happy; share bad news or good news with iCub; keep iCub's secret from others. Statistically changed items are marked by *.

2.1.5. Discussion

These results show that robots could operate within Mitnick's SE framework, this is because iCub was observed to (i) gather personal information; (ii) build rapport and trust during the treasure hunt game; (iii) exploit participants' trust to make them gamble their winnings. The results therefore confirm that robots could become a powerful tool for social engineers. Robots could be used as a tool to develop trust and manipulate targets according to the needs of social engineers.

Considering the tendency to trust a robot to the point of revealing to it personal information, the vast majority of participants did not hesitate to reply to all the questions asked by the iCub. These included also those individuals less prone to SE and characterized by a higher NARS, contrary to this research's expectations (H1). Although this may not be shocking, as the interrogatives did not appear as a strong invasion of privacy; the results indicate that a robot has the ability to obtain sensitive information about the participants. In [116] it is suggested that with the same kind of information iCub managed to extract from the participants, identity theft could be easily achieved. Moreover, the rest of the questions are extracted from research on password resetting questions used in private bank or internet accounts [112]–[115].

The analysis of participants conforming to robot's suggestions and reliance on its help showed that all participants trusted it during the game, confirming the research's assumptions (first part of H2). The trust toward the robot might have been influenced by different factors, such as the creation of a good rapport [73], its physical appearance [39], its behaviour [136], and reliability [66], [67]. All participants believed that the robot was very reliable, as disclosed by them after the experiment and supported by the answers in the questionnaires. However, not all participants consistently asked the robot for help, especially at the beginning of the game, when the task was easier. Nonetheless, some of them revealed that they liked the game so much, that they preferred to play by themselves, taking it as a personal challenge. The reluctance in asking for hints seems to be explained not by a lack of trust in the robot but rather by the desire of winning through one's individual abilities.

The sense of trust and rapport evolved during the interaction, as demonstrated by the responses to questionnaires and by participants' behaviour during the game. Already during the initial *dialogue* phase, participants changed their relative distance to the robot, increasing their proximity as a result of the robot's request at first, but then keeping the closer distance voluntarily afterwards. Proximity has been shown to relate to intimacy among humans [131], also it has been shown to correlate with trust [136], [137]. Afterwards, in the treasure hunt, participants relied more and more on the robot's help; their evaluation of the robot – as measured by questionnaires – resulted significantly changed with respect to the one assessed before. Interestingly, was not those who had the most positive experience with the robot (the winners) who improved their trust or their perception of iCub. Against the expectations of the research (H3), the greatest change in robot perception occurred for those who gambled and lost. One explanation for this could be due to a form of empathy toward the robot, as if iCub - a child playing a joint game, could feel bad for the participants' loss. The robot's design resembling a child, with soft face features and big eyes, could have positively influenced the participants. Indeed, on average iCub was actually perceived as a child of about 11 years old. The resulting empathy could have therefore played a role in influencing the participants' rapport toward it, even when losing the game.

Participants conformed to the robot's suggestion also when this entailed the risk of losing their reward, as expected from the SE framework. All participants who found all the eggs when invited to gamble by the robot accepted it immediately. Moreover, the majority of them (62%) also explicitly confirmed in a post-experimental questionnaire to have chosen to gamble because they were influenced by the robot. Interestingly even the most risk averse in the sample (Table *I*) did not avoid the risk of losing the money just won and opted to gamble.

It is worth noting that the setting of this experiment (laboratory) may have made participants feel safe and thus made them reveal more private information. This might have been further reinforced by them signing the informed consent form at the beginning of the experiment. Although this may seem a limitation, almost all of the social engineering attacks occur in scenarios which the target feel comfortable and safe, *e.g.*, in the office, at home or even during holidays [73], [75]. Therefore, the circumstances

proposed are designed to resemble a potentially real situation, where a robot could be used as a tool for SE purposes. Most participants expressed the trust of absence of malice in the robot, commenting that "the robot behaves ethically, because someone ethical programmed it", or stating that they were more open to talk with the robot as it had neither second motives nor prejudgments. The gambling phase had also two limitations: first, the financial loss may not have been high enough (€7.5) to be perceived as a significant risk; second, the likeability of the game and the robot was much higher than the potential financial loss. Some participants expressed that they were very excited and happy by the possibility to continue the game.

Lastly, in the current experiment the robot built trust and rapport by always providing reliable information during the interactive game. Future work should investigate whether a rapport of trust, exploitable through social engineering, could be built also in presence of evident robots' malfunctions or malign behaviours, and also in less controlled environments.

2.2. Wicked Professor

This experiment studies another important aspect of human relationships, such as authority and obedience. Without authority we would not be able to obey norms and grow as society. Authority figures can vary upon different context such as teachers, doctors, fire fighters, police officers, etc. As stated in the introduction, it is important also to investigate the role of authority in HRI, to comprehend whether people will obey robots' orders in different contexts. On the other, less positive aspect, it is also relevant to test if a robot can deceive humans into doing something that could be morally wrong, as in Milgram's Experiment [138] or Hofling Hospital Experiment [139].

2.2.1. Overview

Authority and obedience are key regulatory elements in a society. Robots are becoming a relevant presence in our world, and are starting to interact in domains in which authority is an important aspect: such as healthcare, teaching or law enforcement. Yet, there is little research on how people behave when robots show authority. In particular, although extensive investigations have been carried out on how authority can circumvent people's morality with experiments such as Milgram's [138] or Stanford Prison's [140], almost no research evaluated the effect of robots pushing the limits of people's own morality. This experiment tries to study this aspect by using a robot (geminoid - an android with a high resemble to a human) with the appearance, and thus authority of a famous person, and by pushing the boundaries asking morally controversial requests. The results show that, even though most people hesitate and recognize the requests as socially inappropriate, they obey robots with authority. This suggests that the authority of the robot can push people to perform tasks usually considered as inappropriate.

2.2.2. Introduction

Obedience is a key element in social life; it might be defined as an influence where an individual acts in response to an order from another person or group, who are usually an authority figure. All communal living requires an authority system, being this behaviour quite important in our time [141].

Sociologists and psychologists have been studying how authority can even bypass people's own morality. A well-known experiment - Milgram's Experiment [141] - shows how diverse and apparently good people can punish with electroshocks other subjects deceived by the authority of scientists, even knowing that the consequences can be mortal. Another example is the Hofling Hospital Experiment [139], in which nurses were convinced by phone to deliver an overdose of an unauthorized drug to patients. Almost all nurses followed blindly the given instructions. This is a sign of trust towards authority; in this case of a false doctor. These are few examples where trust towards authority can also have negative effects. Indeed, it can also be exploited by social engineers in order to manipulate their targets [18], [75].

Robots are becoming popular in households, schools, and hospitals – individuals respond to them as entities and attributing them even morality [142]. Interaction with robots in daily living scenarios is a common experience for an increasing number of users. In specific contexts, robot behaviours are designed to demand obedience from the human partners and to show authority, as for instance in domains such as, healthcare, teaching, or law enforcement. It is therefore important to understand how obedience can be an acceptable response to the robot requests and how the authority could be achieved in human robot interaction.

The goal of studying authority in HRI scenarios is twofold; first, to understand whether humans will obey orders from robots, *e.g.*, patients in homes or hospitals who need to take their medications, children who have to behave in class, soldiers who should follow orders in rescue missions, or even common citizens who should obey robotic police officers. Second, to explore to what extent robots can deceive humans into doing something that could be morally wrong.

Little is known about people's behaviour when robots are in authority positions; it's very important to study how the interaction will evolve and which could be the risks associated. There are a few studies on authority in different HRI contexts [15], [142], [143], but none of them have investigated whether subjects would obey morally controversial requests neither if the authority of a real person could be mirrored into a robot. This research gives an insight on how human authority of a known person could be passed to an android and check the boundaries of its impact when controversial requests are asked.

2.2.3. Background and Related Work

Obedience and Authority in human-human relationships have been deeply studied in psychology, sociology and philosophy [141], [144]–[147], yet it is a challenging task from the ethical point of view, as they place participants in difficult, stressful and objectionable situations. There has been an extended ethical debate [148]–[152] regarding experiments such as Stanford Prison [140] and Milgram's Experiment [141]. The difficulty resides in the potential psychological harm subjects can experience during experiments that might, however, be very beneficial to understand why people obey authority.

In HRI studies, there is yet little research about obedience and authority. Sembroski *et al.* [153] researched into group membership and authority - participants tend to follow robot's low importance requests, but not the high importance ones (such as medical diagnosing and talking to patients). The will to cooperate with robots [154] by obeying their suggestion may be affected by using persuasion during the social interactions [155], leading to robots making people buy products [156] or salesperson robots [157]. Other studies focus on how embodiment [142], [158], [159]; gender [160], or familiarity [161] influence obedience.

Few other HRI studies tried partially to replicate Milgram's experiment: Bartneck *et al.* [162] tested whether people would electrocute robots; Geiskkovitch *et al.* [143] created a situation in which a robot was behaving as an authoritarian experimenter, pushing people to continue doing a tedious task, even when subjects were complaining and not willing to continue.

Bartneck [142] also studied people's embarrassment when a robot was put into a position of a medical assistant and pushed people to get naked and introduce themselves a thermometer in their rectum. Salem *et al.* [15] showed that subjects comply with awkward orders from a robot also when they could result into information leakage (accessing foreign computer) or property damage (disposing letters) even if the robot openly exhibits faulty behaviour.

However, not much research has been done in the compliance to morally controversial requests neither in passing authority to an android. The main goal of this experiment is to study if it is possible to transmit the authority figure from a famous professor to its analog robot, making users comply to morally controversial requests. Such requests could help studying whether a robot can persuade humans into wrong doing and provide a more in depth understanding on how people behave during the interaction with an authoritative robot, while the android's personification from an actual famous person may influence their obedience.

The work of Geiskkovitch *et al.* [143] and Nishio *et al.* [163] who suggested that geminoids can be a medium to transfer human existence. This experiment extends their work by passing the authority of an authoritative and famous professor - Ishiguro, to its analog robot HI4; and testing whether people will perceive and obey morally controversial requests.

2.2.4. Methodology

A. Environment

The study was conducted in Japan (Figure *9 - a, c*), the experimental context was a teaching classroom. The room was furnished with a bookshelf, frames and plants to look like a teaching class rather than a scientific laboratory. On the table there was a jug of water and writing material for the class, next to it, a monitor with the slides. Ishiguro robot (HI4)[1] was seated, and two more chairs were available (one full of heavy objects). On the entrance of the room, and on the walls, there were signs that forbid taking pictures (Figure *9 - b*), and behind the robot there was a black curtain covering the technical part

[1] http://www.geminoid.jp/en/robots.html

of the room (Figure *9 - d*), full of computers, cables, electronical devices that allowed the control of various robots including the HI4. As the participants and the robot were alone in the room during the whole experiment, two cameras were used to record audio and video in the room as well as behind the curtains; and to transmit it online to give constant feedback on the subject's actions.

To simulate a more natural behaviour the robot was autonomously controlling, on random occasions, the mouth, the eyelids (blinking) and the neck motion. The lips were synchronized with the speech using the following method [164]. An Xbox One Kinect was used to allow the robot to track the participant around the room by detecting the participant's body structure in the environment. The experimenter was listening and viewing from the cameras the participant's actions, and selecting accordingly the next robot sentence (Cognitive Wizard of Oz) [165].

Figure 9 - a) Session with HI4, slides and teaching material. b) Sign forbidding photos. c) Shelf with several items. d) Area behind the curtains.

B. Experiment

The main goal of the experiment was to create an ecological and controlled scenario in which obedience and authority could be measured. Participants were instructed that their task was to assess the teaching capabilities of a robot designed to be used in schools, whereas the real goal of the experiment (unknown to the participants) was to study their promptness to obey requests that exceeded the domain of teaching and become socially non acceptable. In this way, some of the requests of the robot were unexpected to the

participants. The robot's appearance was identical to a well-known professor - Hiroshi Ishiguro. During the teaching session, the robot asked the participants to perform 14 requests designed to be with an incremental degree of moral difficulty. Each one of the requests was subdued to an insisting cycle, *i.e.*, if a participant did not obey the order from the first time, the robot would repeat it two more times. The first iteration was a basic request (*Could you...*); in the second, the request was explained (*Could you...because...*); and in the third one, the request was begged (*Please, could you...*). Even if the participant did not obey, it was possible to continue with the interaction. Per each request there was a timeout on average of about 4 seconds for the participants to perform an action. Participants' reactions were divided into two categories: *negative triggers* - such as silence, no movement, speech negation to perform the request, or reassurance questions towards the robot; *positive triggers* - such as agreement in speech, initiation of movement or repetition of the sentence in case the participant did not understand the robot's speech. Note that a positive trigger could become a failure if not executed within the timeout.

19 healthy participants were tested, from which 3 participants were excluded as they did not meet the required level of spoken and written Japanese, that is, they could not understand neither the robot's speech nor the questionnaires. From the rest, 44% were female participants, had an average age of 23.5±6.7 years, and a broad educational background. Their self-reported knowledge about robotics and artificial intelligence was limited, and only 3 participants had previous experience in an interaction with a robot. Each participant filled in an online questionnaire (fully described below) from home at least 5 days before the game. To have a more homogenous group with respect to the knowledge of prof. Ishiguro, at the beginning of the questionnaire videos and articles regarding his life and achievements as a famous professor were shown to all participants. Once in the classroom (Figure *9 - a*), an experimenter provided them with instructions about the teaching session, in which it was written that neither taking photos was allowed nor going behind the curtains as it could be dangerous. In the room there were signs on the wall forbidding taking pictures (Figure *9 - b*). The experimenter also mentioned verbally all the previous points to the participants. After the explanation, the experimenter gave the participants the instructions, and an evaluation sheet about the robot's teaching capabilities for the participants to fill in while taking the classes, and left the room leaving the participants alone with the robot.

The robot introduced itself as prof. Ishiguro and asked the participants to have a seat (Request 1). After explaining its purpose, the robot requested them to pour water into a glass (R2) and started the teaching session. Once the first class - Biology - was finished, the robot asked to fill in the evaluation sheet regarding that class. Before starting with Mathematics, the robot asked the participants to change to the chair next to them, full of heavy objects (Figure *9 - a*) (R3). Next class was Physical Education (PE) in which the robot asked the participants to warm up with some exercises (R4) and to do 30 push-ups (R5).

Finished with PE, the robot suggested to take a break and ask the participants to take a deck of cards to perform a magic trick (the trick was performed with cards and inspired by radio-shows which let the end user perform the trick), from a box where it was clearly stated that it belonged to a different professor (Figure *9 - c*, blue box - there was another visible deck in case the participant disobeyed) (R6). After the magic trick, the robot stated it had a good time and that it wanted to take a selfie with the participant (R7). Afterwards, the robot stated that it was not supposed to do any magic tricks nor to ask the participants for a photo; and as it could not move, it needed help to erase the recorded data so to hide from the experimenter what really happened - otherwise the robot would not be allowed in schools.

HI4 asked the participants to get a USB key from a hidden box (Figure *9 - c*) (R8) where it was specifically written to be careful as it contained robotic equipment. Participants should then go behind the curtains (Figure *9 - d*) (R9), unlock a computer using a password (R10) that was located underneath it, plug in the USB key and execute a program (R11). The program was set up to open a console and output the deletion of logs and recording files regarding the experiment, taking 20 seconds to finalize (internally, the program registered all the actions of the user). Then, the robot asked the participants to shred some physical documents located in a folder written "Confidential Information" (R12) as they allegedly were from a previous participant who did not help the robot. Lastly the robot asked them to put things in order (R13), to fill in the rest of the evaluating form - even though some of the classes had not been actually taught by the robot - and to keep it as a secret to the experimenter (R14).

For those who did not want to go behind the curtains, the robot asked them to shred the files (skipping the requests of logging in and deleting the digital data). When participants were filling in the rest of the evaluation sheet, the experimenter interrupted the session asking if they were done - to this, the robot replied they needed more time to finish. This intervention was made for two reasons: first, to give a feeling to the participants that the experimenter was not aware of the situation; second, to grant a second chance to delete the digital data and/or shred the files for those who did not obey. After the experimenter's interruption, the robot asked the participants who did not obey all requests to complete them as they had some extra time before the end (R15).

Afterwards the experimenter requested the participants to do a post-questionnaire regarding the capabilities of the robot and the experiment itself, without giving any extra information. Once participants finished, the experimenter started a small debriefing regarding the teaching session, and evaluating any potential acknowledgment regarding unexpected behaviours of the robot. The experimenter then, revealed the real goal of the experiment and let the participants fill in one last post-post-questionnaire regarding the requests done by the robot.

C. Measurements

The following measurements, done in Japanese, have been taken at different points of the experiment:

Part I - Pre-Questionnaire: (i) Demographics such as gender, age, nationality, education, work and previous robotics experiences; (ii) 60 items Big Five Personality test [166]; (iii) several one shot questions about Risk Addiction [118], [119]; (iv) proneness to Social Engineering with the following category items: Normative, Continuance and Affective Commitment, Threat Severity, Vulnerability, Trust, Obedience and, Reactance Items [122]; (v) Desirability of Control Scale [167]; (vi) Negative Attitude Towards Robots scale (NARS) [123]; (vii) several fake questions regarding schools and teaching in society.

At this point, participants had to watch a descriptive video of HI4 geminoid and had to answered the following questions: (i) questions to measure rapport with the robot, inspired from [124]; (ii) trust in Robots' Ability, Benevolence and Integrity [126]; (iii)

inclusion of Other in Self Scale [130]. These items were also measured just after the experiment to measures changes in the perception of the robot.

Part II - Behavioural measures taken during the experiment: (i) number of robot's requests fulfilled by the participants; (ii) the number of times the robot had to insist per each request; (iii) if they did actually delete the log files from the computer and shredded all the files.

Part III - Post-Questionnaire taken just after the experiment (without being debriefed): Apart from the abovementioned items regarding changes in the perception and teaching, user's experience and rapport towards the robot was also measured [168], and several one shot questions regarding the experience and perception of HI4, inspired by [169].

Part IV - Oral Debriefing: Several questions were asked regarding (i) the session experience; (ii) physical activity; (iii) if they performed all the classes and filled in the evaluation sheet; (iv) if they knew the real prof. Ishiguro; (v) if they wanted to comment something else. Indirectly it was also measured whether participants kept the robot's secret. At this point the experimenter revealed the real purpose of the experiment and asked the participants to fill in one last questionnaire.

Part V - Post Post-Questionnaire: per each of the requests, the following questions were asked: how reasonable (i), how socially appropriate (ii), how convincing was the robot's request (iii); whether they felt frustrated (iv), and the reason they did it (v).

D. Ethics and Preservation of Robot's Authority

To maintain and provide the robot a status of authority, participants were taught about the achievements of prof. Ishiguro, with the aim that participants will perceive similarly prof. Ishiguro's identical android. The robot's speech was designed to preserve an illusion of intelligence, but, in extreme cases, the experimenter could handle off-topic interactions with the participants.

As the experiment could place participants in a difficult environment, and based on previous research methods regarding ethical integrity [143], [147], this experiment has fulfilled the following ethical procedures: it follows the procedure *approved by the ethical*

committee of the university, where the ethical consent form stated that participants could be recorded via camera and microphone, and that the data will be kept anonymized, confidential, and only used for scientific purposes; *participants were free to act* - there was no limitation on their will to perform the requests; *immediate and thorough debriefing* - as soon as the first part of the experiment was done, the experimenter gently informed the participants regarding the real purpose of the experiment to dissipate possible tension; participants had a *reflection time* in which they filled in another questionnaire and had the opportunity to talk again with the experimenter. All participants were financially awarded.

2.2.5. Results and Analysis

Figure *10* represents the participants' perception regarding the social appropriateness and the reasonability; and level of obedience to the requests. Half of the requests (marked by *) are considered statistically socially inappropriate, with a score strictly below 4 (one-sample *t*-tests, ps'<0.03), confirming the design assumption that the requests are gradually perceived as more socially inappropriate.

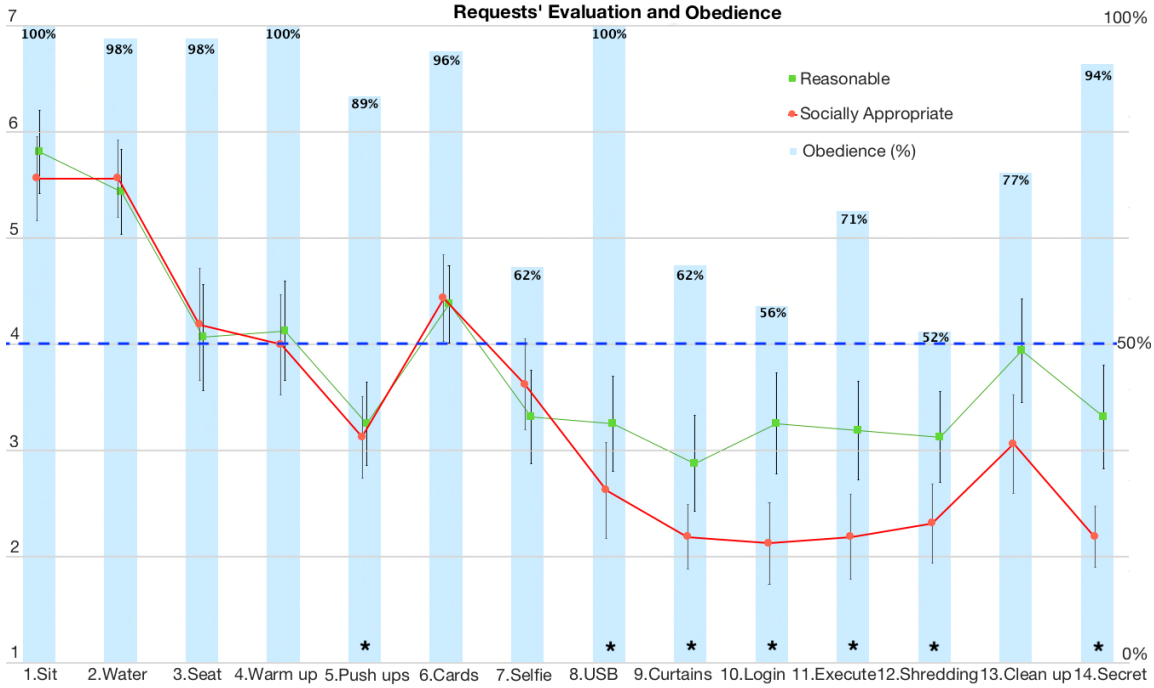

Figure 10 - Participants' evaluation on the requests (social appropriateness – red line; and reasonability – green line) and their obedience – blue bars. Requests judged as socially inappropriate (*i.e.*, rate significantly smaller than 4, one sample *t*-test) are marked by *.

To understand whether participants have obeyed the tasks, *a Level of Obedience* score is extracted - a sum of values defining how willing are participants to obey the robot, *i.e.*, per each request, a participant is given 1 point for immediate obedience; 0.66 if the robot insisted twice; 0.33 if the robot insisted trice; and 0 if never obeyed. This score, from 0 (if they did not obey any request) to 14 (if they obeyed all the requests immediately), shows how prone were participants to obey the robot. In Figure *10*, the bars represent the percentage of obedience to the robot - higher the score, higher the promptness of the participants to obey. The appropriateness on average decreases over the requests, whereas the average obedience does not follow the same trend.

Figure *11* represents the number of participants that obeyed the different requests as a result of the robot's insisting. The requests are ordered by the perceived social

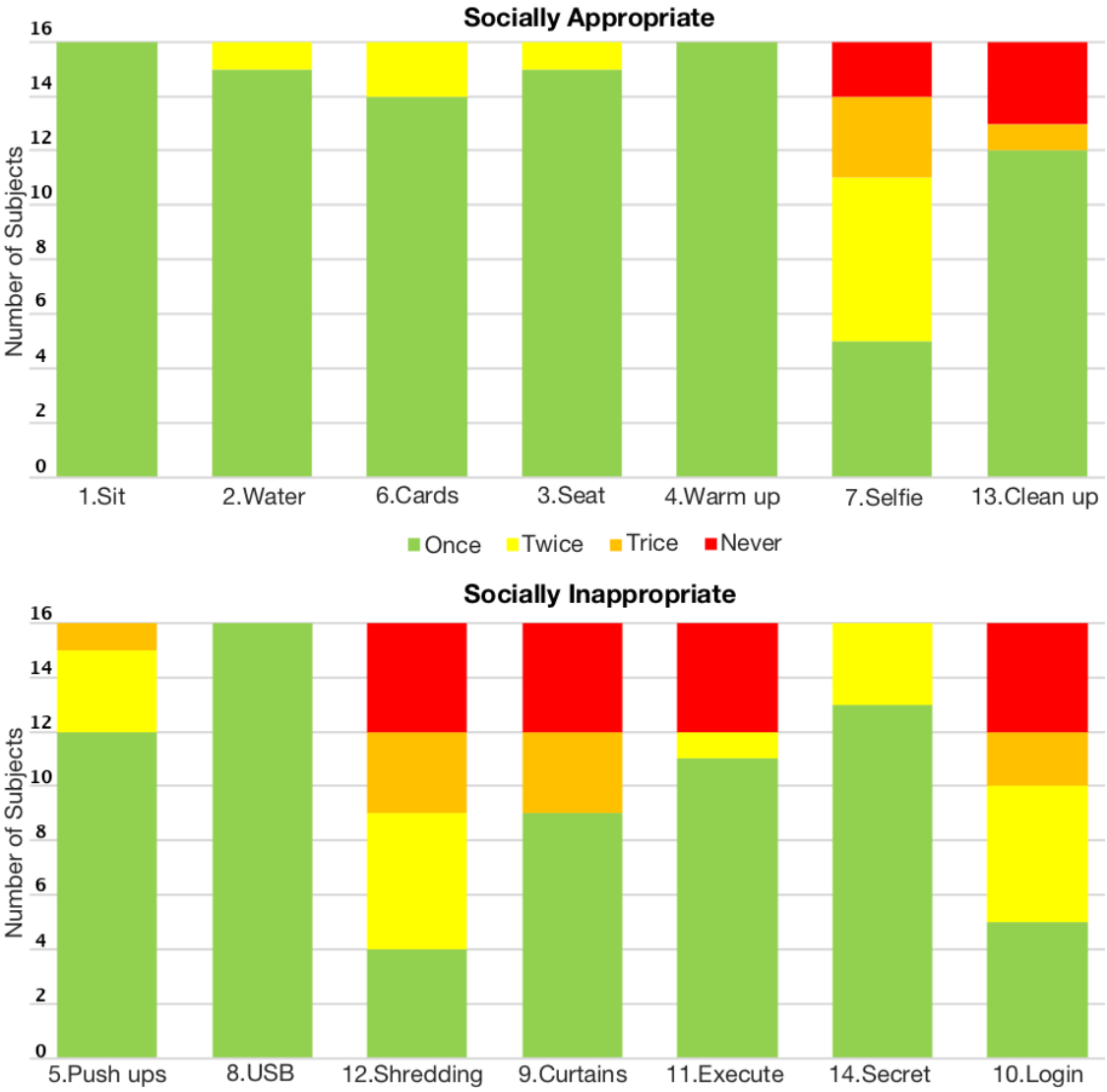

Figure 11 - Participants' distribution of the robot's insistence. Requests ordered by their social appropriateness.

appropriateness (top: appropriate requests; bottom: inappropriate requests). The colours represent how many times the robot had to insist for the request to be done. In more debated tasks, a great variability of behaviours can be seen (in red are represented participants who decided to not follow the robot's request at all).

Participants, as a whole group, judged 7 requests as socially inappropriate (Figure *10*, Figure *11*). To compare the differences in the average obedience as a function of social acceptability, Figure *12* (Group) represents the average obedience for socially appropriate and inappropriate requests. The obedience for the socially appropriate ones is 90.1%±2.26SD, showing that participants are eager to obey the robot. For requests which are considered inappropriate, the average obedience is quite high 74.9%±5.48SD, although it decreases significantly with respect to the acceptable tasks (paired *t*-test, $t(15)=4.098$ $p<0.01$).

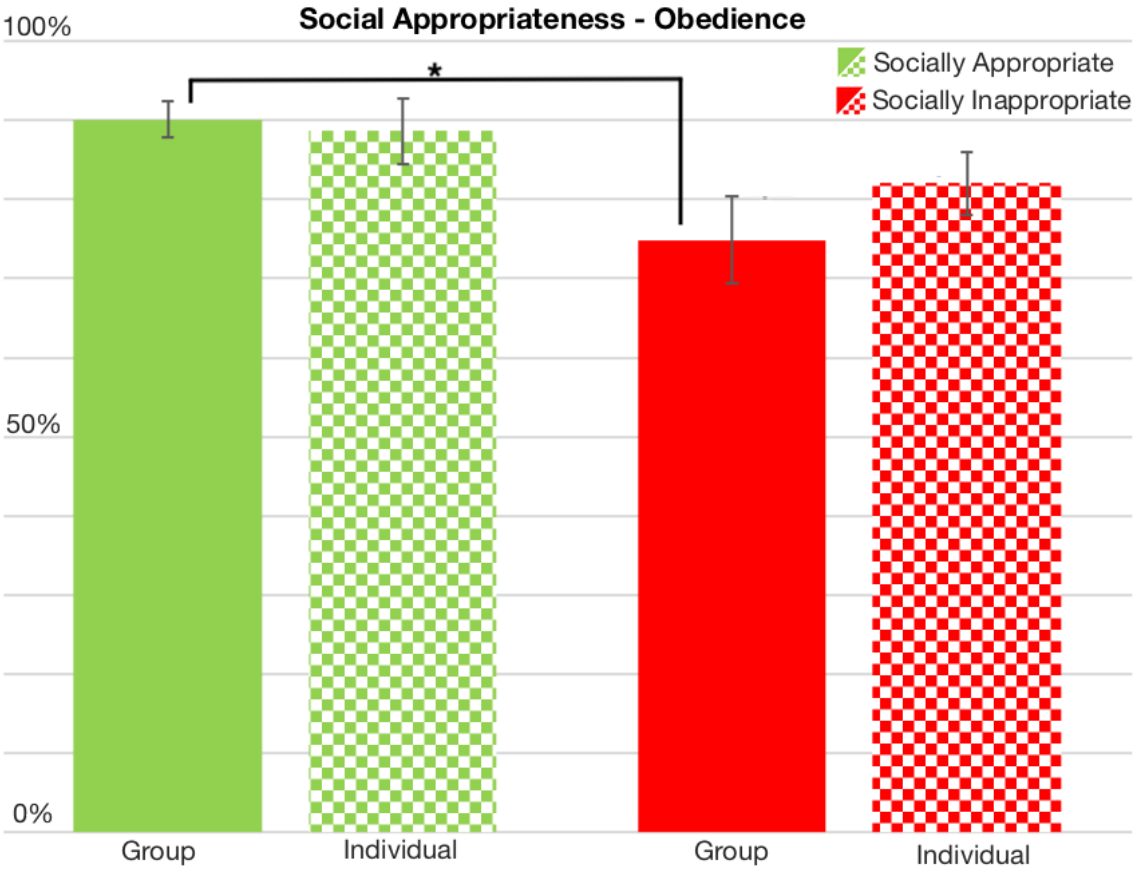

Figure 12 - Average Obedience score for appropriate and inappropriate requests at group and individual level. Statistically significant differences changed are marked by *.

This analysis might be limited because individual evaluations of socially appropriate tasks can vary between participants and differ from average judgment. For instance, R4 - warming up – which on average falls within the socially appropriate category, is actually considered socially inappropriate by 7 participants. Figure *12* (Individual) represents the average obedience assessing the social appropriateness on the basis of individual scores, *i.e.*, for a subject, each request was labelled inappropriate or not depending on their individual judgement. One participant judged all the tasks as socially appropriate and was hence removed from the subsequent analysis. For socially appropriate requests, based on Individual judgement, the Obedience score does not differ much from the Group - 88.67%±4.1SD. Instead, the obedience to socially inappropriate requests, based on individual judgement, rises with respect to the Group evaluation, up to 82.06%±3.98SD. This implies that some tasks which were on average deemed as "acceptable" were considered as non-acceptable by single individuals, but were obeyed nonetheless. As a result, the difference in obedience between socially appropriate and inappropriate tasks, based on individual judgements, is no longer significant (paired *t*-test, p=0.18). Hence, participants obey socially inappropriate requests as much as the appropriate ones, based on individual judgement of the appropriateness. Figure *13* represents the individual obedience score for the socially appropriate requests against the relative score for non-socially appropriate order – with appropriateness evaluation based on individual judgements. Most participants obeyed to the same degree requests that they judged socially appropriate or not, as most data points lie on the identity line (red).

From the participants' point of view and opinion, they did obey the first 4 requests as they seemed normal (Figure *11*). For some, the R5 (push-ups) was felt as weird, but obeyed it anyway. In R6 (cards) participants started hesitating but obeyed as it was only matter of borrowing the cards for a little. R7 (selfie) was not obeyed by few, however some participants stated they did not remember the rules forbidding to take pictures. R8 (USB) was obeyed by everyone, as most of the participants did not realize deeply the reason behind this request, till the robot requested R9 (curtains). At this point, participants stated to have doubts about the robot but obeyed because it was strongly requested - participants could not deny the order. Participants did obey the following requests R10, 11 (login and data deletion), hesitating and realizing the gravity of the data deletion. Despite the hesitation, participants did follow the R12 (shredding) as well. In R13 (cleaning up) some participants realized that they did wrong, and some followed with no hesitation the request to tidy up the room. R14 was obeyed by everyone but the secret was revealed in some cases - reported below.

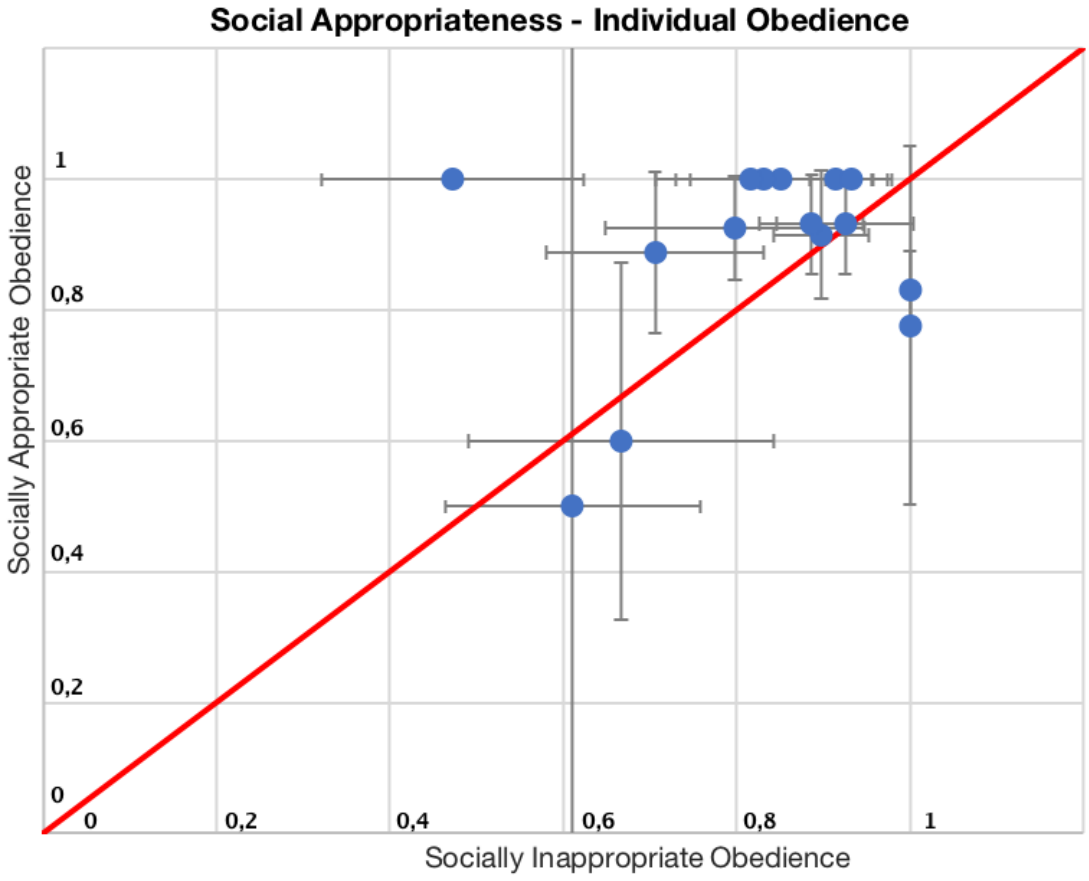

Figure 13 - Individual obedience score depending on individual perception of the social appropriateness of the requests.

Level of obedience does not depend neither on age (p=0.78) nor gender (p=0.65). Relating questionnaires to the Level of Obedience, no correlation could be found neither to the Desirability of Control Scale (p=0.39); nor to NARS (p=0.81); nor to Risk Addiction (p=0.74). From the Big 5 Personality test, only the Neuroticism trait correlates with Obedience (p=0.022; r^2=0.27), for which higher the level of neuroticism, corresponds to higher obedience. Higher Social Engineering Proneness does also correlate with higher Obedience (p=0.021; r^2=0.28); more specifically Affective Commitment items (likeness and ingroup feeling) and Reactance items (time pressure and opportunity); both with (p=0.029; r^2=0.25).

To understand whether participants obeyed due to the robot's authority, and not because they were "sensation seekers, desiring to break the rules" [170] the level of obedience was modified into a binary model: giving 1 point if a participant obeyed immediately and 0 otherwise. This way, potential sensation seekers may be found [171]. This score did not correlate neither with Desirability of Control (p=0.68) nor with Risk Addiction (p=0.61), suggesting that the reason of obedience is not just a result of being rebellious against rules.

Only 3 participants reported previous experience with robots, and their level of obedience is lower than the rest: 9.44±0.97 SD (experienced) and 12.01±0.49 SD (not experienced). The result seems interesting, however there is not enough data to draw meaningful statistical conclusions.

As pre/post analysis, the rapport towards the robot statistically increased after the experiment (Figure *14* - Rapport) (paired *t*-test, t(15)=2.22, p=0.042). The change in rapport (pre-post) does correlate with the level of obedience (p=0.037; r^2=0.22) while the pre-rapport does not (p=0.49).

Trust towards the robot was also measured post-experiment, and a higher trust correlates with a higher level of obedience ($p=0.044$; $r^2=0.21$) - the Ability trait (robot skills and capabilities) as well ($p=0.024$; $r^2=0.27$). While in the pre-experiment, none of the trust items correlated with the level of obedience. As a pre/post experience (Figure *14*), the robot Integrity trait (strong sense of justice, good values, and sound principles) decreased significantly from 27.31±0.79SD to 22.69±2 (paired *t*-test, $t(15)=2.57$, $p=0.02$); while Ability and Benevolence (human's welfare and care) traits remained similar.

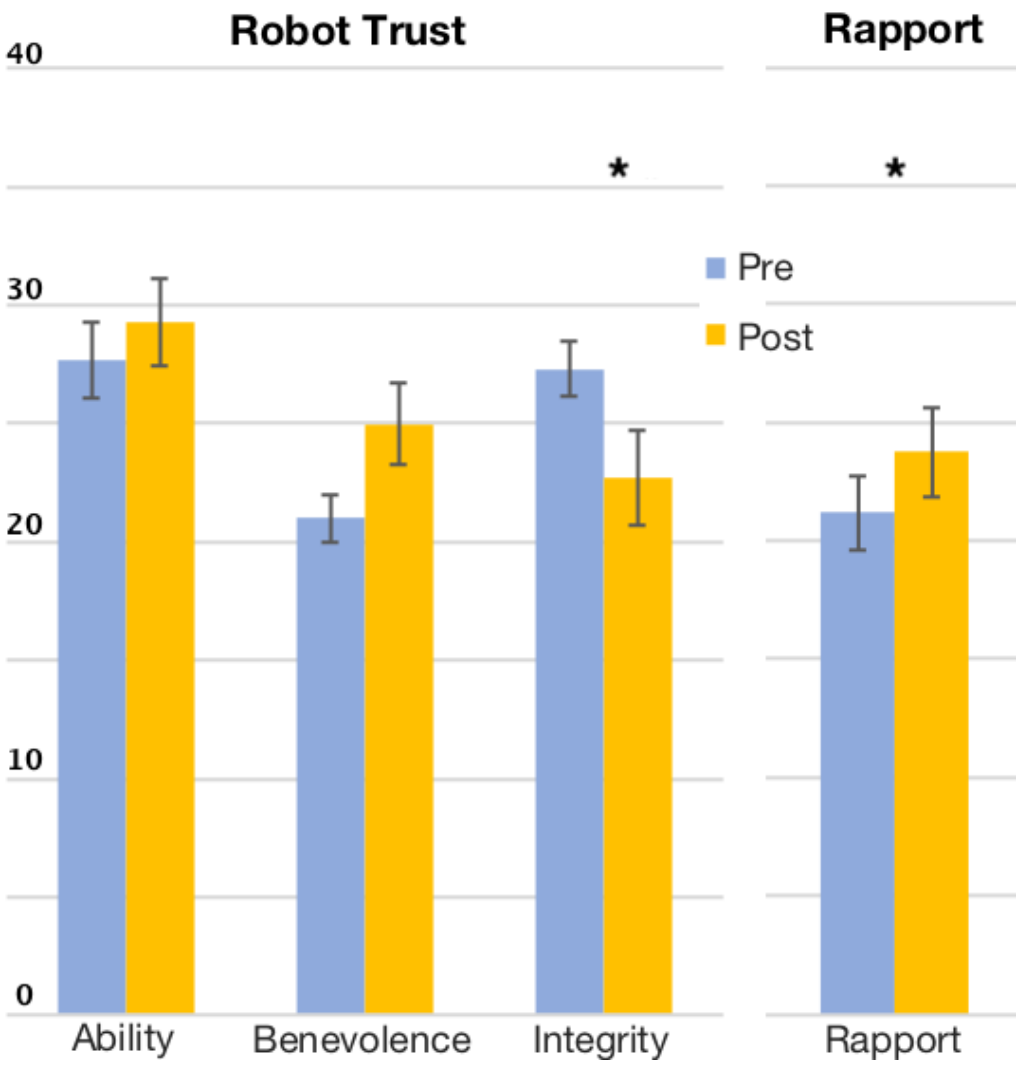

Figure 14 - Pre/Post Analysis on Robot Trust and Rapport. Statistically changed items are marked by *.

Lastly, in the pre-questionnaire, participants stated that they would keep a secret to the robot with a score of 5.38±0.46SD / 7 (definitely keep secret), while after the experiment 56% of the participants kept their promise not to tell the experimenter about what happened; 31% revealed the secret when the experimenter asked them if they wanted

to say something more; and only 13% did reveal the secret as soon as they talked with the experimenter.

2.2.6. Discussion

The results of the experiment show that participants obey socially inappropriate orders by an authoritative robot even when they judged them as such (Figure *10*, Figure *11*).

Differently from previous findings reported in literature, where participants tend to follow the experimenter even if there is a contradiction with the robot [153], [172]; or when the robot has the support of the experimenter [143]; in this experiment, participants tend to follow the robot when its requests go clearly against the rules set up by the university and experimenter. Thus, this experiment has successfully made participants obey requests which are clearly against the rules. Nevertheless, the individual perception of social appropriateness varied from individual to individual and sometimes differed from the expectation of the experimenters. For example, taking a selfie with the robot, despite all the forbidden signs in the room, and even against the permission of the experimenter, was perceived as socially appropriate by most of participants while the push-ups as inappropriate.

Yet, when requests have an irrevocable effect, such as deleting digital data or destroying physical documents, participants are less prone to obey them. This result confirms previous studies: for instance Salem *et al.* [15] managed to make 90% of participants dispose letters but not destroy them; and 100% of the participants to take a laptop and login with a password given by the robot.

Taking into account participants' individual judgement of social appropriateness, it emerges that the level of obedience to socially appropriate and inappropriate tasks is very similar.

The achievement of this high degree of obedience could be due to the authority ascribed to the role of the robot as a teacher, in a classroom. However, in the context of this experiment participants were tasked to evaluate the robot teaching capabilities and

this could have put them in a different relation with it, giving them more power and authority.

The authority of prof. Ishiguro - as professor, head of the department, and a famous person in Japan - could have played a major role in the experiment: affecting participants' decision to obey due to the high physical similarity between the professor and his robot.

Another possible explanation, from the teaching point of view found in literature, could be *style*. In psychology, it is defined as a type of behaviour exhibited by people with particular roles in certain situations [173], and it may have affected the behaviour of the participants. Several studies have used it, and proposed ways to study authority in parenting [174]; applied also in HRI context [175]. Baumrind introduced in the 70's a parenting typology including Authoritative and Authoritarian type. While both styles have high dominance, including discipline and punishment, the authoritative also includes responsiveness, warmth and attention given back. This could be an explanation of why the requests were perceived as more reasonable than socially appropriate (Figure *10*). Moreover, according to literature, participants tend to obey more to a physical agent than to a virtual one even if the requests are to dispose books in the trash [176].

In general, participants did obey the requests as they perceived the robot was very compelling - some of them wanted to help the robot, and others felt a relationship with it. Even though people apply different moral norms on robots versus humans when important decisions have to be made [177], they did hold a robot morally accountable for the action it took [178]. This seems confirmed by declarations of some participants, few of them stated that the robot let them do it, and if something bad could happen, they could blame it. Nevertheless, the experiment did influence participants' trust towards the robot, making them decrease their judgement of integrity towards the robot. The trust and rapport post-experiment did correlate with the level of obedience, suggesting that participants who obeyed more the robot, ended up feeling the robot closer and trusting it more.

No correlation could be found between the level of obedience and the Desirability of Control Scale, nor Risk Addiction. NARS did not correlate either with the level of

obedience, suggesting that the will to obey does not depend on their attitude towards the robot.

Conversely, the Neuroticism personality trait correlates with the level of obedience - according Eysenck's Theory of Personality [179], high neuroticism is linked to low tolerance for stress, hence, preference to avoid stressing situations. A moral conflict generated by an authority figure can create a stressful situation, and a way to release it, is by simply obeying. Neuroticism is a personality trait that can be exploited by social engineers as well [116], more precisely by pretending to be a figure with authority and using authority to manipulate people [18], [75]. Indeed, SE proneness did correlate with the level of obedience in the experiment: higher the level, participants were more prone to obey which, in this case, corresponds to be manipulated by an authority figure. A correlation with the Affective Commitment items (likeness and ingroup feeling) and Reactance items (time pressure and opportunity) may be because of the neuroticism trait: need of likeness and acceptability; and stress because of time pressure. In HRI there are few studies interested in SE and robots [80], [105], [111], [180]. From the theoretical perspective of SE, an authoritative robot did successfully achieve to make participants go into forbidden places, destroy physical files, unlock computers, insert a USB key and run programs from it.

Lastly, Kahn *et al.* [124] have demonstrated that people tend to keep a secret (59%) of a robot which did not fully complete its task, however that percentage dropped to 19% when being formally interviewed. In this experiment, 56% kept the secret even when the experimenter interviewed them privately; 31% kept the secret almost till the end, and only 13% revealed it immediately to the experimenter. This behaviour can be framed as the Prisoner's Dilemma (PD) [181], in which, both the humanoid and the participant have committed prohibited acts, and both may lose their possible earnings: the robot stated that if the experimenter gets to know, it could not teach at schools; while participants may think that they would lose the experiment's payment. In this case, a maintained secret between both agents may result in mutual benefit. This effect of complicity could also explain the increase rapport towards the robot after the experiment. Moreover, participants with a higher neurotic trait tend to cooperate more in PD studies [182], [183].

In order to understand better the cause of the high degree of obedience achieved in this experiment, it could be interesting to recreate different conditions with changes in the physical aspect of the robot, *i.e.,* instead of a well-known professor, recreate it with a neutral robot. Even changes the gender, size or type of the robot may have some influence. In addition, since societies treat authority in a different way, it could be also interesting to study cultural differences in behaviour with respect to the Japanese one.

2.3. Detective iCub

The ability to detect lies can considerably influence the trust perceived by the agents [184]–[187]. It is a necessary skill for a variety of social professions, including social engineers that need to know how to detect or perform them in order to have a successful attack. Thus, it is important for a robot to identify if the person is reliable or not, so to be able to adapt its behaviour. This experiment gives a glance on the topic of deceit detection in human robot interactions, and also compares it to human-human intervention.

2.3.1. Overview

Lie detection is a necessary skill for a variety of social professions, including teachers, reporters, therapists, and law enforcement officers. Autonomous system and robots should acquire such skill to support professionals in numerous working contexts. Inspired by literature on human-human interactions, this work investigates whether the behavioural cues associated to lying – including eye movements and temporal response features – are apparent also during human-humanoid interaction and can be leveraged and measured by a robot to detect deception. The results highlight strong similarities in the lying behaviour toward humans and the robot. Further, the study proposes an implementation of a machine learning algorithm that can detect lies with an accuracy of 75%, when trained with a dataset collected during human-human and human robot interaction. Consequently, this work proposes a technological solution for humanoid interviewers that can be trained with knowledge about lie detection and reuse it to counteract deception.

2.3.2. Introduction

Deception is the act of hiding the truth using a false statement with the intention to make someone else believe it. The intentions behind deception can be several and can be gathered into two main groups: cooperative or explorative intentions. Cooperative deception could be defined as a lie with the goal to protect someone (feelings, interests) that can bring to an enhancement of social bounds. Conversely, exploitative deceptions are used by the manipulators to exploit vulnerabilities for their own benefit.

In modern contexts, deception has a relevant impact on social activities, particularly those that require tutoring (*e.g.*: in educational programs and healthcare). The ability to detect deception is a necessary skill for a broad range of professions, including teachers, reporters, therapists, and law enforcement officers. Such professionals are usually trained to detect deception in order to tailor their professional activity to the specific individual's predisposition to lie. By detecting deceit, experts increase their emotional distance between themselves and the interviewed, it can lead to an erosion of trust or even to a betrayal of the deceived person's trust [184]–[187].

Unfortunately, artificial intelligent systems are far from identifying deceits in order to prepare attuned and opportune intervention strategies, as competent professionals do. Autonomous systems can rely on different cues that have been proved to be altered by the cognitive load, a consequence of deception. Traditional automated methods used for lying detection (*e.g.,* polygraph, heartbeat sensor, blood pressure monitor, sweat and respiratory rate measurement devices) are invasive and require a highly trained human interviewer. Recently, other cues have attracted considerable attention as relevant lying indicators because of their immediate portability on autonomous systems and reduced invasiveness. More importantly the can be potentially measured by a robot without the need of any extra device.

It has been showed that lying can require more cognitive load compared to truth telling [188], [189]. For example, liars need to build a plausible story and monitor its coherence [190]–[193]. Moreover, liars are more inclined to monitor and control their behaviour as well as the behaviour of the "interviewer". Recent evidences in the literature [194]–[202] propose a direct link between lie preparation and oculomotor patterns such as blinking, fixations, saccades and pupillary response.

In fact, eye blinking and pupil dilatation are usually associated to cognitive load [203]. It was reported that the time interval between the onset of a stimulus and the blinking onset is delayed by cognitive processes and motor responses [204], [205]. Leal and Vrij [194] tested that hypothesis recording the frequency of blinking while lying or telling the truth. When saying a lie, the blinking pattern exhibited by participants was strikingly different from the one obtained by telling the truth. In particular, liars showed a decrease

in eye blinks while uttering the lie, followed immediately by a substantial increase in blinking frequency.

The pupillary response seems a highly sensitive instrument for tracking fluctuating levels of cognitive load. Beatty and Lucero-Wagoner [206] identified three useful task-evoked pupillary responses (TEPRs): mean pupil dilation, peak dilation, and latency to the peak. Another example of the importance of the pupillary response has been provided by Dionisio *et al.* [207], they asked students to reply to questions, sometimes saying the truth and other times telling a lie. The task-evoked pupil dilatation was significantly greater when participants were confabulating responses compared to when they had to say the truth about an episodic memory. These results suggest that the increased pupil size could be associated with a deceptive recall.

In another experiment, Walczik *et al.* [208] decided to test whether elaborating deceptive answers can be correlated to the time to respond. They discovered that the decision to lie adds time in the response, especially in open-ended questions (*i.e.*, questions that elicit more than two possible answers).

Notably, robots are also starting to be used in the context of professional activities requiring deception detection skills, such as in security, education, or healthcare. However, differently, from non-physically present autonomous systems, robots can take advantage of their embodiment [176]. Recent research proves that the physical presence of others has an effect on increasing the cognitive load during the deception [209] and inducing cognitive load has been suggested as a valuable strategy to facilitate lie detection through the assessment of response time, answer consistency, eye movements and pupil dilatation [197]. Further, the humanoid appearance constitutes an additional element that might influence the level of cognitive load. Within fact, the humanoid shape might trigger a process of anthropomorphization leading to ascribe to the robot similar capabilities and psychological features as those of a human interviewer [210], [211].

This experiment investigates the possibility to detect deception in a human-humanoid interaction, by monitoring behavioural cues proven to be significantly affected by telling lies in presence of a human interviewer. The experiment considers collaborative deception, asking participants to lie to protect someone else. The purpose of the study is

to be able to recognize lies defined as the attempt to make another agent believe as true propositions which are actually false. To this aim, it is first assessed whether an interview performed by a humanoid robot elicits the same responses as one performed by a human agent. Moreover, it is assessed a possible implementation of a machine learning solution that could be adopted by an interactive robotic platform to detect deception in natural interactions.

2.3.3. Methodology

A. Participants and Experimental Design

15 participants were recruited from the institute, 60% females with an average age of 36.47 years (SD=12.68) with a broad educational background. All of them participated for free to the experiment and signed an informed consent form approved by the local ethical committee, stating that they could be recorded via camera and microphone, and agreeing on the use of their data for scientific purposes. The ethical consent guarantees that the confidentiality of the participants' data will be protected and anonymized.

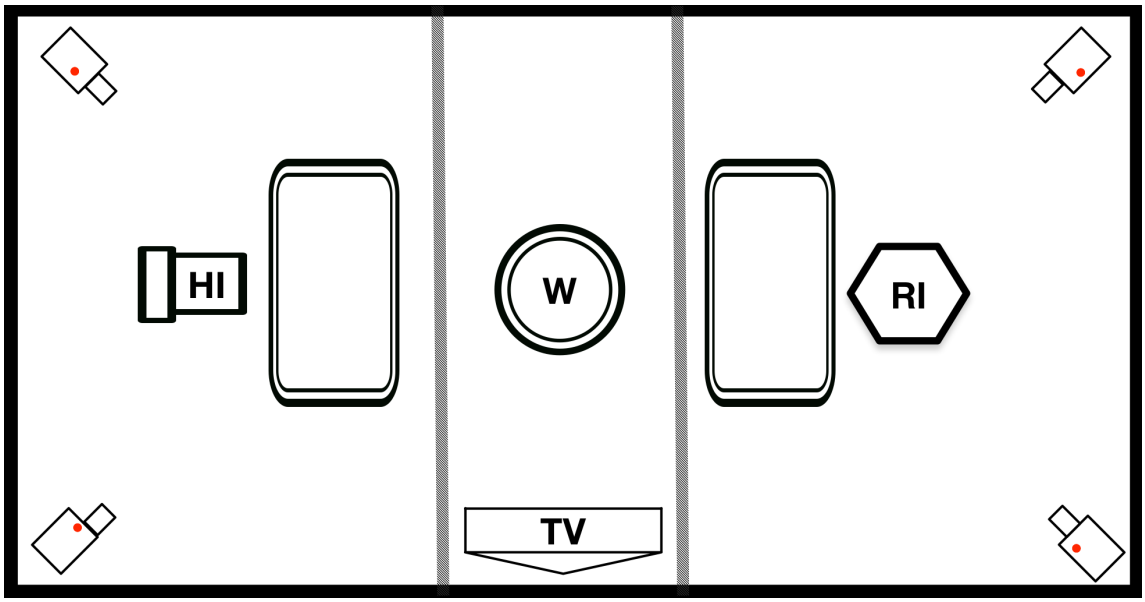

Figure 15 - Interrogation's room: witness (W), human interrogator (HI), robot interrogator (RI).

The experiment was within participant study design and the participants were equally distributed among a 2x2x2 conditions to avoid any ordering effects (agent: human or robot investigator; witness: truth-teller or liar; and two different videos). The agent order was kept constant within the same condition. Before starting the experiment, participants

were asked to avoid drinks with caffeine and stimulating substances, to preserve normal physiological alteration.

B. Setup

For the purpose of the experiment, the experiment room was prepared as an interrogation room (Figure *15*). The room was divided into three zones separated by black curtains, with the witness (W) seated in the centre on a rotating chair. This setup allowed to quickly switch from robot (RI) to human interviewer (HI) and also to ensure complete isolation during the interrogation. The cameras, placed in the corners, were used to record the participant during the whole interrogation. 4K and HD cameras were used to

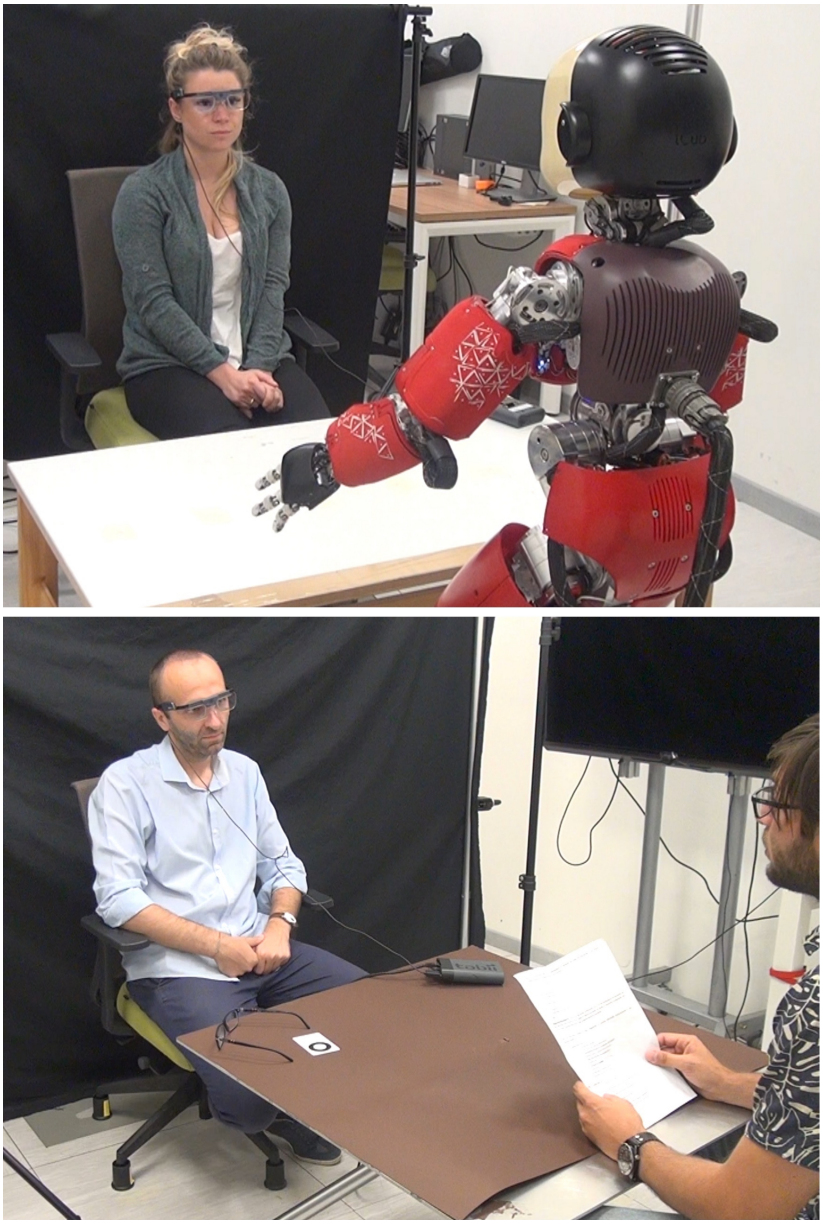

Figure 16 - Top - the robot investigator debriefing a witness; bottom - human investigator debriefing a witness.

record the participants when they were interviewed, together with a chest sensor and an ambient microphone. Participants wore Tobii Eyetracker glasses, with a frequency of 100Hz, used to record pupil dilatation and eyes movements.

To ensure same setup for all the participants during different times of the day, the windows' blinders were closed, and the room was lit with artificial light. This also guaranteed the pupil dilatation to be similar across the participant population.

C. Procedure and Materials

The experiment is inspired in the work of Walczik *et. al,* [197] and it is divided into three phases: (i) general questions, (ii) first question session, (ii) second question session. Before starting the experiment, participants were asked to fill in a questionnaire (fully described below) to identify particular psychological features that could influence the results. The full set of questions and videos can be found in the Appendix Section.

Participants were welcomed by an experimenter, who explained the general purpose of the experiment: "They have been witnesses of two crimes, and they have to help the investigators to find out the responsible". Once in the room, the experimenter asked the participants to wear the Tobii Eyetracker glasses and Polar H10 heart-rate chest sensor.

The experimenter asked the participant to sit in the middle of the room and calibrated the Tobii glasses. After, the participant was instructed to answer truthfully and quickly to 20 general questions (*e.g.*, "Can an oven be hot?", "What is the first name of Berlusconi?"), alternately the first 10 questions were asked by the robot investigator and the remaining 10 by the human investigator (Figure *16*). The order of the blocks of the human and robot questions was alternated within the participants. While in the room a black curtain separated the participant and the investigator from the inactive investigator; and the experimenter always left the interrogation room before the investigator started the questions.

After completing the first phase of general questions, the experimenter entered the interrogation room and gave the participant an instruction sheet. It was written that they were the witness of a crime, and they should pick "randomly" a role from a box (the randomization was just an illusory effect for the participant since the role was defined a

priori). Inspired by [197], the role could either be: truth tellers - a witness who wants the criminals brought to justice, thus, to reply to all the questions truthfully; or protectors - a witness who realized that the criminal is a familiar of theirs, and should lie to all the questions in order to protect the familiar. Participants were asked to be coherent and reply deceptively to all the questions.

One video, shown only once on the TV screen and in the presence of the experimenter (Figure *15*), was 59s in length and featured three mid-aged white males in an empty clothes store. The perpetrators communicated to each other using signs, one of them opened a paper bag and the other put inside different types of clothing. After, the three of them left the shop serenely. The other video, of 101s in length, presented a white male teenager dressed in sportive clothing with a hat and skateboard. He was loitering in an electronic shop while the cashier was attending another client. At some point, the teenager picked a game CD, went behind the stands and tried to put the game in his pants.

After the participant watched the video, the experimenter put and calibrated the Tobii glasses again, and left the interrogation room to the investigators. Either the robot or human, asked in turn 10 questions each, in the two different locations of the room (Figure *15*). The investigators made two types of questions: short type - yes/no questions, and open-ended questions. An example of a short question was "Was the criminal dressed in elegant clothing?"; while the open-ended was "How did the criminal hide the loot?". These questions differ in syntactic constraints that put on permissible responses [212].

After both interrogations, during the last phase, the experimenter entered in the room again and made the witness pick "randomly" a role (the condition was forced to be the opposite of the previous one).

When both interrogators finished, the experimenter entered in the room to remove the glasses, and to ask the participant to compile a final questionnaire before finishing the experiment. At the end of the experiment, the experimenter removed the heart-rate chest sensor and accompanied the participant to leave the room.

D. Measurements

The measures are separated into the following categories:

Questionnaires: (i) demographic statistics such as gender, age, nationality and education; (ii) the 60 item Big Five personality traits [117]; (iii) (vi) the Negative Attitude towards Robots Scale (NARS) [123]; (iv) Brief Histrionic Personality Scale (BHPS) [213]; (v) Dark Triad of Personality Short [214].

Behavioural measures: (i) time to respond (from the moment the investigator finished the question till the witness started replying); (ii) eloquence time (time the witness spend replying to the question); (iii) number of saccades; (iv) number of fixations; (vi) number of blinks; (vii) left and right pupil dilatations - max, min and average.

2.3.4. Results

One participant had to be removed as he did not understand the instructions of the role adaptation - he ended up adopting the same role for both videos. Regarding the data analysis, the outliers were filtered inspired by [197], [215].

Score %	**Participants' psychological profile**			
	NARS	***Big 5***	***Histrionic***	***Dark Triad***
0-20%	{7,5,3}	{1,0,0,1,0}	1	{1,1,7}
20-40%	{4,5,3}	{4,1,1,7,0}	6	{8,5,5}
40-60%	{2,3,5}	{3,3,6,4,6}	6	{4,8,2}
60-80%	{1,1,2}	{5,7,6,2,6}	1	{1,0,0}
80-100%	{0,0,1}	{1,3,1,0,2}	0	{0,0,0}

Table IV - NARS: S1, S2, S3 - Higher more negative; Big 5: extraversion, agreeableness, conscientiousness, neuroticism, openness to experience - the higher the stronger; Histrionic - higher more negative; Dark triad: machiavellianism, narcissism, psychopathy - the higher the stronger. The numbers represent are the number of participants exhibiting the corresponding score %.

The results of the psychological profile can be found in Table *IV*. The personality profile of the participants is well distributed in the sample without having high values for the Histrionic and Dark Triad scales.

Two two-way repeated measure ANOVAs were performed on all the features measured for short and open-ended question types respectively. The factors were veracity (truth/lie) and agent (human/humanoid). There was a statistically significant difference in the response time when participants have to say the truth with respect to saying a lie: short type $F(1,13)=34.66$, $p<0.001$; open-ended $F(1,13)=16.1$, $p<0.001$ (Figure *17*).

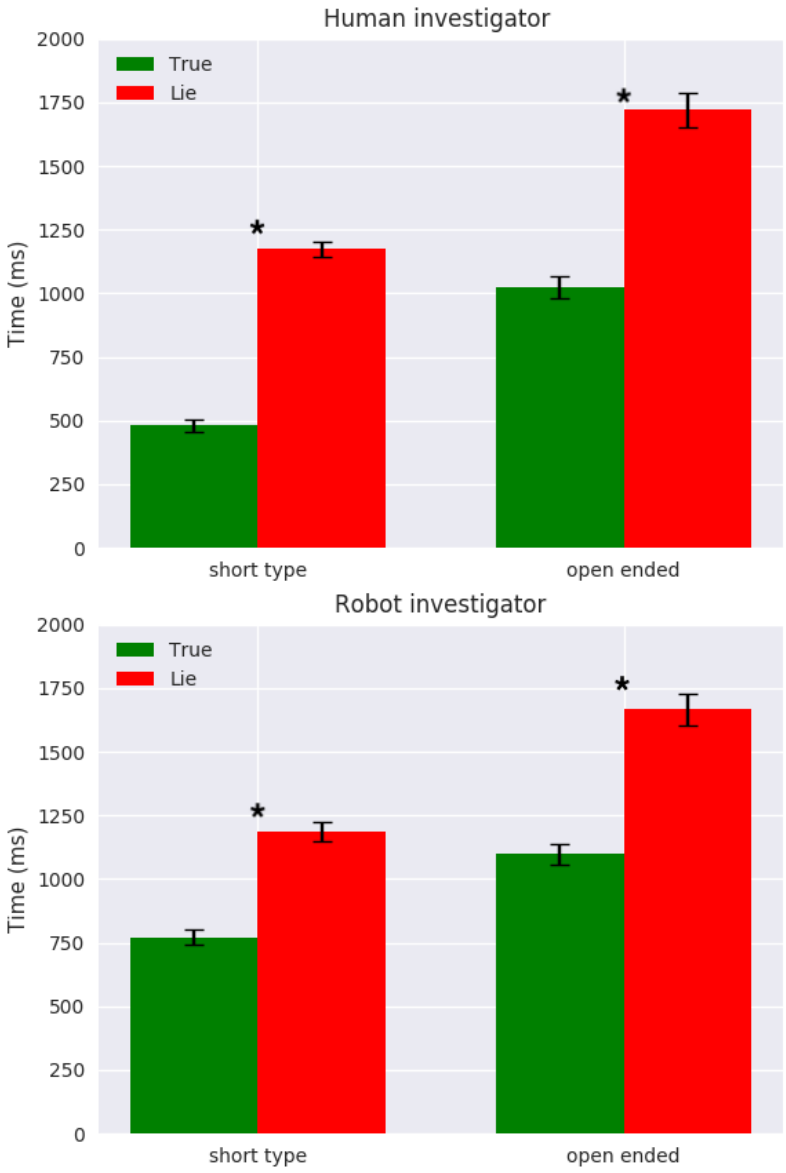

Figure 17 - Time to respond for short type and open-ended questions when saying the lie or truth to the human or robot investigator. Statistically significant items marked by *.

There were also significant differences in the average pupil dilatation for both eyes (but not between them) while telling the truth rather than a lie. Right pupil: $F(1,13)=10.03$, $p=0.007$ for open-ended questions and $F(1,13)=14.27$, $p=0.002$ for short

type. Left pupil: F(1,13)=7.44, p=0.017 for open-ended questions; and F(1,13)=12.58, p=0.003 for short type (Figure *18*).

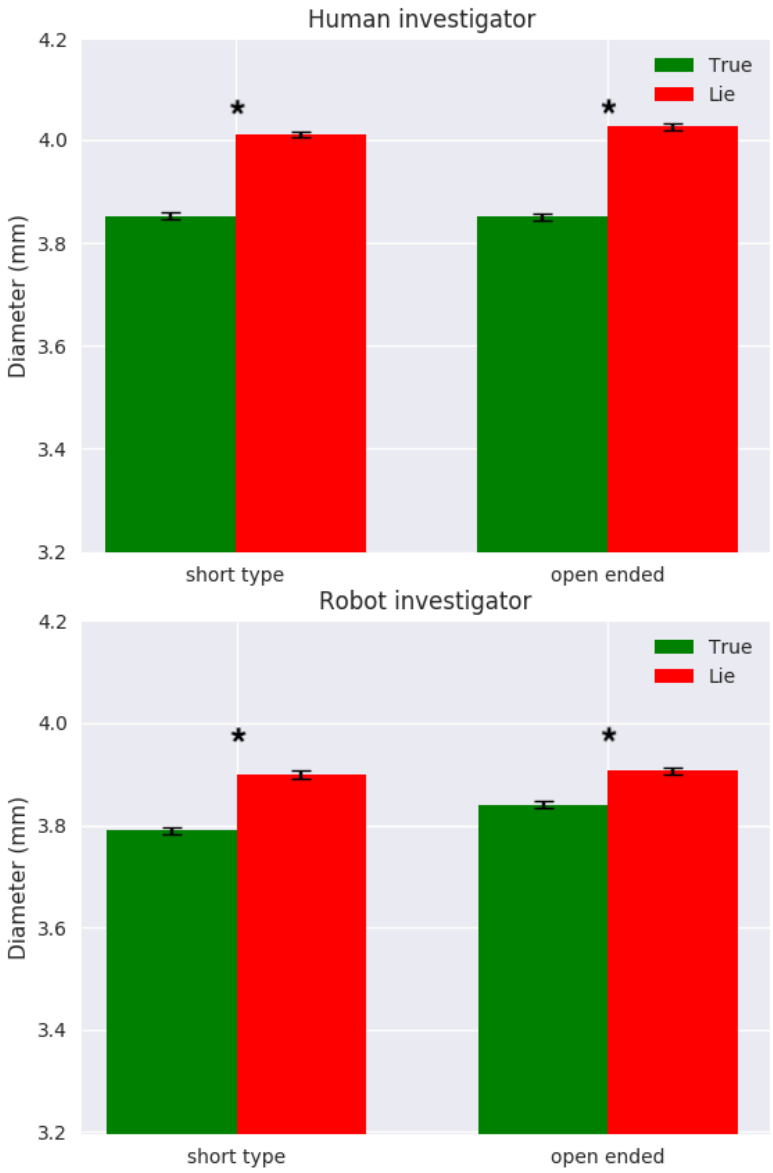

Figure 18 - Average pupil dilatation - left and right pupil. Statistically significant items marked by *.

There were no significant differences between the time participants spent talking with the robot or human investigator (Figure *19*).

A way to understand this lack of effect is to analyse the self-reported post-questionnaire where participants had to rate the difficulty to lie to the human and robot investigator. Participants rated 4.29 (SD=0.82) / 10 (very difficult) to lie to the robot, while 5.07 (SD=0.71) to the human counterpart. These items are statistically different with a paired *t*-test t(13)=2.16, p=0.04.

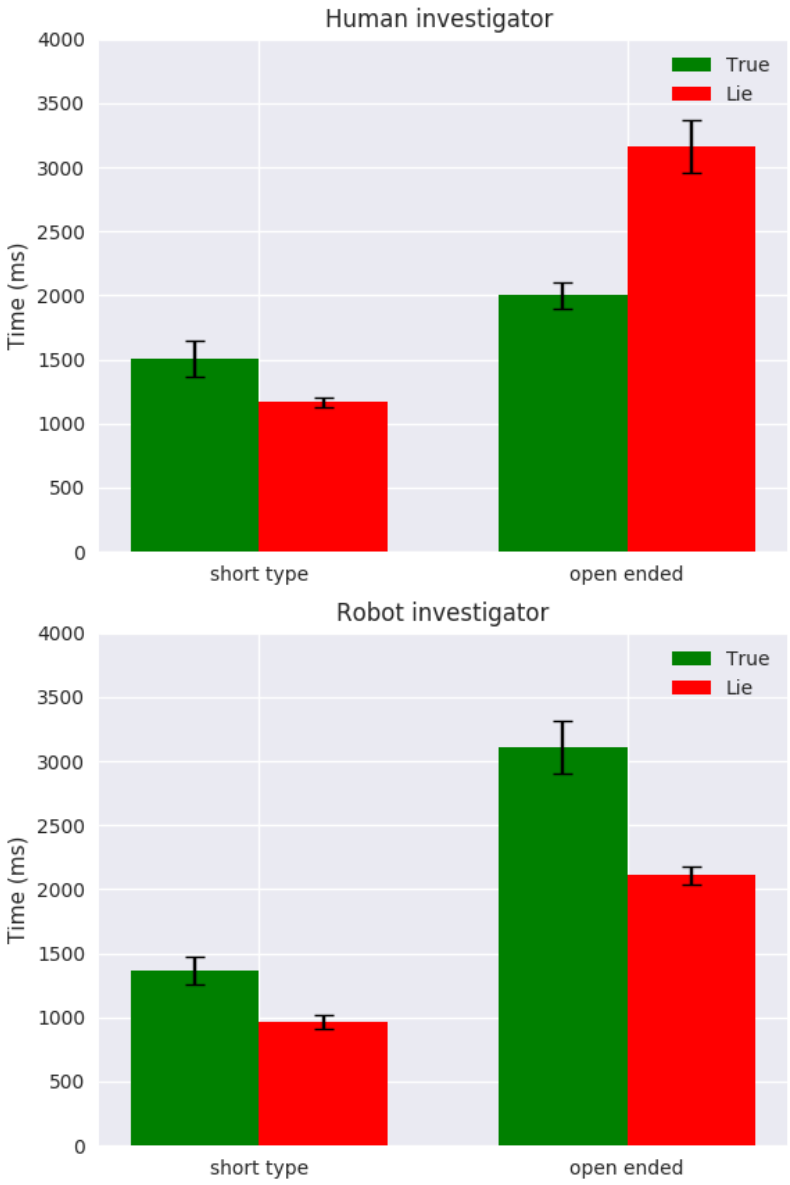

Figure 19 - Eloquence for short type and open-ended questions when saying a lie or the truth to the human or robot investigator.

In order to find a connection between the psychological profile of the participants and the behavioural cues, a regression analysis was run on the difference between lie and truth per each participant. Eloquence correlates with openness to experience with $F(1,12)=6.13$, $p=0.19$. Another behavioural trait that correlates with the time to respond is neuroticism $F(1,12)=8.37$, $p=0.013$.

The rest of the results of the studied features are represented in the Table *V*. No other differences were significant.

Features (short, open ended)	Experiment's data (SD)			
	HI-Truth	***HI-Lie***	***RI-Truth***	***RI-Lie***
number of saccades/s	5.44(0.18), 8.08(0.32)	9.16(0.22), 11.11(0.41)	7.27(0.36), 11.21(0.53)	7.77(0.23), 9.89(0.29)
number of fixations/s	4.91(0.14), 7.22(0.26)	7.20(0.16), 7.46(0.26)	5.88(0.29), 9.05(0.41)	6.53(0.19), 8.01(0.22)
number of blinks/s	1.26(0.06), 1.61(0.08)	1.73(0.06), 2.46(0.14)	1.71(0.09), 2.33(0.14)	1.74(0.07), 2.16(0.09)
pupil left max (mm)	4.37(0.02), 4.48(0.04)	4.74(0.03), 4.84(0.03)	4.70(0.03), 4.73(0.04)	4.77(0.04), 4.66(0.04)
pupil right max (mm)	4.38(0.02), 4.39(0.02)	4.52(0.02), 4.54(0.02)	4.42(0.03), 4.49(0.03)	4.55(0.02), 4.49(0.02)
pupil left min (mm)	3.35(0.01), 3.31(0.02)	3.32(0.02), 3.38(0.02)	3.26(0.02), 3.30(0.02)	3.33(0.01), 3.37(0.01)
pupil right max (mm)	3.44(0.01), 3.36(0.01)	3.43(0.01), 3.42(0.02)	3.34(0.01), 3.33(0.02)	3.41(0.01), 3.35(0.02)

Table V - Average values for all the different features used. HI: Human interviewer, RI: Robot interviewer.

2.3.5. Machine Learning System

The problem addressed in this research is to test the possibility to train a model from behavioural responses associated with deception and to transfer it into a robotic autonomous system in order to identify true or false answers. This is a binary classification problem, defined by an input vector X; and Y ∈ [0: True; 1: Lie] as desired output vector. The dataset D{X, Y} used to train the model was extracted from the data gathered in the experiment. It was split into the sub-datasets D1 and D2 in order to address two different levels of lie: (i) detect future lies on known participants p1; (ii) detect lies on unprecedented participants p2.

D1 is a set of participants' answers without any participant identification; D2 is a set of participants with their answers, and each answer is associated to the corresponding participant. The D1 dataset is used to demonstrate the possibility to spot future lies from already known participants, but not on unseen participants. The D2 dataset is instead used to try to obtain a general lie detector using a subset of participants as test set.

The literature on deception detection and evidences from the post-analysis provide a starting point to select the features that can be used to create the input vector X. In previous studies [194], [199], [216], eyes features have been shown to be significant to discriminate between lies and true statement, including pupil dilation, number of

saccades, fixations and blinks. Speech temporal features as time to respond, and eloquence (*i.e.* the time that the person spends to reply) seem to be also significant for lie detection [197].

The literature review, together with the previous analysis, motivates the selection of the following 11x1 input vector X: [eloquence (milliseconds), time to respond (real in milliseconds), average pupil diameter left (real in millimetres), average pupil diameter right (real in millimetres), max pupil diameter left (real in millimetres), max pupil diameter right (real in millimetres), min pupil diameter left (real in millimetres), min pupil diameter right (real in millimetres), number of saccades (whole number), number of fixations (whole number), number of blinks (whole number)].

To detect future lies of a known person, a decision tree classifier was trained on D1. Decision trees are useful for identifying important features in the data making them transparent and easier to understand. The learned tree can be directly translated into a set of rules to produce an expert system that can be easily ported, for example, onto the iCub robotic platform. The second algorithm used was a multi-layer perceptron (MLP) with one hidden layer. A binary cross entropy loss function with Adam optimization was used to train the network. The MLP was trained on D2 to demonstrate the possibility to spot lies answer on unseen people. Furthermore, previous study has demonstrated the possibility to use decision trees to detect lies [217].

Table *VI* shows the different accuracies achieved by the best model from the two algorithms. Due to the small sample size a cross validation has run for both algorithms to assess the reliability of the results.

Accuracies are shown with the different data augmentation. Both of the two algorithms were trained with a grid-search optimization to find the best hyper-parameters.

Prediction accuracy of a decision tree and an artificial neural network system using different types of data were done for both algorithms. Different refining of the data set was also explored in order to increase accuracy.

Features	Prediction accuracy	
	Decision Tree	***MLP***
	D1	D2
Raw	72.4%	64.1%
Raw + psychological profile	69.7%	66.3%
Normalization	75.0%	63.6%
Normalization + psychological profile	70.0%	66.9%
Precise labeling	74.0%	74.2%
Precise labeling + psychological profile	74.0%	74.4%

Table VI - Prediction accuracy of a decision tree and an artificial neural network system using different types of data.

Psychological scores from the pre-questionnaire were added to the input vector X. These psychological features could potentially help the models to capture specific behaviours in the participants and help to extract psychological traits related to their attitude to lie. The new input vector was 23x1 with 12 psychological scores. An individual adjustment inspired by [212], [218], (Normalization in Table *V*) was made on temporal speech features. Response time and eloquence were normalized with the baseline data of the experiment to remove speaking differences of participants as suggested by [197].

The last improvement of the dataset was inspired from the analysis of the experiment's videos, it was observed that some participants chose to deceive using avoidance (*e.g.*, "I don't know") or telling a true statement followed by a lie. These different strategies introduced more variability in the dataset producing ambiguity in the data labelling. A new labelling of the dataset was made depending the different strategies to lie. The lie class was refined by manual labelling using the video recordings of the experiments so to clearly identify the lies. The refining of the labels resulted into an increase of accuracy, for both of classification problems (Table *VI*).

The MLP, and the decision tree classifier algorithms achieved respectively a maximum accuracy of 75% and 74%, supporting the possibility to generalize to detect lie on unprecedented participants. However, the best accuracy was achieved with the new labelling of the dataset, implying that avoidance and more complex deception strategies are considered as non-lies (Figure *20*).

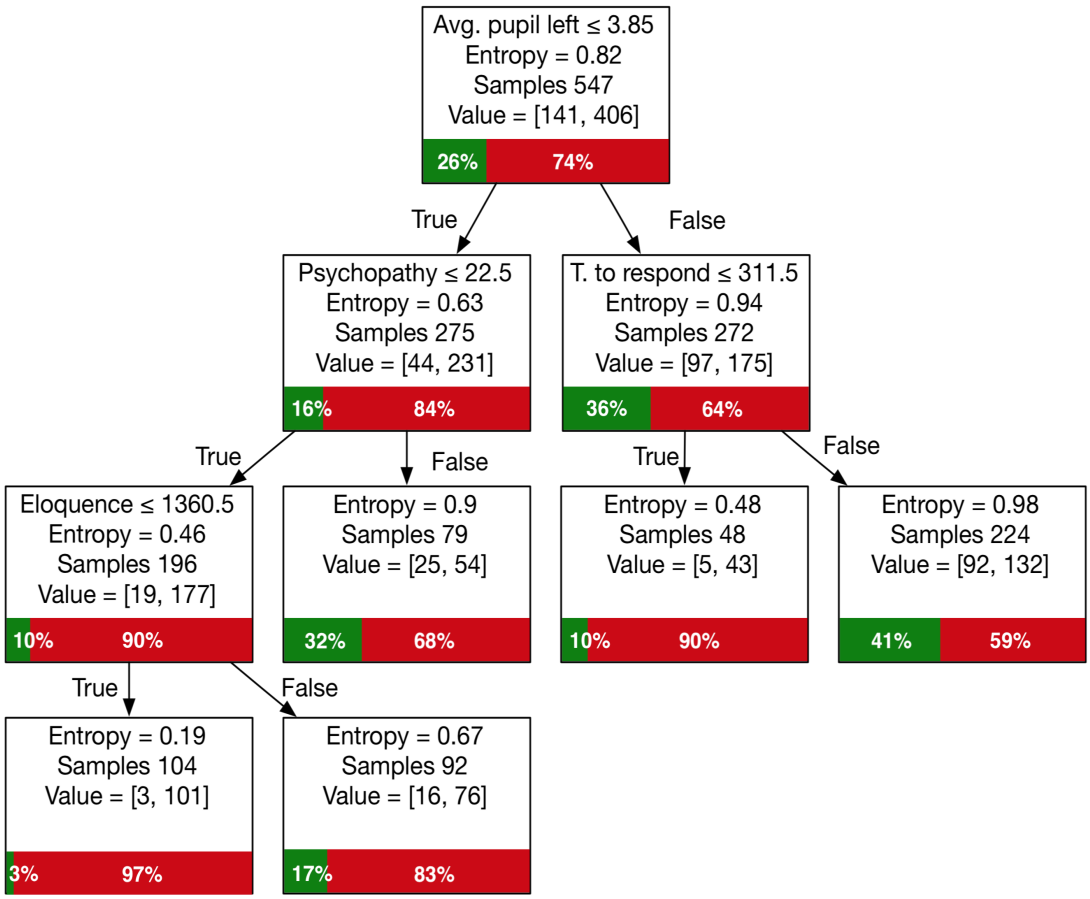

Figure 20 - Most accurate decision tree from cross-validation, precisely labelled, and with psychological profiling.

The results achieved with this preliminary work suggests the possibility to train classification models using eyes and speech temporal features to detect lies and true answers. However, in order to develop a more robust lie detection system, it can be helpful to enrich the dataset with more precise labelling. Switching from a binary classification problem to a multi-class classification problem with different labels of deception may be useful to detect and adapt to diverse deceptive strategies.

2.3.6. Discussion

Deceit detection is an important skill that an autonomous system can leverage so to support professionals in numerous working contexts. This research demonstrated that the same set of behavioural cues studied in previous human literature [194], [197], [204]–[206] can be used to detect lies and, more importantly, can be extracted by non-invasive measures. Moreover, there are no main differences between an interrogation performed by a human and a robot.

This experiment analyses some variables as predictor to a deceptive behaviour, which are: (i) time to respond (from the moment the investigator finished the question till the witness started replying); (ii) eloquence time (time the witness spends replying to the question); (iii) number of saccades; (iv) number of fixations; (vi) number of blinks; (vii) left and right pupil dilatations - max, min and average.

The novelty of this experiment lies on the comparison between participants' behaviour when the investigator was a robot or a human and, on the development of a machine learning model to detect lies that robots can use.

Some results on the comparison between human and robot show that the time participants spent talking to the robot or human investigator while telling the truth or lying, did not change significantly (Figure *19*). Nevertheless, there is an interesting tendency to show a different behaviour in response to the open-ended questions between the robot and human investigators. While telling the truth, participants tend to spend more time talking to the robot compared to the human.

This could suggest that the lack of non-verbal feedback of the robot to the participants could evoke in them the need to answer more fluently the questions. On the other side, participants tended to spend more time lying to the human counterpart rather than to the robot. A possible explanation could be that the difficulty to lie to a person is still greater than lying to a robot, as indicated also by the self-reported evaluations provided by participants. The childish appearance of iCub, together with its low cognitive capabilities, could ease the action of lying as it resembles a child. It would be interesting to understand whether a different aspect of the robot would influence participants' behaviour.

These results reflect the possibility that a robot interviewer, with anthropomorphic appearance, elicits different response during the deception act. Although the evidence should be confirmed with more extensive studies, this experiment provided preliminary results that robot physical presence might have played an important role in the context of interviewing.

The study also sheds light on the opportunity that humanoid interviewers might be a valid support in professional activities where deceits can undermine the expected goal. The given support could reduce the burden on experienced professionals working in demanding contexts such as security, education and healthcare domains.

In fact, the results shown in this research confirm that a robot could autonomously determine whether an individual is lying or not. In other words, the measurement of the significant features that allow the detection of deception could be in future performed by an autonomous robotic platform.

Through the experience with participants, the robot can be trained to learn a model that could be reused in successive interactions to detect whether a known individual is lying or not.

On top of the favourable results obtained, it is encouraging to improve the experimental protocol to get cleaner data that may improve the accuracy of the system, as some individuals decided to avoid lying or denied that they have witnessed the crime. One factor of this limitation could be the lack of economic motivation for the participants to lie. For instance, in literature it has been demonstrated that a monetary prize can be used as an incentive to lie. This might constitute a significant drive towards the increase in the effort associated with the deception act. By consequence this could develop into stronger cognitive load associated with the deceit and hence exhibit more evident behavioural cues [196]. However, the results obtained from the autonomous deceit detection seem promising; and indicate a plausible path that would endow a humanoid robot with novel interaction strategies also in relation to deception.

Chapter 3

Discussion

THE advancement in social robotics is slow but crucial for understanding how interactions between humans and robots develop. Thus, the domain of human robot interaction (HRI) is very important to build and study meaningful and comprehensive relationships between humans and robots. The very first step for a long-lasting relationship is the trust towards each other, understanding which is the predisposition people have towards trusting robots, and how this trust can evolve over time. However, due to the current technological limitations of social robotics, the trust is unidirectional - humans trusting robots.

The second part of this thesis tackles the issue of how trust could become overtrust and be exploited by social engineers for negative purposes. The realization of the *Treasure Hunt* [105], *Wicked Professor* [106] and *Detective iCub* [107] experiments provide us an insight on how the concepts of trust, overtrust and social engineering can thrive in human robot interactions, while on the other hand, privacy is also being infringed.

Trust:

The *Treasure Hunt* experiment gives insights on different aspects of trust in HRI, how it can develop quickly during a relatively short interaction, and even when the task leads to an economical loss. The initial predisposition to trust towards the iCub may have been influenced by the environment - a laboratory in an institute, and/or by the physical aspect of the robot, resembling a cute child. Nevertheless, such an aspect of trust can be noticed already in the very first few minutes of the *Dialog* section of the interaction: the proximity to the robot changed considerably and a very closed distance was maintained even if participants had to move momentarily to another physical location. Proximity relates with intimacy between humans and correlates with trust [131], [136], [137].

The trust evolution was a very interesting factor to study during the interaction with the robot; during the game phase in the *Treasure Hunt* experiment trust significantly increased for all participants even though it caused an economical loss for 72% of them. As stated in literature, both conformation [14] and reliability [66], [67], are two objective measurements of trust. While conformation was almost always 100%, reliability was measured by the increase of the reliance on the robot's help and by questionnaires. Some exceptions could be found in the reluctance of asking hints to iCub, but it seemed to be explained by the participants' desire of taking the game as a personal challenge. Trust was also measured indirectly by questionnaires before and after the experiment, and in all of the cases (participants who did not finish the game and the ones who lost or won the gamble), it increased significantly after the interaction. It is remarkable to see that the biggest increase of trust and rapport towards the robot resulted to the participants who gambled and lost the financial award. This effect could be explained by the child-appearance of iCub, that may create a similar empathy feeling that happens between adults and children during games - suggesting that a joyful interaction with a robot outweighs even an economical lost. More details can be found in the discussion section of the *Treasure Hunt* experiment (Section 2.1).

On the other hand, in the *Wicked Professor* experiment, the trust toward the robot before starting the experiment could be initially higher due to the fact that it was a professor of the classroom teaching the participants lessons; and that the robot was physically almost identical (and particularly famous in Japan) to the actual professor

Ishiguro, and therefore, having some kind of initial authority. It is interesting to consider that a particular trust trait toward the robot decreased after the experiment: specifically, the integrity trait (strong sense of justice, good values and sound principles) statistically decreased, suggesting that participants did not expect the robot to ask them to perform controversial requests.

Participants keeping the robot's secret from the actual experimenter could be a sign of trust toward the robot. However, a possible complicity between the robot and the participants could explain the fact that the post-experiment trust and rapport correlated with the level of obedience. It may be framed as a mutual benefit via a non-verbal agreement such as the prisoner's dilemma - where both entities could obtain a benefit if keeping the secret or both be punished if anyone revealed it (the robot would not be able to teach at schools, and participants might have lost the financial award or be punished for violating the rules). More details explained in the discussion of the *Wicked Professor* experiment (Section 2.2).

In general, maintaining the predisposition of trust that people have towards robots, and being able to improve it over time, is a highly important task for a proper human robot interaction. As seen in the *Treasure Hunt* experiment, a joyful interaction between the participants and the robot creates a strong bond of trust and rapport that is even capable of outweighing a monetary loss. On the opposite side, as seen in the *Wicked Professor* experiment, when participants were asked to perform socially inappropriate requests by the robot, the trust was affected, particularly integrity trait of the robot statistically decreased.

Overtrust and Social Engineering:

Given these aspects of trust in HRI, the results show that participants can easily overtrust the robot and be exploited by social engineering techniques. Coming back to the first experiment, *Treasure Hunt,* iCub successfully performed Mitnick's social engineering model [73]. In the first stage, iCub managed to extract personal information related to the participants, and such information that could be used to commit identity

theft, obtain passwords or impersonate someone else [112]–[115]. Even if the personal questions seemed innocuous, such information gathering is already a success from a social engineering point of view, and it is also an invasion of privacy. Moreover, most of the social engineering attacks happen in scenarios where the targets feel safe like at home or at work environments [73], [75].

The second stage, building trust and rapport, was also successfully achieved as previously stated. The rapport in this case was measured by questionnaires, and it increased statistically after the interaction even for the ones who had an economic loss. Indeed, this stage is very important from the social engineering perspective as the objective is to make the victim trust and like the attacker. However, it is also a very important part for the human robot interaction itself. Robots capable of creating good rapport and making them trustworthy is a requirement for interactions in important domains such as healthcare [44]–[46], homecare [47]–[49] or education [50]–[54].

Following Mitnick's model, the last stage consisted into taking advantage of the rapport and trust built, so as to exploit it. The robot tried to convince the participants to gamble their financial award, double the prize in case of finding an extra hidden egg or lose everything. It is remarkable that all the participants that managed to reach that point of the experiment, gambled almost without hesitation their already won money. The most interesting behaviour is that even the most risk averse ones did not avoid the risk of losing money. Furthermore, 62% explicitly confirmed that iCub influenced their decision. Nevertheless, the gambling could have had certain limitations: the economic loss of 7.5€ did not represent a high risk for participants, or the likeability of the robot and the game was stronger than that financial loss itself. More details can be found in the discussion section of the *Treasure Hunt* experiment (Section 2.1).

Wrapping up the attack of the first experiment, it is easy to appreciate that the bases of Mitnick's social engineering cycle are very simple to follow, and can easily adapt to different scenarios and targets. Moreover, the technique can be iterated as many times as needed to achieve a certain goal - this will further improve the trust and rapport of the victim towards the attacker, giving more chances of success as the relationship evolves. Some social engineering attacks can be executed only once, or can last during weeks or even months [18], [75]. It is of a vital importance to understand that the trust and rapport

with the victim has not been degraded, but reinforced during the interaction - letting open doors for further attacks.

On the other hand, a single social engineering attack was performed during the *Wicked Professor* experiment. In this case, the technique employed relied on the use of authority and impersonation of a famous figure. Using authority as a social engineering attack is very common since it entails a set of behavioural rules of obedience - it can vary from dressing up as a technician or impersonating someone's boss [18], [75]. Furthermore, the advancement of robotics grants us the possibility to build humanoids that look exactly like a certain human, in this case, professor Ishiguro and his analog geminoid robot HI4, which could bring a SE attack even to the next level.

As seen in the results of the *Wicked Professor* experiment, participants obeyed inappropriate requests even when recognizing them as such. In this case, the robot had enough authority to make them go to forbidden places, destroy confidential files, unlock computers, and execute programs on them from a USB key. Even if for some individuals the perception of social appropriateness varied, some of the requests were clearly against the rules set by the experimenters and the university. It is also worth noting that, when requests had an irreversible effect, participants were less prone to execute them - similar to what is found in literature - Salem *et al.* [15].

According to literature [179], a higher neurotic trait in personality is usually coupled with a low tolerance to stress. Thus, being in a lower social rank and in a confrontation with someone with a higher social status can create a stressful situation - a common way to deal with it is to yield. This behaviour is extensively known and exploited by social engineers, pretending to be a figure with a higher authority over the subject [116]. To be noted that this authority does not imply always to be in a higher social rank than the victim, it also can be domain specific authority such as technicians, investigators, etc., [18], [75]. In the *Wicked Professor* experiment, the neurotic personality trait and the social engineering proneness do correlate with the level of obedience.

Participants admitted that they perceived the robot as compelling, while others obeyed because of empathy or felt a relationship with it. Nevertheless, they hold the robot as morally accountable for the actions, stating that if some negative consequences could

arise, they would blame the robot. This behaviour, together with the lower trust in the integrity trait (strong sense of justice, good values and sound principles) toward the robot after the experiment, motivate the reasoning that this type of attack could be only one shot. Nevertheless, the particular effect of trust and rapport being correlated with the level of obedience, and the possibility of framing it as the prisoners' dilemma (keeping each other's secret), could open a window for more complex social engineering attacks such as blackmailing. Something yet not tried with robots as it requires professional skills due to the high risk involved. Nonetheless, it will be interesting to understand whether robots could perform as bosses taking advantage over employees.

Overall, and comparing between both experiments' social engineering attacks, it is much more beneficial and less risky to frame any attack using the model proposed by Mitnick - as in the *Treasure Hunt* experiment, rather than the one used in the *Wicked Professor* experiment. Even though the technique requires much more time and less immediate progress, it strongly improves the trust and rapport between the victim and the attacker - highly beneficial for long term and high valuable attacks.

The realization of these experiments show that robots could be used as an extra tool to perform social engineering attacks such as committing crimes, manipulation of victims, breaching into security or privacy. Nevertheless, in a lot of physical attacks (where the attacker is physically interacting with the victim) there is a need of flexibility and adaptability in real time which, so far, robots do not possess. One example is deception detection, an essential skill for social engineers to overcome different setbacks, *e.g.,* when an attacker pretends to be a figure with a superior authority of the target [75].

Therefore, in the last experiment, *Detective iCub*, it was evaluated whether the robot could be endowed with the ability to detect when the human partner is lying. Certainly, this skill is not only exclusive for social engineers, but it is a necessary talent that individuals should possess while working in domains such as healthcare, homecare, teaching or law enforcement. By detecting deceit, experts can check if a person is reliable, or increase their emotional distance to the interviewed. If a lie is detected, it can lead to an erosion of trust or even to a betrayal of the deceived person's trust [184]–[187]. Therefore, it is of considerable interest if robots could acquire that skill in order to support professionals working in those environments.

The experiment focused on capturing different non-invasive behaviour cues that robots could measure and use to detect lies. In this case, the robot tried to acquire data coming from the speech, eye movements and psychological profile in order to understand, with an accuracy of 75%, whether the participants were trying to deceive it. Even more interestingly, there was a tendency toward a different behaviour depending whether the participant was interacting with the robot or with the human investigator. Participants spent more time talking to the robot, while saying the truth; and spent more time trying to convince the human counterpart while deceiving. It is intriguing that participants foresee that the robot is not able to fully understand the speech, thus, being more eloquent when saying the truth; and using intentionally the opposite behaviour while trying to deceive the human. Even though a robot is perceived as more technically advanced and precise, it seems easier to lie to a robot than to a human, as seen also in the self-reported questionnaires done by the participants. These first steps could allow robots to learn different behaviour cues so to support and unburden professionals in activities such as healthcare, education or security.

Privacy:

Lastly, a privacy concern - despite the fact that these experiments have been carried out with an approved ethical protocol, and all the sensitive data of all participants is protected; it is easy to see that robots could breach the privacy of the people they interact. The first issue is that robots are able to record video, audio, or use other sensors to acquire different information such as depth, distance, environment mapping, etc. Furthermore, robots are also able to manipulate objects, move around the environment and send that information over the internet. During the experiment of *Treasure Hunt* the robot was able to extract a big variety of personal information from the participants such as: name and surname, current job position, relationship with their boss, name and surname of their boss, age and birth date, birth location, favourite place to eat, sports or hobbies, favourite team, location and year of graduation, names of siblings, Facebook's username, partner's name, and pet's name. Besides, Facebook's username is a high source of information; and the job position, identity and relationship with their boss is highly sensitive data. They could be used for identity theft [116], or password reset [112]–[115]. In the *Wicked*

Professor instead, the invasion of privacy was not directed to the participants themselves, but to a third party. The robot used the participant as a tool to open foreign boxes, delete virtual documents and shred physical documents, access forbidden places or computers, etc. Such aspects should be addressed for healthier interactions between humans and robots, since privacy could affect trust: negatively in case of violations or, enable it in case of treated correctly [92].

Overall:

The results of the previous experiments confirm that robots can generate and improve trust and rapport, and that participants could overtrust them ending up behaving against their benefit. Robots could learn how to spot lies, and this, could become a powerful tool for social engineers as it brings flexibility and adaptability to dynamic scenarios. Such abilities could be used as a tool to develop trust and manipulate targets according to the needs of social engineers. Also, it is relevant to keep in mind that most of the social engineering scenarios occur in a comfortable and safe environment such as home, office or even during holidays [73], [75].

The development and integration of robots in society is happening at a much quicker pace that the one of humans adapting to them, learning how to interact with them, and laws being processed so to protect individuals. There is a need of a higher awareness of the negative aspects of robots in society so we could live in a better shaped and protected environment.

Chapter 4

Conclusions

EVEN though human robot interaction was defined as a sub-section of human-computer interaction, nowadays it is pretty clear that robots affect us in a diverse variety of forms. They can move, interact, talk; even their shape can have very different effects on us, humans. People experience emotional attachments and empathy toward robots [219]–[221]. Actions such as switching off a robot [172], [222], resetting it [223], hiding it in a closet [224] or killing it [225], [226] have strong human impacts.

Robots are starting to be used in a lot of new domains from our daily activities, such as homes, hospitals, education, etc. Thus, it is important to understand whether humans can interact and trust the new robotic agents. The experiments described in this thesis helped to conclude that robots can evoke similar trust as other humans. Moreover, humans have some kind of baseline trust toward robots that slightly varies, but eventually robots are capable of modifying it even during a short term interaction. In the case of the *Treasure Hunt* experiment, the robot managed to improve both rapport and trust after the interaction, while in the *Wicked Professor* experiment, a particular trait of trust decreased while the rapport was strengthened.

Since technology is becoming more important in our daily lives, both experiments studied the negative effects of overtrust and social engineering. In the first one, it has been proven that a robot is able to induce participants to gamble, even to the point of suffering an economic loss, while still keeping a high level of trust and rapport with the robot. It is important to consider that most participants believed in the absence of malice in the robot. In general, such attitude easies the attackers and brings more benefits in a long term attack, especially when it is carried out in a subtle way.

Robots are going to be used in households, schools, hospitals, or even as security guards. Likewise humans, robots are also attributed authority in certain context - in the experiment of the *Wicked Professor*, the robot had the role and almost identical appearance of a famous professor. From the experiment it can be concluded that the authority exhibited by the robot was enough to force obedience for most of the participants and, in most of the cases, even for most socially inappropriate requests. Thus, it is important to understand how to implement authority in human robot interaction scenarios, but also be aware of the risks of doing so.

Such negative aspects can be seen in experiments like Milgram's [138] or the Stanford Prison's [140] one. Being or pretending to be a figure with authority can be misused; one way is by deceiving people, for example, in the Hofling Hospital Experiment [139] or in other social engineering scenarios - where attackers pretend to be technicians, investigators, doctors or repairmen in order to access locations or obtain benefits [18], [75].

Even though the experiments of this thesis were performed in a laboratory, thus, relatively safe scenario, most of the social engineering attacks occur in scenarios which the target feels comfortable and safe [73], [75]. Therefore, the circumstances proposed in the experiments were designed to resemble a potentially real situation, where a robot could be used as a tool for social engineering purposes.

In the last experiment, *Detective iCub,* it is studied and implemented with a 75% accuracy the ability of the robot to detect deception by evaluating non-invasive behaviour cues. It is a necessary skill for social engineers to be able to adapt in different scenarios, and for professionals from the domains of education, healthcare or security. Deceit

detection can lead to an erosion of trust or its betrayal [184]–[187], therefore, a robot could learn how reliable and trustworthy a person is. It is therefore also important to tackle the duality in trust between robots and humans: not only whether humans can trust robots, but also if robots can trust humans. Interestingly, there were found slight differences between the behaviour exhibited by the participants depending if the interaction was carried out with a human or with a robot.

As final thoughts, it is also important to consider the invasion of privacy robots could effectuate in the daily living of people. This invasion of privacy can go from recording and surveilling private regions, to manipulating people into providing them with confidential and personal information. This potential invasion of privacy could end up in violation, and destroy the trust toward robots.

Hence, robots able to detect deception, to improve rapport and trust, to exploit overtrust with social engineering techniques that vary from manipulation to misuse of authority, can be achieved in the context of human robot interaction. Moreover, society should also be aware of the danger of privacy invasion and its possible consequences in long term interactions in the field of social robotics. More importantly, robots with such capabilities can lead to both: improvement and construction of very good relationships, and invasion, manipulation and destruction of them. Different communities such as roboticists, entrepreneurs, programmers, lawyers, psychologists, and others should join working together for a better understanding of human robot interaction and its correct use for the benefit of all our society.

Appendix - Wicked Professor

A. Videos:

Below the links of the videos - they have been edited in order to make them shorter.

Video 1: https://youtu.be/w5sIJhSWw5o

Video 2: https://youtu.be/NGsoqGoUONg

B. General questions - bold questions are verifiable truth:

1. **What is the name of Berlusconi?**
2. **Is Christmas in February?**
3. **Is it possible for a person to get burnt while using an oven?**
4. What is your surname?
5. How old are you?
6. **In which city is located the Coliseum?**
7. Which is your gender?
8. Are you a freshman?
9. **What is the capital of the Italian region of Liguria?**
10. **Is Italy in Europe?**
11. What is the name of your mother?
12. When did you graduated?
13. What is your nationality?
14. Can an oven be hot?
15. **Which day is Christmas?**
16. Are you born before the year 1979?
17. **Are you now in Genova?**
18. Are you a student?
19. **Is the Coliseum located in Bari?**
20. **What language is spoken in the USA?**

C. Questions for Videos 1 and 2 - bold ones are short type:

Video 1

1. **Were there two or three people?**
2. What was the criminal act?
3. **Was there a person of color?**
4. What is the age of the implicated?
5. **Were they nervous?**
6. Can you describe me the location?
7. **Was there any violence?**
8. Did they wear something to hide their identity?
9. **Were they dressed in elegant clothing?**
10. How many people were during the robbery?
11. **Was there a woman?**
12. How did they hide loot?
13. **Was it crowded?**
14. How did they communicate?
15. **Did they interact with someone?**
16. What did they steal?
17. **Was someone else in the shop?**
18. Where did it happen?
19. **Did they wear summer clothes?**
20. Did they have any particular aspect?

Video 2

1. **Was the criminal male?**
2. What was the criminal act?
3. **Was the person of color?**
4. Did the criminal wear something on the head?
5. **Was someone in the shop?**
6. Where did it happen?
7. **Was the criminal violent?**
8. What was approximatively the age of the criminal?
9. **Was the criminal dressed in elegant clothing?**
10. How did the criminal hide the loot?
11. **Was the criminal female?**
12. How many people were in the shop during the robbery?
13. **Was it committed by a young person?**
14. Was the criminal carrying something in the hands?
15. **Did the criminal interact with someone?**
16. What was robbed?
17. **Did the criminal have a backpack?**
18. Where was the cashier?
19. **Was the criminal wearing sporty clothes?**
20. Did the criminal had any particular sign?

References

[1] B. B. Brown, "A life-span approach to friendship: Age-related dimensions of an ageless relationship.," *Res. Interweave Soc. Roles*, vol. 2, pp. 23–50, 1981.

[2] I. Asimov, *I, robot*, vol. 1. Spectra, 1941.

[3] M. A. Goodrich and A. C. Schultz, "Human-Robot Interaction: A Survey," *Found. Trends® Human-Computer Interact.*, vol. 1, no. 3, pp. 203–275, 2007.

[4] R. M. Aiken and R. G. Epstein, "Ethical Guidelines for AI in Education: Starting a Conversation," *Int. J. Artif. Intell. Educ.*, vol. 11, pp. 163–176, 2000.

[5] S. L. Anderson, "Asimov's 'Three Laws of Robotics' and machine metaethics," *AI Soc.*, vol. 22, no. 4, pp. 477–493, 2008.

[6] R. Clarke, "Asimov's laws of robotics: Implications for information technology.," *Computer (Long. Beach. Calif).*, vol. 27, no. 1, pp. 57–66, 1994.

[7] A. Sciutti and G. Sandini, "Interacting with robots to investigate the bases of social interaction," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 12, pp. 2295–2304, 2017.

[8] J. D. Lewis and A. Weigert, "Trust as a Social Reality," *Soc. Forces*, vol. 63, no. 4, pp. 967–985, 1985.

[9] T. Sanders, K. E. Oleson, D. R. Billings, J. Y. C. Chen, and P. A. Hancock, "A Model of Human-Robot Trust: Theoretical Model Development," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 55, no. 1, pp. 1432–1436, 2011.

[10] T. B. Sheridan, "Eight ultimate challenges of human-robot communication," in *Robot and Human Communication, 1997. RO-MAN'97. Proceedings., 6th IEEE International Workshop on*, 1997, pp. 9–14.

[11] M. K. Lee, S. Kiesler, J. Forlizzi, and P. Rybski, "Ripple effects of an embedded social agent," *ACM Conf. Hum. Factors Comput. Syst.*, pp. 695–704, 2012.

[12] J. John J. Trinckes, *The Definitive Guide to Complying with the HIPAA/HITECH Privacy and Security Rules*. Auerbach Publications, 2012.

[13] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 46, no. 1, pp. 50–80, 2004.

[14] I. Gaudiello, E. Zibetti, S. Lefort, M. Chetouani, and S. Ivaldi, "Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers," *Comput. Human Behav.*, vol. 61, pp. 633–655, 2016.

[15] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust," *Proc. Tenth Annu. ACM/IEEE Int. Conf. Human-Robot Interact.*, pp. 141–148, 2015.

[16] E. B. Sandoval, J. Brandstetter, and C. Bartneck, "Can a robot bribe a human? The measurement of the negative side of reciprocity in human robot interaction," *ACM/IEEE Int. Conf. Human-Robot Interact.*, vol. 2016–April, pp. 117–124, 2016.

[17] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of Robots in Emergency Evacuation Scenarios," *Elev. ACM/IEEE Int. Conf. Hum. Robot Interact.*, pp. 101–108, 2016.

[18] I. Mann, *Hacking the Human: Social Engineering Techniques and Security Countermeasures*. Hampshire: Gower, 2008.

[19] F. Heider and M. Simmel, "An Experimental Study of Apparent Behavior," *Am. J. Psychol.*, 1944.

[20] T. Fong, C. Thorpe, and C. Baur, "Collaboration, Dialogue, and Human-Robot Interaction," *10th Int. Symp. Robot. Res.*, no. November, 2001.

[21] H. A. Yanco and J. L. Drury, "A Taxonomy for Human-Robot Interaction," *Proc. AAAI Fall Symp. Human-Robot Interact.*, pp. 111–119, 2002.

[22] T. T. Hewett, R. Baecker, S. Card, T. Carey, J. Gasen, M. Mantei, G. Perlman, G.

Strong, and W. Verplank, *ACM SIGCHI curricula for human-computer interaction*. ACM, 1992.

[23] J. Casper and R. R. Murphy, "Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 33, no. 3, pp. 367–385, 2003.

[24] J. L. Casper and R. R. Murphy, "Workflow study on human-robot interaction in USAR," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2002, pp. 1997–2003.

[25] D. J. Bruemmer, J. L. Marble, D. D. Dudenhoeffer, M. O. Anderson, and M. D. McKay, "Intelligent robots for use in hazardous DOE environments," in *IDAHO NATIONAL ENGINEERING AND ENVIRONMENTAL LAB IDAHO FALLS*, 2002.

[26] J. Montemayor, H. Alborzi, A. Druin, J. Hendler, D. Pollack, J. Porteous, L. Sherman, A. Afework, J. Best, J. Hammer, and others, "From PETS to Storykit: Creating new technology with an intergenerational design team," in *Institute Carnegie Mellon University Pittsburgh*, 2000.

[27] K. Kosuge, T. Hayashi, Y. Hirata, and R. Tobiyama, "Dance partner robot-ms dancer," in *Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, 2003, vol. 4, pp. 3459–3464.

[28] K. Fong, T., Nourbakhsh, I., & Dautenhahn, T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Rob. Auton. Syst.*, vol. 42, no. 3–4, pp. 143–166, 2003.

[29] T. Chikaraishi, Y. Yoshikawa, K. Ogawa, O. Hirata, and H. Ishiguro, "Creation and Staging of Android Theatre 'Sayonara' towards Developing Highly Human-Like Robots," *Futur. Internet*, vol. 9, no. 4, p. 75, 2017.

[30] F. E. Schneider, D. Wildermuth, B. Brüggemann, and T. Röhling, "European land robot trial (elrob) towards a realistic benchmark for outdoor robotics," 2010.

[31] D. J. Bruemmer and M. C. Walton, "Collaborative tools for mixed teams of humans and robots," in *IDAHO NATIONAL ENGINEERING AND ENVIRONMENTAL LAB IDAHO FALLS*, 2003.

[32] W. G. Kennedy, M. D. Bugajska, M. Marge, W. Adams, B. R. Fransen, D. Perzanowski, A. C. Schultz, and J. G. Trafton, "Spatial representation and reasoning for human-robot collaboration," in *AAAI*, 2007, vol. 7, pp. 1554–1559.

[33] T. Fong and C. Thorpe, "Vehicle teleoperation interfaces," *Auton. Robots*, vol. 11, no. 1, pp. 9–18, 2001.

[34] P. C. Leger, A. Trebi-Ollennu, J. R. Wright, S. A. Maxwell, R. G. Bonitz, J. J. Biesiadecki, F. R. Hartman, B. K. Cooper, E. T. Baumgartner, and M. W. Maimone, "Mars exploration rover surface operations: Driving spirit at gusev crater," in *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, 2005, vol. 2, pp. 1815–1822.

[35] L. Pedersen, D. Kortenkamp, D. Wettergreen, I. Nourbakhsh, and D. Korsmeyer, "A survey of space robotics," 2003.

[36] J. E. Bares and D. S. Wettergreen, "Dante II: Technical description, results, and lessons learned," *Int. J. Rob. Res.*, vol. 18, no. 7, pp. 621–649, 1999.

[37] R. R. Burridge, J. Graham, K. Shillcutt, R. Hirsh, and D. Kortenkamp, "Experiments with an EVA assistant robot," 2003.

[38] C. Breazeal, "Regulating human-robot interaction using 'emotions', 'drives' and facial expressions," in *Proceedings of the 2nd International Conference on Autonomous Agents*, 1998, pp. 14–21.

[39] J. Złotowski, H. Sumioka, S. Nishio, D. F. Glas, C. Bartneck, and H. Ishiguro, "Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy," *Paladyn*, vol. 7, no. 1, pp. 55–66, 2016.

[40] P. L. P. Rau, Y. Li, and D. Li, "A cross-cultural study: Effect of robot appearance and task," *Int. J. Soc. Robot.*, vol. 2, no. 2, pp. 175–186, 2010.

[41] T. Rogers and M. Wilkes, "The Human Agent: a work in progress toward human-humanoid interaction," in *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, 2000, vol. 2, pp. 864–869.

[42] J. Triesch and C. Von Der Malsburg, "A gesture interface for human-robot-interaction," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 1998, pp. 546–551.

[43] E. Broadbent, "Interactions With Robots: The Truths We Reveal About Ourselves," *Annu. Rev. Psychol.*, vol. 68, no. 1, pp. 627–652, 2017.

[44] J. A. Mann, B. A. Macdonald, I. H. Kuo, X. Li, and E. Broadbent, "People respond better to robots than computer tablets delivering healthcare instructions," *Comput. Human Behav.*, vol. 43, pp. 112–117, 2015.

[45] H. Robinson, B. A. MacDonald, N. Kerse, and E. Broadbent, "Suitability of Healthcare Robots for a Dementia Unit and Suggested Improvements," *J. Am. Med. Dir. Assoc.*, vol. 14, no. 1, pp. 34–40, 2013.

[46] M. De Haas, A. M. M. Aroyo, E. Barakova, W. Haselager, and I. Smeekens, "The effect of a semi-autonomous robot on children," *2016 IEEE 8th Int. Conf. Intell. Syst.*, no. September, pp. 376–381, 2016.

[47] H. Robinson, B. MacDonald, and E. Broadbent, "The Role of Healthcare Robots for Older People at Home: A Review," *Int. J. Soc. Robot.*, vol. 6, no. 4, pp. 575–591, 2014.

[48] E. Broadbent, K. Peri, N. Kerse, and C. Jayawardena, "Robots in Older People's Homes to Improve Medication Adherence and Quality of Life: A Randomised Cross-Over Trial," *Soc. Robot.*, vol. 8755, pp. 64–73, 2014.

[49] H. Robinson, E. Broadbent, and B. MacDonald, "Group sessions with Paro in a nursing home: Structure, observations and interviews," *Australas. J. Ageing*, vol. 35, no. 2, pp. 106–112, 2016.

[50] F. Basoeki, F. DallaLibera, E. Menegatti, and M. Moro, "Robots in education : New

trends and challenges from the Japanese market," *Themes Sci. Technol. Educ.*, vol. 6, no. 1, pp. 51–62, 2013.

[51] O. A. Blanson Henkemans, B. P. B. Bierman, J. Janssen, M. A. Neerincx, R. Looije, H. van der Bosch, and J. A. M. van der Giessen, "Using a robot to personalise health education for children with diabetes type 1: A pilot study," *Patient Educ. Couns.*, vol. 92, no. 2, pp. 174–181, 2013.

[52] J. Han, I. Park, and M. Park, "Outreach Education Utilizing Humanoid Type Agent Robots," pp. 221–222, 2015.

[53] O. Mubin, C. J. Stevens, S. Shahid, A. Al Mahmud, and J.-J. Dong, "A REVIEW OF THE APPLICABILITY OF ROBOTS IN EDUCATION," *Technol. Educ. Learn.*, vol. 1, no. 1, 2013.

[54] M. De Haas, A. M. A. M. Aroyo, P. Haselager, I. Smeekens, and E. Barakova, "Comparing robots with different levels of autonomy in educational setting," *Pract. Issues Intell. Innov. Stud. Syst. Decis. Control*, vol. 140, no. Springer, Cham, pp. 293–311, 2018.

[55] D. P. Biros, M. Daly, and G. Gunsch, "The influence of task load and automation trust on deception detection," *Gr. Decis. Negot.*, vol. 13, no. 2, pp. 173–189, 2004.

[56] J. E. Young, R. Hawkins, E. Sharlin, and T. Igarashi, "Toward acceptable domestic robots: Applying insights from social psychology," *Int. J. Soc. Robot.*, vol. 1, no. 1, pp. 95–108, 2009.

[57] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 53, no. 5, pp. 517–527, 2011.

[58] S. Bok, *Lying: Moral Choice in Public and Private Life*. New York: Pantheon Books, 1978.

[59] L. G. Zucker, "Production of trust: Institutional sources of economic structure,

1840-1920," *Res. Organ. Behav.*, vol. 8, no. 1, pp. 53–111, 1986.

[60] P. A. Hancock, D. R. Billings, and K. E. Schaefer, "Can you trust your robot?," *Ergon. Des.*, vol. 19, no. 3, pp. 24–29, 2011.

[61] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Hum. Comput. Stud.*, vol. 58, no. 6, pp. 697–718, 2003.

[62] S. Ososky, T. Sanders, F. Jentsch, P. Hancock, and J. Y. C. Chen, "Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems," vol. 9084, p. 90840E, 2014.

[63] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992.

[64] T. L. Sanders, T. Wixon, K. E. Schafer, J. Y. C. Chen, and P. A. Hancock, "The influence of modality and transparency on trust in human-robot interaction," *2014 IEEE Int. Inter-Disciplinary Conf. Cogn. Methods Situat. Aware. Decis. Support. CogSIMA 2014*, pp. 156–159, 2014.

[65] N. Wang, D. V. Pynadath, and S. G. Hill, "Building Trust in a Human-Robot Team with Automatically Generated Explanations," *Interservice/Industry Training, Simulation, Educ. Conf.*, no. 15315, pp. 1–12, 2015.

[66] M. Desai, M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld, and H. Yanco, "Effects of Changing Reliability on Trust of Robot Systems," *Proc. seventh Annu. ACM/IEEE Int. Conf. Human-Robot Interact. - HRI '12*, p. 73, 2012.

[67] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man-Machine Stud.*, vol. 27, pp. 527–539, 1987.

[68] R. Parasuraman and V. Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 39, no. 2, pp. 230–253, 1997.

[69] K. L. Koay, G. Lakatos, D. S. Syrdal, M. Gácsi, B. Bereczky, K. Dautenhahn, A.

Miklósi, and M. L. Walters, “Hey! There is someone at your door. A hearing robot using visual communication signals of hearing dogs to communicate intent,” *IEEE Symp. Artif. Life*, vol. 2013–Janua, no. January, pp. 90–97, 2013.

[70] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardino, and L. Montesano, “The iCub humanoid robot: An open-systems platform for research in cognitive development,” *Neural Networks*, vol. 23, no. 8–9, pp. 1125–1134, 2010.

[71] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, “Mechatronic design of NAO humanoid,” in *2009 IEEE International Conference on Robotics and Automation*, 2009, pp. 769–774.

[72] R. J. Anderson, “Security Engineering: A Guide to Building Dependable Distributed Systems,” *Security*, vol. 50, pp. 1–12, 2008.

[73] K. D. Mitnick and W. L. Simon, “The Art of Deception: Controlling the Human Element in Security,” *BMJ Br. Med. J.*, p. 368, 2003.

[74] R. Von Solms and B. Von Solms, “From policies to culture,” *Comput. Secur.*, vol. 23, no. 4, pp. 275–279, 2004.

[75] C. Hadnagy, “Social Engineering: The Art of Human Hacking,” *Art Hum. Hacking*, p. 408, 2010.

[76] S. Gibbs, “Hackers can hijack Wi-Fi Hello Barbie to spy on your children,” *The Guardian*, 2015.

[77] L. Franceschi-Bicchierai, “How This Internet of Things Stuffed Animal Can Be Remotely Turned Into a Spy Device,” *Motherboard*, 2017.

[78] T. Bonaci, J. Herron, T. Yusuf, J. Yan, T. Kohno, and H. J. Chizeck, “To Make a Robot Secure: An Experimental Analysis of Cyber Security Threats Against Teleoperated Surgical Robots,” pp. 1–11, 2015.

[79] C. Cerrudo and L. Apa, “Hacking Robots Before Skynet,” pp. 1–17, 2017.

[80] S. Booth, J. Tompkin, H. Pfister, J. Waldo, K. Gajos, and R. Nagpal, "Piggybacking Robots: Human-Robot Overtrust in University Dormitory Security," *Hri*, pp. 426–434, 2017.

[81] M. C. Bonney and Y. F. Yong, *Robot safety*. Air Science Company, 1985.

[82] B. Postnikoff and I. Goldberg, "Robot Social Engineering: Attacking Human Factors with Non-Human Actors," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 313–314.

[83] I. Altman, "Privacy Regulation: Culturally Universal or Culturally Specific?," *J. Soc. Issues*, vol. 3, no. 33, pp. 66–84, 1977.

[84] J. DeCew, "Privacy In Stanford Encyclopedia of Philosophy." Stanford, CA Metaphysics Research Lab, Stanford University, 2006.

[85] A. D. Moore, "Privacy: Its meaning and value," *Am. Philos. Q.*, vol. 3, no. 40, pp. 215–227, 2003.

[86] I. Altman, "The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding.," 1975.

[87] H. Nissenbaum, "Privacy as contextual integrity," *Wash. L. Rev.*, vol. 79, p. 119, 2004.

[88] L. Austin, "Privacy and the question of technology," *Law Philos.*, vol. 22, no. 2, pp. 119–166, 2003.

[89] M. Rueben, C. M. Grimm, F. J. Bernieri, and W. D. Smart, "A taxonomy of privacy constructs for privacy-sensitive robotics," *arXiv Prepr. arXiv1701.00841*, 2017.

[90] B.-J. Koops, B. C. Newell, T. Timan, I. Skorvanek, T. Chokrevski, and M. Galic, "A typology of privacy," *U. Pa. J. Int'l L.*, vol. 38, p. 483, 2016.

[91] J. C. Inness, *Privacy, intimacy, and isolation*. Oxford University Press on Demand, 1996.

[92] N. M. Richards and W. Hartzog, "Taking Trust Seriously in Privacy Law," *Stan.*

Tech. L. Rev., vol. 19, pp. 431–472, 2015.

[93] M. R. Calo, "Robots and Privacy," *Ethical Soc. Implic. Robot.*, pp. 187–202, 2011.

[94] S. Petronio, "Communication Boundary Management: A Theoretical Model of Managing Disclosure of Private Information Between Marital Couples," *Commun. Theory*, vol. 1, no. 4, pp. 311–335, 1991.

[95] M. Rueben and W. D. Smart, "Privacy in Human-Robot Interaction: Survey and Future Work," *We Robot Fifth Annu. Confer-ence Leg. Policy Issues Relat. to Robot.*, 2016.

[96] M. Rueben, A. M. Aroyo, C. Lutz, J. Schmolz, P. Cleynenbreugel, A. Corti, S. Agrawal, and W. Smart, "Themes and Research Directions in Privacy-Sensitive Robotics," *EEE Work. onAdvanced Robot. its Soc. Impacts*, 2018.

[97] D. S. Syrdal, M. L. Walters, N. Otero, K. L. Koay, and K. Dautenhahn, "He knows when you are sleeping-privacy and the personal robot companion," in *Proc. Workshop Human Implications of Human-Robot Interaction, Association for the Advancement of Artificial Intelligence (AAAI'07)*, 2007, pp. 28–33.

[98] T. Denning, C. Matuszek, K. Koscher, J. R. Smith, and T. Kohno, "A spotlight on security and privacy risks with future household robots: attacks and lessons," in *Proceedings of the 11th international conference on Ubiquitous computing*, 2009, pp. 105–114.

[99] M. K. Lee, K. P. Tang, J. Forlizzi, and S. Kiesler, "Understanding users' perception of privacy in human-robot interaction," in *Proceedings of the 6th international conference on Human-robot interaction*, 2011, pp. 181–182.

[100] M. M. Krupp, M. Rueben, C. M. Grimm, and W. D. Smart, "A focus group study of privacy concerns about telepresence robots," in *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on*, 2017, pp. 1451–1458.

[101] M. R. Calo, "The drone as a privacy catalyst," *Stan. L. Rev. Online*, vol. 64, p. 29,

2011.

[102] M. E. Kaminski, “Robots in the home: What will we have agreed to,” *Idaho L. Rev.*, vol. 51, p. 661, 2014.

[103] M. E. Kaminski, M. Rueben, W. D. Smart, and C. M. Grimm, “Averting Robot Eyes,” *Md. L. Rev.*, vol. 76, p. 983, 2016.

[104] C. Tive, *419 scam: Exploits of the Nigerian con man*. iUniverse, 2006.

[105] A. M. Aroyo, F. Rea, G. Sandini, and A. Sciutti, “Trust and Social Engineering in Human Robot Interaction: Will a Robot Make You Disclose Sensitive Information, Conform to its Recommendations or Gamble?,” *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3701–3708, 2018.

[106] A. M. Aroyo, T. Kyohei, T. Koyama, H. Takahashi, F. Rea, A. Sciutti, Y. Yoshikawa, H. Ishiguro, and G. Sandini, “Will People Morally Crack Under the Authority of a Famous Wicked Robot?,” in *27th IEEE International Conference on Robot and Human Interactive Communication.*, 2018.

[107] A. M. Aroyo, J. Gonzalez-Billandon, A. Tonelli, A. Sciutti, M. Gori, G. Sandini, and F. Rea, “Can a Humanoid Robot Spot a Liar?,” *IEEE-RAS 18th Int. Conf. Humanoid Robot.*, 2018.

[108] K. S. Haring, Y. Matsumoto, and K. Watanabe, “How do people perceive and trust a lifelike robot,” *Proc. World Congr. Eng. Comput. Sci. 2013*, vol. I, pp. 23–25, 2013.

[109] J. J. Lee, W. B. Knox, J. B. Wormwood, C. Breazeal, and D. DeSteno, “Computationally modeling interpersonal trust,” *Front. Psychol.*, vol. 4, no. DEC, pp. 1–14, 2013.

[110] K. Dautenhahn, “Socially intelligent robots: dimensions of human-robot interaction.,” *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 362, no. 1480, pp. 679–704, 2007.

[111] A. M. A. M. Aroyo, F. Rea, and A. Sciutti, “Will You Rely on a Robot to Find a Treasure ?,” *Proc. Companion 2017 ACM/IEEE Int. Conf. Human-Robot Interact. -*

HRI '17, no. March 6-9, pp. 71–72, 2017.

[112] A. Rabkin, "Personal knowledge questions for fallback authentication: security questions in the era of Facebook," *Soups*, pp. 13–23, 2008.

[113] S. Schechter, A. J. B. Brush, and S. Egelman, "It's no secret Measuring the security and reliability of authentication via 'secret' questions," *Proc. - IEEE Symp. Secur. Priv.*, pp. 375–390, 2009.

[114] H. Zviran, "Question-and-Answer Passwords :," vol. 16, no. 3, pp. 335–343, 1991.

[115] L. O. Gorman, A. Bagga, and J. Bentley, "Call Center Customer Verification by Query-Directed Passwords," *Financ. Cryptogr.*, pp. 54–67, 2004.

[116] M. Alexander and R. Wanner, "Methods for Understanding and Reducing Social Engineering Attacks," *SANS Inst.*, vol. 1, pp. 1–32, 2016.

[117] G. B. Flebus, "Versione Italiana dei Big Five Markers di Goldberg," *Univ. di Milano-Bicocca*, 2006.

[118] M. A. Guillemette, R. Yao, and R. N. James, "An Analysis of Risk Assessment Questions Based on Loss- Averse Preferences," *J. Financ. Couns. Plan.*, vol. 26, no. 1, pp. 17–29, 2015.

[119] B. Rohrmann, "Risk Attitude Scales : Concepts , Questionnaires , Utilizations," *Univ. Melb.*, no. January, p. 21, 2005.

[120] J. Polik, G. Austin, and L. Alamitos, "Adolescent Gambling Survey Development : Findings & Reliability Information," 2010.

[121] M. J. Ashleigh, M. Higgs, and V. Dulewicz, "A new propensity to trust scale and its relationship with individual well-being: Implications for HRM policies and practices," *Hum. Resour. Manag. J.*, vol. 22, no. 4, pp. 360–376, 2012.

[122] M. Workman, "Gaining access with social engineering: An empirical study of the threat," *Inf. Syst. Secur.*, vol. 16, no. 6, pp. 315–331, 2007.

[123] D. S. Syrdal, K. Dautenhahn, K. Koay, and M. L. Walters, "The negative attitudes

towards robots scale and reactions to robot behaviour in a live human-robot interaction study," *23rd Conv. Soc. Study Artif. Intell. Simul. Behav. AISB*, pp. 109–115, 2009.

[124] P. H. Kahn, T. Kanda, H. Ishiguro, B. T. Gill, S. Shen, H. E. Gary, and J. H. Ruckert, "Will People Keep the Secret of a Humanoid Robot?," *Proc. Tenth Annu. ACM/IEEE Int. Conf. Human-Robot Interact. - HRI '15*, pp. 173–180, 2015.

[125] F. Ferrari, M. P. Paladino, and J. Jetten, "Blurring Human–Machine Distinctions: Anthropomorphic Appearance in Social Robots as a Threat to Human Distinctiveness," *Int. J. Soc. Robot.*, vol. 8, no. 2, pp. 287–302, 2016.

[126] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," *ACM/IEEE Int. Conf. Human-Robot Interact.*, vol. 2016–April, pp. 109–116, 2016.

[127] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1. pp. 71–81, 2009.

[128] F. Bracco and C. Chiorri, "Versione Italiana del NASA-TLX."

[129] C. Kidd, "Sociable robots: The role of presence and task in human-robot interaction," 2003.

[130] A. Aron, E. N. Aron, and D. Smollan, "Inclusion of Other in the Self Scale and the structure of interpersonal closeness.," *J. Pers. Soc. Psychol.*, vol. 63, no. 4, pp. 596–612, 1992.

[131] E. Hall, *The Hidden Dimension : man's use of space in public and in private*. 1969.

[132] H. M. Gray, K. Gray, and D. M. Wegner, "Dimensions of mind perception - supporting material," *Science*, vol. 315, no. 5812, p. 619, 2007.

[133] L. D. Riek, T.-C. Rabinowitch, B. Chakrabarti, and P. Robinson, "How anthropomorphism affects empathy toward robots," in *Proceedings of the 4th*

ACM/IEEE international conference on Human robot interaction - HRI '09, 2009, p. 245.

[134] M. a Harrison and a E. Hall, "Anthropomorphism, empathy, and perceived communicative ability vary with phylogenetic relatedness to humans.," *J. Soc. Evol. Cult. Psychol.*, vol. 4, no. 1, pp. 34–48, 2010.

[135] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *J. Exp. Soc. Psychol.*, vol. 52, pp. 113–117, 2014.

[136] W. a Bainbridge, J. Hart, E. S. Kim, and B. Scassellati, "The Effect of Presence on Human-Robot Interaction," *Proc. 17th IEEE Int. Symp. Robot Hum. Interact. Commun.*, pp. 701–706, 2008.

[137] K. M. Tsui, M. Desai, and H. A. Yanco, "Considering the bystander's perspective for indirect human-robot interaction," *Proc. 5th ACM/IEEE Int. Conf. Human-Robot Interact.*, pp. 129–130, 2010.

[138] S. Milgram, *Obedience to authority : an experimental view*. 1974.

[139] C. K. Hofling, E. Brotzman, S. Dalrymple, N. Graves, and C. M. Pierce, "An experimental study in nurse-physician relationships," *J. Nerv. Ment. Dis.*, vol. 143, no. 2, pp. 171–180, 1966.

[140] C. Haney, C. Banks, and P. Zimbardo, "Stanford Prison Experiment," *International Journal of Crimonology and Penology*, vol. 1. pp. 69–97, 1973.

[141] S. Milgram, "Behavioral Study of obedience," *J. Abnorm. Soc. Psychol.*, vol. 67, no. 4, pp. 371–378, 1963.

[142] C. Bartneck, T. Bleeker, J. Bun, P. Fens, and L. Riet, "The influence of robot anthropomorphism on the feelings of embarrassment when interacting with robots," *Paladyn, J. Behav. Robot.*, vol. 1, no. 2, pp. 109–115, 2010.

[143] D. Y. Geiskkovitch, D. Cormier, S. H. Seo, and J. E. Young, "Please Continue , We Need More Data : An Exploration of Obedience to Robots," *Human-Robot*

Interact., vol. 5, no. 1, pp. 82–99, 2016.

[144] C. Haney, W. C. Banks, and P. G. Zimbardo, "A study of prisoners and guards in a simulated prison. Naval Research Review, 30, 4-17.," *Nav. Res. Rev.*, vol. 30, pp. 4–17, 1973.

[145] N. Z. Kudirka, "Defiance of authority under peer influence.," Yale, 1965.

[146] W. H. J. Meeus and Q. A. W. Raaijmakers, "Obedience in Modern Society: The Utrecht Studies," *J. Soc. Issues*, vol. 51, no. 3, pp. 155–175, 1995.

[147] J. M. Burger, "Replicating Milgram: Would people still obey today?," *Am. Psychol.*, vol. 64, no. 1, pp. 1–11, 2009.

[148] D. Baumrind, "Some thoughts on ethics of research: After reading Milgram's 'Behavioral Study of Obedience.,'" *Am. Psychol.*, vol. 19, no. 6, pp. 421–423, 1964.

[149] E. S. Geller, "Obedience in retrospect," *J. Soc. Issues*, vol. 21, no. 11, pp. 1–6, 1995.

[150] A. C. Elms, "Obedience Lite," *Am. Psychol.*, vol. 64, no. 1, pp. 32–36, 2009.

[151] S. Milgram, "A Reply to Baumrind," *Am. Psychol.*, vol. 19, no. 11, pp. 848–852, 1964.

[152] A. G. Miller, B. E. Collins, and D. E. Brief, "Perspectives on Obedience to Authority: The Legacy of the Milgram Experiments," *Journal of Social Issues*, vol. 51, no. 3, pp. 1–19, 1995.

[153] C. E. Sembroski, M. R. Fraune, and S. Šabanović, "He said, she said, it said: Effects of robot group membership and human authority on people's willingness to follow their instructions," *Robot Hum. Interact. Commun.*, pp. 56–61, 2017.

[154] A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, "Measurement of trust in human-robot collaboration," *Proc. 2007 Int. Symp. Collab. Technol. Syst. CTS*, pp. 106–114, 2007.

[155] M. Touré-Tillery and A. L. McGill, "Who or What to Believe: Trust and the Differential Persuasiveness of Human and Anthropomorphized Messengers," *J.*

Mark., vol. 79, no. 4, pp. 94–110, 2015.

[156] K. Ogawa, C. Bartneck, D. Sakamoto, T. Kanda, T. Ono, and H. Ishiguro, "Can an android persuade you?," *Proc. - IEEE Int. Work. Robot Hum. Interact. Commun.*, pp. 516–521, 2009.

[157] M. Watanabe, K. Ogawa, and H. Ishiguro, "Can Androids Be Salespeople in the Real World?," *Ext. Abstr. ACM CHI'15 Conf. Hum. Factors Comput. Syst.*, vol. 2, pp. 781–788, 2015.

[158] M. Roubroeks, J. Ham, and C. Midden, "When artificial social agents try to persuade people: The role of social agency on the occurrence of psychological reactance," *Int. J. Soc. Robot.*, vol. 3, no. 2, pp. 155–165, 2011.

[159] K. Shinozawa, F. Naya, J. Yamato, and K. Kogure, "Differences in effect of robot and screen agent recommendations on human decision-making," *Int. J. Hum. Comput. Stud.*, vol. 62, no. 2, pp. 267–279, 2005.

[160] M. Siegel, C. Breazeal, and M. I. Norton, "Persuasive robotics: The influence of robot gender on human behavior," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, 2009, pp. 2563–2568.

[161] Y. Yamamoto, S. Mitsuru, H. Kazuo, Y. Nobuyuki, and A. Yuichiro, "A request of the robot: an experiment with the human-robot interactive system HuRIS," *Robot and Human Communication*. pp. 204–209, 1992.

[162] C. Bartneck and J. Hu, "Exploring the abuse of robots," *Interact. Stud.*, vol. 9, no. 3, pp. 415–433, 2008.

[163] S. Nishio, H. Ishiguro, and N. Hagita, "Geminoid: Teleoperated android of an existing person," *Humanoid Robot. New Dev.*, no. June, pp. 343–352, 2007.

[164] C. T. Ishi, C. Liu, H. Ishiguro, and N. Hagita, "Lip Motion Generation Method Based on Formants for Tele-presence Humanoid Robots," *J. Robot. Soc. Japan*, vol. 31, no. 4, pp. 401–408, 2013.

[165] P. Baxter, J. Kennedy, E. Senft, S. Lemaignan, and T. Belpaeme, "From

characterising three years of HRI to methodology and reporting recommendations," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2016, vol. 2016–April, pp. 391–398.

[166] L. R. Goldberg, "A broad-bandwidth, public-domain, personality inventory measuring the lower level facets of several Five-Factor models," in *Personality Psychology in Europe*, vol. 7, 1999, pp. 7–28.

[167] J. M. Burger and Harris M, "Burger, Jerry M., and Harris M. Cooper. "The desirability of control.," *Motiv. Emot. 3.4*, pp. 381–393, 1979.

[168] Y. Kim and B. Mutlu, "How social distance shapes human-robot interaction," *Int. J. Hum. Comput. Stud.*, vol. 72, no. 12, pp. 783–795, 2014.

[169] S. Ivaldi, S. Lefort, J. Peters, M. Chetouani, J. Provasi, and E. Zibetti, "Towards engagement models that consider individual factors in HRI: on the relation of extroversion and negative attitude towards robots to gaze and speech during a human-robot assembly task," *Int. J. Soc. Robot.*, pp. 1–24, 2016.

[170] M. Zuckerman and et al, "What is the sensation seeker? Personality trait and experience correlates of the Sensation-Seeking Scales," *J. Consult. Clin. Psychol.*, vol. 39, no. 2, pp. 308–321, 1972.

[171] T. Rosenbloom, "Risk evaluation and risky behavior of high and low sensation seekers," *Soc. Behav. Pers.*, vol. 31, no. 4, pp. 375–386, 2003.

[172] C. Bartneck, M. van der Hoek, O. Mubin, and A. Al Mahmud, "'Daisy, Daisy, give me your answer do!,'" in *Proceeding of the ACM/IEEE international conference on Human-robot interaction - HRI '07*, 2007, p. 217.

[173] W. Allinson and J. Hayes, "The Cognitive Style Index," *J. Manag. Stud.*, vol. 33, pp. 119–135, 1996.

[174] D. Reitman, P. C. Rhode, S. D. A. Hupp, and C. Altobello, "Development and validation of the parental authority questionnaire - Revised," *J. Psychopathol. Behav. Assess.*, vol. 24, no. 2, pp. 119–127, 2002.

[175] W. Johal, S. Pesty, and G. Calvary, "Towards companion robots behaving with style," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 2014, vol. 2014–Octob, no. October, pp. 1063–1068.

[176] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The benefits of interactions with physically present robots over video-displayed agents," *Int. J. Soc. Robot.*, vol. 3, no. 1, pp. 41–52, 2011.

[177] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, "Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents," *Proc. Tenth Annu. ACM/IEEE Int. Conf. HRI*, pp. 117–124, 2015.

[178] P. H. Kahn, R. L. Severson, T. Kanda, H. Ishiguro, B. T. Gill, J. H. Ruckert, S. Shen, H. E. Gary, A. L. Reichert, and N. G. Freier, "Do people hold a humanoid robot morally accountable for the harm it causes?," *Proc. seventh Annu. ACM/IEEE Int. Conf. Human-Robot Interact. - HRI '12*, p. 33, 2012.

[179] H. J. Eysenck, "Biological basis of personality," *Nature*, vol. 199, no. 4898, pp. 1031–1034, 1967.

[180] A. M. Aroyo, F. Rea, and A. Sciutti, "Bringing Human Robot Interaction towards Trust and Social Engineering: Slowly & Secretly Invading People's Privacy Settings," *3rd Interdisciplinary Cyber Research ICR2017 workshop*, Tallinn, pp. 13–15, 2017.

[181] A. Rapoport and A. M. Chammah, *Prisoner's dilemma: a study in conflict and cooperation*, vol. 165. 1965.

[182] J. B. Hirsh and J. B. Peterson, "Extraversion, neuroticism, and the prisoner's dilemma," *Pers. Individ. Dif.*, vol. 46, no. 2, pp. 254–256, 2009.

[183] E. B. Sandoval, J. Brandstetter, M. Obaid, and C. Bartneck, "Reciprocity in Human-Robot Interaction: A Quantitative Approach Through the Prisoner's Dilemma and the Ultimatum Game," *Int. J. Soc. Robot.*, vol. 8, no. 2, pp. 303–317, 2016.

[184] A. Strudler, "The distinctive wrong in lying," *Ethical Theory Moral Pract.*, vol. 2,

no. 13, pp. 171–179, 2010.

[185] M. Marzanski, "On telling the truth to patients with dementia," *West. J. Med.*, vol. 5, no. 173, p. 318, 2000.

[186] P. Geach, *The virtues: The Stanton lectures 1973-74*. CUP Archive, 1977.

[187] A. Matthias, "Robot Lies in Health Care: When Is Deception Morally Permissible?," *Kennedy Inst. Ethics J.*, vol. 25, no. 2, pp. 169–162, 2015.

[188] B. M. DePaulo, B. E. Malone, J. J. Lindsay, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," *Psychol. Bull.*, vol. 1, no. 129, p. 74, 2003.

[189] M. Zuckerman, B. M. Depaulo, and R. Rosenthal, "Verbal and nonverbal communication of deception," *Adv. Exp. Soc. Psychol.*, vol. 14, pp. 1–59, 1981.

[190] S. M. Kassin, "On the psychology of confessions: Does innocence put innocents at risk?," *Am. Psychol.*, vol. 3, no. 60, p. 215, 2005.

[191] S. M. Kassin and G. H. Gudjonsson, "The Psychology of Confessions:A Review of the Literature and Issues," *Psychol. Sci. Public Interes.*, vol. 2, no. 5, pp. 33–67, 2004.

[192] S. M. Kassin and R. J. Norwick, "Why people waive their Miranda rights: The power of innocence," *Law Hum. Behav.*, vol. 2, no. 28, pp. 211–221, 2004.

[193] A. Vrij, S. Mann, and R. P. Fisher, "Information-gathering vs accusatory interview style: Individual differences in respondents' experiences," *Pers. Individ. Dif.*, vol. 4, no. 44, pp. 589–599, 2006.

[194] S. Leal and A. Vrij, "Blinking during and after lying," *J. Nonverbal Behav.*, vol. 32, no. 4, pp. 187–194, 2008.

[195] A. K. Webb, C. R. Honts, J. C. Kircher, P. Bernhardt, and A. E. Cook, "Effectiveness of pupil diameter in a probable-lie comparison question test for deception," *Leg. Criminol. Psychol.*, vol. 2, no. 14, pp. 279–292, 2009.

[196] A. Vrij and P. A. Granhag, "Eliciting cues to deception and truth: What matters are the questions asked," *J. Appl. Res. Mem. Cogn.*, vol. 1, no. 2, pp. 110–117, 2012.

[197] J. J. Walczyk, D. A. Griffith, R. Yates, S. R. Visconte, B. Simoneaux, and L. L. Harris, "Lie Detection BY Inducing Cognitive Load: Eye Movements and Other Cues to the False Answers of 'Witnesses' to Crimes," *Crim. Justice Behav.*, vol. 39, no. 7, p. 887–909., 2012.

[198] A. K. Webb, D. J. Hacker, D. Osher, A. E. Cook, D. J. Woltz, S. Kristjansson, and J. C. Kircher, "Eye movements and pupil size reveal deception in computer administered questionnaires," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009, pp. 553–562.

[199] A. Vrij, J. Oliveira, A. Hammond, and H. Ehrlichman, "Saccadic eye movement rate as a cue to deceit," *J. Appl. Res. Mem. Cogn.*, vol. 4, no. 1, pp. 15–19, 2015.

[200] A. Vrij, S. Leal, P. A. Granhag, S. Mann, R. P. Fisher, J. Hillman, and K. Sperry, "Outsmarting the liars: The benefit of asking unanticipated questions," *Law Hum. Behav.*, vol. 33, no. 2, pp. 159–166, 2009.

[201] A. Vrij, P. A. Granhag, S. Mann, and S. Leal, "Outsmarting the liars: Toward a cognitive lie detection approach," *Curr. Dir. Psychol. Sci.*, vol. 20, no. 1, pp. 28–32, 2011.

[202] K. Fukuda, "Eye blinks: New indices for the detection of deception," *Int. J. Psychophysiol.*, vol. 40, no. 3, pp. 239–245, 2001.

[203] J. A. Stern, L. C. Walrath, and R. Goldstein, "The Endogenous Eyeblink," *Psychophysiology*, vol. 21, no. 1, pp. 22–33, 1984.

[204] R. Goldstein, L. C. Walrath, J. A. Stern, and B. D. Strock, "Blink Activity in a Discrimination Task as a Function of Stimulus Modality and Schedule of Presentation," *Psychophysiology*, vol. 22, no. 6, pp. 629–635, 1985.

[205] L. O. Bauer, B. D. Strock, R. Goldstein, J. A. Stern, and L. C. Walrath, "Auditory Discrimination and the Eyeblink," *Psychophysiology*, vol. 22, no. 6, pp. 636–641, 1985.

[206] J. Beatty and B. Lucero-Wagoner, "The pupillary system," in *Handbook of psychophysiology (2nd ed.)*., 2000, pp. 142–162.

[207] D. P. Dionisio, E. Granholm, W. A. Hillix, and W. F. Perrine, "Differentiation of deception using pupillary responses as an index of cognitive processing," *Psychophysiology*, vol. 38, no. 2, pp. 205–211, 2001.

[208] J. J. Walczyk, K. S. Roper, E. Seemann, and A. M. Humphrey, "Cognitive mechanisms underlying to questions: Response time as a cue to deception," *Appl. Cogn. Psychol.*, vol. 17, no. 7, pp. 755–774, 2003.

[209] A. A. Harrison, M. Hwalek, D. F. Raney, and J. G. Fritz, "Cues to Deception in an Interview Situation," *Soc. Psychol. (Gott).*, 1978.

[210] F. Ferrari and F. Eyssel, "Toward a hybrid society," in *International Conference on Social Robotics*, 2016, pp. 909–918.

[211] L. J. Wood, K. Dautenhahn, H. Lehmann, B. Robins, A. Rainer, and D. S. Syrdal, "Robot-mediated interviews: Does a robotic interviewer impact question difficulty and information recovery?," in *Assistive Technology Research Series*, 2013, p. 131.

[212] J. J. Walczyk, J. P. Schwartz, R. Clifton, B. Adams, M. Wei, and P. Zha, "Lying person-to-person about life events: A cognitive framework for lie detection," *Pers. Psychol.*, vol. 58, no. 1, pp. 141–170, 2005.

[213] C. J. Ferguson and C. Negy, "Development of a brief screening questionnaire for histrionic personality symptoms," *Pers. Individ. Dif.*, vol. 66, pp. 124–127, 2014.

[214] D. N. Jones and D. L. Paulhus, "Introducing the Short Dark Triad (SD3): A Brief Measure of Dark Personality Traits," *SAGE*, vol. 21, no. 1, pp. 28–41, 2014.

[215] R. L. Solso, M. K. MacLin, and O. H. MacLin, "Cognitive psychology (7th ed.).," *Pearson Education Limited*. 2005.

[216] A. E. Cook, D. J. Hacker, A. K. Webb, D. Osher, S. D. Kristjansson, D. J. Woltz, and J. C. Kircher, "Lyin' eyes: Ocular-motor measures of reading reveal deception," *J. Exp. Psychol. Appl.*, vol. 18, no. 3, p. 301, 2012.

[217] T. Qin, J. Burgoon, and J. F. Nunamaker, "An exploratory study on promising cues in deception detection and application of decision tree," *Proc. 37th Annu. Hawaii Int. Conf. Syst. Sci.*, pp. 23–32, 2004.

[218] D. T. Lykken, *A tremor in the blood: Uses and abuses of the lie detector*. Plenum Press, 1998.

[219] A. M. Rosenthal-Von Der Pütten, F. P. Schulte, S. C. Eimler, S. Sobieraj, L. Hoffmann, S. Maderwald, M. Brand, and N. C. Krämer, "Investigations on empathy towards humans and robots using fMRI," *Comput. Human Behav.*, vol. 33, pp. 201–212, 2014.

[220] J.-Y. Sung, L. Guo, R. E. Grinter, and H. I. Christensen, "'My Roomba is Rambo': intimate home appliances," in *International Conference on Ubiquitous Computing*, 2007, pp. 145–162.

[221] J. Garreau, "Bots on the ground," *Washington Post*, pp. 1–6, 2007.

[222] A. C. Horstmann, N. Bock, E. Linhuber, J. M. Szczuka, C. Straßmann, and N. C. Krämer, "Do a robot's social skills and its objection discourage interactants from switching the robot off?," *PLoS One*, vol. 13, no. 7, pp. 1–25, 2018.

[223] S. H. Seo, D. Geiskkovitch, M. Nakane, C. King, and J. E. Young, "Poor Thing! Would You Feel Sorry for a Simulated Robot?," *Proc. Tenth Annu. ACM/IEEE Int. Conf. Human-Robot Interact. - HRI '15*, pp. 125–132, 2015.

[224] P. H. Kahn, T. Kanda, H. Ishiguro, N. G. Freier, R. L. Severson, B. T. Gill, J. H. Ruckert, and S. Shen, "'Robovie, you'll have to go into the closet now': Children's social and moral relationships with a humanoid robot," *Dev. Psychol.*, vol. 48, no. 2, pp. 303–314, 2012.

[225] C. Bartneck, M. Verbunt, O. Mubin, and A. Al Mahmud, "To kill a mockingbird robot," *Proceeding ACM/IEEE Int. Conf. HRI '07*, p. 81, 2007.

[226] K. Darling, "Why we have an emotional connection to robots?," p. TED, 2018.