UNIVERSITY OF GENOVA

PHD PROGRAM IN BIO-ENGINEERING AND ROBOTICS
ISTITUTO ITALIANO DI TECNOLOGIA
INFORMATION AND COMMUNICATION TECHNOLOGIES (ICT)
ROBOTICS BRAIN AND COGNITIVE SCIENCES (RBCS)

Social Engineering Defense Solutions Through Human-Robot Interaction

by

Dario Pasquali

Thesis submitted for the degree of *Doctor of Philosophy* (34° cycle)

July 2022

Dr. Francesco Rea — Supervisor
Dr. Stefano Bencetti — Supervisor

Thesis Jury:

Prof. Richard Matthews, *University of Adelaide* — External examiner
Prof. Eduardo Benitez Sandoval, *University of New South Wales Sydney* — External examiner
Prof. Nicoletta Noceti, *University of Genova* — Internal examiner

Dibris

Department of Informatics, Bioengineering, Robotics and Systems Engineering

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Dario Pasquali
June 2022

Abstract

Social Engineering is the science of using social interaction to influence others on taking computer-related actions of attacker's interest. It is used to steal credentials, money, or people's identities. After being left unchecked for a long time, social engineering is raising increasing concerns.

Despite its *social* nature, state-of-the-art defense systems mainly focus on *engineering* factors. They detect technical features specific to the medium employed in the attack (e.g., phishing emails), or they train final users on detecting them. However, the crucial aspects of social engineering are *humans*, their vulnerabilities, and how attackers leverage them, gaining victims' compliance. Recent solutions involved victims' explicit perception and judgment in technical defenses (Humans-as-a-Security-Sensor paradigm). However, humans also communicate implicitly: gaze, heart rate, sweating, body posture, voice tone, ..., are physiological and behavioral cues that implicitly disclose humans' cognitive and emotional state. In literature, expert social engineers reported monitoring such cues from the victims continuously to adapt their strategy (e.g., in face-to-face attacks); also, they stressed the importance of controlling them to avoid revealing the attacker's malicious intentions.

This thesis studies how to leverage such behavioral and physiological cues to defend against social engineering. Moreover, it researches humanoid social robots - more precisely the iCub and Furhat robotic platforms - as novel agents in the cybersecurity field. Humans' trust in robots and their role are still debated: attackers could hijack and control them to perform face-to-face attacks from a safe distance. However, this thesis speculates robots could be helpers, everyday companions able to warn users against social engineering attacks, better than traditional notification vectors could do. Finally, this thesis explores leveraging game-based entertaining human-robot interactions to collect more realistic, less biased data. For this purpose, I performed four studies concerning different aspects of social engineering.

Firstly, I studied how the trust between attackers and victims evolves and can be exploited. In a Treasure Hunt game, players had to decide whether trust the hints of iCub. The robot

showed four mechanical failures designed to mine its perceived reliability in the game, and it could provide transparent motivations for them. The study showed how players' trust in iCub decreased only if they perceived all the faults or the robot explained them; i.e., they perceived the risk of relying on a faulty robot.

Then, I researched novel physiological-based methods to unmask malicious social engineers. In a Magic Trick card game, autonomously led by the iCub robot, players lied or told the truth about gaming card descriptions. ICub leveraged an End-to-end deception detection architecture to identify lies based on players' pupil dilation alone. The architecture enables iCub to learn customized deception patterns, improving the classification over prolonged interactions.

In the third study, I focused on victims' behavioral and physiological reactions during social engineering attacks; and how to evaluate their awareness. Participants played an interactive storytelling game designed to challenge them against social engineering attacks from virtual agents and the humanoid robot iCub. Post-hoc, I trained three Random Forest classifiers to detect whether participants' perceived the risk and uncertainty of Social Engineering attacks and predict their decisions.

Finally, I explored how social humanoid robots should intervene to prevent victims' compliance with social engineering. In a refined version of the interactive storytelling, the Furhat robot contrasted players' decisions with different strategies, changing their minds. Preliminary results suggest the robot effectively affected participants' decisions, motivating further studies toward closing the social engineering defense loop in human-robot interaction.

Summing up, this thesis provides evidence that humans' implicit cues and social robots could help against social engineering; it offers practical defensive solutions and architectures supporting further research in the field and discusses them aiming for concrete applications.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

We live in an era of high connectivity and easy communication; we continuously deal with emails, notifications, and push messages; hundreds of billions of them are sent everyday [1]. However, we are unlikely to pay the same attention to each email opened, login, or website visited [1]. Humans learned to use and trust technology [2]; however, by overtrusting technology and our understanding of it, we risk falling into the so-called *Social Engineering* attacks.

We could define *Social Engineering* as the science - or sometimes even the *art* - of using social interactions to get an individual or an organization to perform specific tech-related actions relevant to the attacker victims did not want to do. Social Engineering, in cybersecurity, is used to steal credentials, money, or even people's identity. For many years, the phenomenon has been overlooked and reduced to simple persuasion tricks to get free goods. However, technology and the Internet are becoming a consistent part of our lives: they define how we work, entertain, and sometimes live our social lives. The network is becoming a breeding ground for social engineers and cyber attacks. Hence, researchers [3] and companies [4], raised their concerns about Social Engineering and called for immediate defensive action.

However, the state-of-the-art Social Engineering defense systems are usually unable to keep up. They still consider it as any other cybersecurity threat; they either develop technical solutions based on the attacks' features (e.g., phishing email filters based on the

[1] https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/

email metadata or content), to prevent them from reaching their target or train users to detect the same red flags. That is it: they mainly focus on the *"engineering"* aspect of the threat.

Social engineering is, first of all, related to sociality. The critical factor is how humans, either the attacker(s) or the victim(s), socially interact. Humans, during social interactions, communicate on many levels: they could *explicitly* verbalize an intention; but show at the same time completely different *implicit* cues (i.e., with their bodies or voice tones) [5]. For instance, victims, when making decisions under uncertainty, such as whether to comply (to accept) or not with the requests of a stranger, travel through a wide variety of emotional states, fluctuations of their cognitive efforts, and stress, which they usually reveal on a non-verbal level [6]. Expert social engineers - like Kevin Mitnick or Christopher Hadnagy - continuously monitor the evolution of victims' implicit non-verbal signals to adapt their attack strategy [7, 6]. Conversely, any social engineering attack starts with formulating a credible pretext meant to gain victims' trust [8]; attackers are humans too: they must fabricate and maintain the pretext during the interaction. This process could lead to stress and variation of cognitive load, reflecting on their implicit communication [9].

The central question leading my research is: *Would it be possible to take advantage of such mental state variations to predict the occurrence of Social Engineering attacks? How novel technologies (e.g., machine and deep learning) can help?*

Recently, another relatively new technology is becoming an essential part of modern society: *Robots*. Robots are slowly moving from research labs and cinematic sci-fi environments to everyday contexts. They are employed in industries, schools, hospitals, and offices and used for entertainment in theme parks or as toys in our homes. Sooner or later, they are expected to become our everyday-life companions. However, as for all new technologies, people need to learn how to use them; this is the primary concern of Human-Robot Interaction (HRI): the science that studies how humans perceive, interact, and adapt to robots, improving them and facilitating their interaction with humans [10, 11]. Furthermore, people will have to decide whether to trust robots or not. Recent studies showed how people tend to trust and humanize robots, forgetting the maze of computers, networks, and informatics systems behind them [12]. Social Engineers could hijack robots, taking advantage of their embodiment and humans' trust bias, to perform face-to-face attacks from a safe distance. For instance, an attacker could remotely control a rover robot and enter a secured area by following an authenticated user (piggybacking) [13]; or it could persuade humans into releasing sensitive information during an informal dialogue [14, 15]. However, this trust bias could be used for

good purposes: artificially intelligent robots, able to detect in advance humans' compliance with social engineering attacks, could exploit social interactions to intervene, preventing users from becoming victims.

This thesis aims to develop Social Engineering defense systems centered on humans rather than technology. While state-of-the-art systems focus on the technical factors of attacks, I speculate the attention should be on the attackers and victims, on how they socially communicate and react during the interactions. For this purpose, I took advantage of two novel technologies in the cybersecurity field: machine learning models and social robots.

The first crucial challenge is evaluating victims' internal states, linking them with Social Engineering threats, and leveraging them to detect the threats in advance. For these purposes, machine learning techniques constitute practical tools to unfold complex and not-straightforward problems like understanding humans' perceptions or predicting their behavior.

Also, social robots (and human-robot interaction, HRI) are particularly relevant in this context. Robots can be experimental tools that provide predictable, repeatable, and reproducible stimuli, more than human actors could do [16]. For instance, they could be used to run social engineering attacks without being blocked by humans' morale or shyness [17]. Robots can learn from and adapt to humans: closing the loop on human-robot interaction; they can perceive humans and affect their intentions and behavior. Such features make them good candidates, for instance, to help humans defend against Social Engineering threats.

The second challenge is how to properly study social engineering or, in general, how to run compliance studies in a laboratory. Indeed, traditional experimental setups could involve several biases that inevitably compromise the realism and generalization of the scientific findings to real-world environments. Participants usually bend their behavior and intentions to the experimental context acting as they assume they are supposed to [18, 19]. Hence, to study the social engineering phenomenon, it is crucial to run in-the-wild studies or immerse participants in ecological contexts that mitigate their perception of being in a laboratory. For this purpose, social robots can be playful companions, engaging participants in ecological, and social interactions, making them forget to be in an experimental setup.

The schema in Figure 1.1 resumes my research approach. Borrowing a cybersecurity term, robots are *"men-in-the-middle"*. The primary interaction occurs between two human (or groups of human) actors: the attacker(s) and the victim(s). Robots are third-part entities, lying between attackers and victims. They monitor both directions of this interaction on an

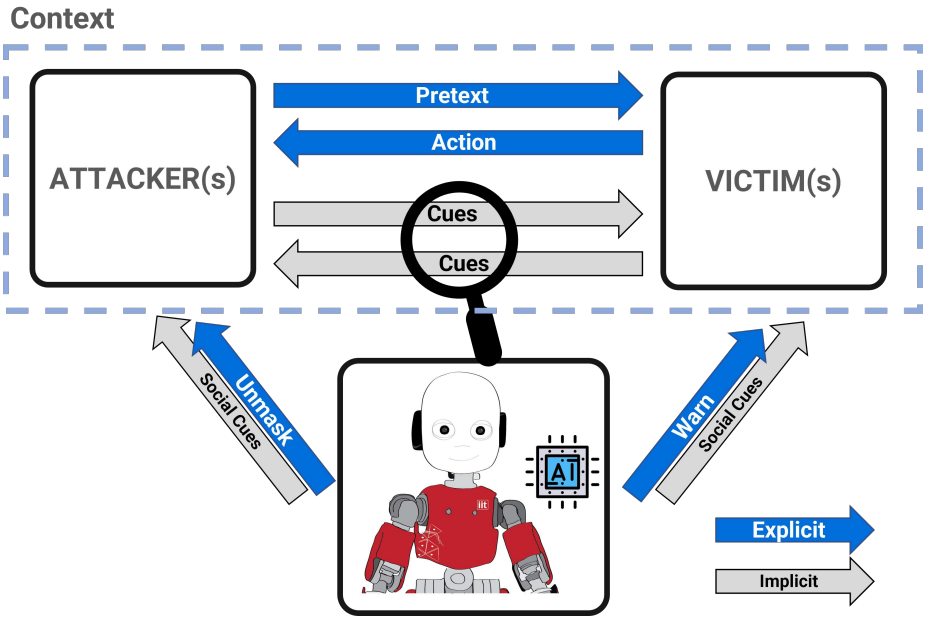

Figure 1.1 Human-Robot Interaction schema with a robotic agent (e.g., iCub) monitoring the implicit communication between the Social Engineering attacker(s) and the victim(s), intervening if necessary.

implicit level, perceiving actors' actions and reactions to spot dangerous situations (i.e., the occurrence of Social Engineering attacks) and potentially intervene.

One could speculate that robots are superfluous from a purely sensory point of view. A camera could monitor a human-human interaction (e.g., the hall of a strategic building), just as a phishing filter could monitor incoming emails. However, it has been proved that robots' physical and social presence can affect the other agents sharing the interacting context [20, 21]. While state-of-the-art warning systems rely on pop-ups, notifications, emails, or other notification forms, humanoid robots can take advantage of a more comprehensive set of communication vectors (i.e., movements, gaze, voice, facial emotion, ...). Also, they can assume a social role based on the interacting context. They are usually not perceived as generic entities, offshoots of systems; robots are agents able to affect and be affected by the context; they elicit the development of social rapport, trust, and friendship, all concepts not easily definable for traditional defense systems.

1.2 Thesis Structure

The remainder of the thesis is organized as follows:

Part I explains the background of my activities, reviews the relative state-of-the-art, and presents the scientific and technological objectives I planned to achieve. The background (see **chapter 2**) focuses on the two core concepts of my research: Social Engineering (SE) and Human-Robot Interaction (HRI). It starts by presenting Social Engineering and its building blocks; it reviews the available state-of-the-art defense methods and discusses how they could be improved by using human physiological and behavioral reactions. Then, it introduces the iCub and Furhat robots, the two humanoid robotic platforms I employed in my research. Finally, it shows how robots and human-robot interaction can grow from tools to companions; it discusses their implication for human privacy and safety and how they could be used in Social Engineering defense. Afterward, **chapter 3** recaps my research questions and the scientific and technological objectives of this thesis.

Part II presents the scientific and technological evidence of different studies, shedding light on how novel methods and technologies can help in social engineering warfare. Following the schema in Figure 1.1, I decomposed the social engineering issue into four topics:

Firstly, I studied the *context* in which SE attacks happen (**Unreliable Treasure Hunt**, **chapter 4**), following a state-of-the-art framework (see section 2.1.1). I found how human participants still trusted the humanoid robot iCub, even if it showed to be unreliable (i.e., with mechanical faults). Participants' trust in the robot decreased if they realized mechanical faults were happening (either by seeing it or thanks to transparent motivations from the robot). This suggested that users' perception and appraisal of the risk involved in a decision (i.e., relying on an untrustworthy agent) are particularly relevant in shaping their behavior.

Then, I focused on the *attackers*' behavior and reaction while acting based on a false pretext (i.e., lying) (**End-to-End Lie Detector**, **chapter 5**). I proved that pupillometry is an effective physiological proxy to evaluate humans' cognitive effort related to deception during informal human-robot interactions. I enabled the humanoid robot iCub to detect humans' lies only based on pupil dilation, autonomously and in real-time. Moreover, I proved how it is possible to train more robust models aiming for real-world applications with machine learning. Starting with the promising results on pupillometry, I further expanded my research on human physiology and behavior.

In the third project (**Social Engineering Adventure, chapter 6**), I studied *victims'* behavioral and physiological reactions while under attack. In particular, I leveraged such reactions to model targets' risk appraisal of SE attacks and potentially predict their behavior against manipulations from the iCub robot. After providing the statistical evidence of this approach, I show how machine learning once again proved effective in modeling humans' perceptions and behavior.

Finally, I looked for novel *robot*-based intervention strategies (**Adventurer Robot Companion, chapter 7**), exploring the different strategies a Furhat robot can use to elicit behavioral changes against Social Engineering attacks. I present the preliminary results of this ongoing project and speculate on how to close the loop in Human-Robot Interaction.

The last **Part III** (**Final Discussion, chapter 8**) comprehensively discusses the achievements of my research with respect to the questions I aimed to answer. Also, it analyzes the limitations still affecting my findings and how I addressed or plan to address them soon. I conclude by speculating how this novel research field could expand in the future and its implication for society, aiming for real-world applications.

Part I

Background

Chapter 2

State of the Art

Social engineering, a cybersecurity threat for a long time left unchecked, is raising more and more concerns. The wild-fire-like diffusion of technology, and the ability to quickly communicate to thousands of people, made the world wide web a playground for attackers. Furthermore, the distribution of intelligent assistants and robots in everyday spaces poses a threat that researchers must consider. This chapter clarifies what social engineering is, which state-of-the-art defense methods are available, and how it would be possible to improve them, focusing more on the human actors of the attacks. In this landscape, robots will be crucial: from one side, it is not clear whether they can be trusted; on the flip side, they could be effective companions able to help human partners.

2.1 Social Engineering definition

The introduction gave an intuitive definition of Social Engineering (SE); however, defining SE is complex due to its multifaceted and multidisciplinary nature. Several researchers from the most diverse fields gave their definition of SE. In this thesis, I will employ the most recent and comprehensive one from Wang et *al.* [3]

"Social engineering, in the context of cybersecurity, is a type of attack wherein the attacker(s) exploit human vulnerabilities using social interaction to breach cyber security [with or without the use of technical means and technical vulnerabilities].

2.1.1 Social Engineering attack framework

Most of the social engineering attacks follow a common framework formalized by Mouton et *al.* in [8] (see Figure 2.1).

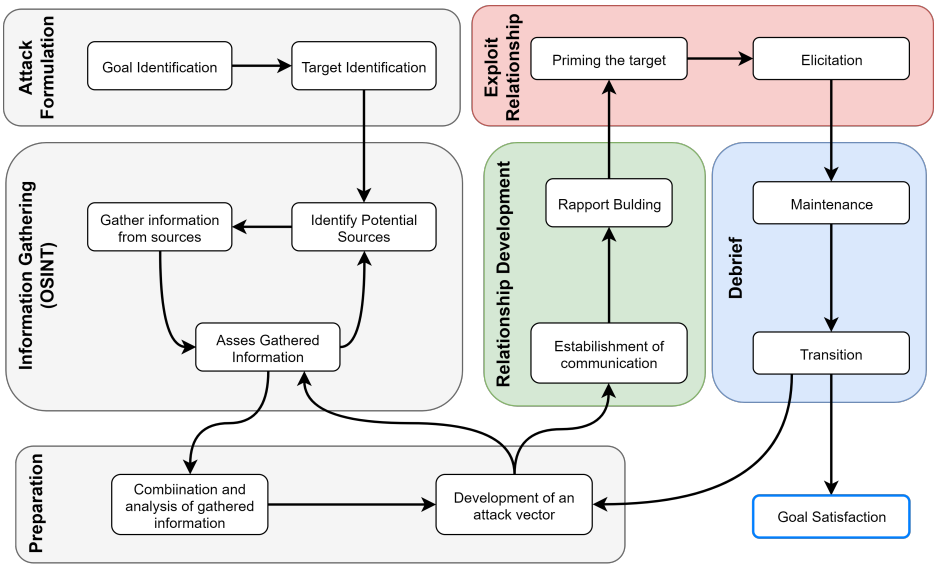

Figure 2.1 The social engineering attack framework formalized by Mouton et *al.* in [8]

Attacks start by identifying the attacker's objective and target (*Attack Preparation (gray)*. They could be composed of single or several actions, each with a specific purpose. Attackers collect Open Source Intelligence (OSINT) for each target (e.g., job, role, social life, social network accounts, habits, contacts, ...); and use them to identify the most suitable attack strategy. The primary outcome of this preparatory phase is a credible and consistent cover story called **pretext**. Then, the attackers start interacting with the victims to build a rapport and gain their trust (*Trust Development (green)*. This phase could occur with just an email or over multiple interactions. Once the trust is gained, the attackers take advantage of it (*Trust Exploitation (red)*): they leverage human vulnerabilities to plant the seed of their request. Here, the "magic" happens: attackers do not force victims to comply; victims **decide** take the attackers' desired action. Finally, the attackers "reset" the victims to a state wherein they would not think about the consequences of their actions *Debriefing (blue)*.

2.1.2 Attacks building blocks

SE attacks are usually composed of 3 building blocks: *attack methods*, *human vulnerabilities*, and *effect mechanisms* [22].

Attack Methods

Attack methods are the mediums (either technical or not) employed to perform the attack; they are *"what"* attacks are about.

In **Phishing** attacks [23, 24], the attacker sends a spoofed email or instant message to the victims. Phishing emails could contain compromised links (usually shortened or embedded in the text to mask them), leading to *waterholes*, fake replicas of trusted websites (e.g., victims' bank), meant, for instance, to steal login credentials. Also, they could include fake attachments, compromised with malware, ransom-wares, or Trojan horses, meant to infect victims' systems and blackmail them or install back-doors for future access. Finally, attackers could elicit the payment of vast amounts of money, usually promising a future gain.

Given the high usage of emails in our everyday lives, phishing is probably the most used attack method. Over the years, attackers developed multiple phishing variants. The most generic one is called **massive phishing**: a compromised email is sent to many targets, hoping to make some of them fail. Massive phishing can also be the first step of more prolonged attacks: attackers could target all the employers of a company, trying to steal the credential of at least one of them, which they will impersonate in further attacks.

Spear phishing happens when the target is a particular individual, or company [25]. The attack is highly customized to the victims, including characterizing information they can understand. Spear phishing becomes **whaling** [23] when the target has a high rank in its context. For instance, an attacker could steal a company's CEO's identity by sending an email to the purchasing office asking to quickly update a supplier's payment contacts and immediately forward a late payment.

Also, the recent adoption of two-factor authentication (2FA) and one-time passwords (OTP) pushed attackers to develop the **real-time phishing** attacks [26]. The attacker works as a "man-in-the-middle" replacing the 2FA login screen of the actual website (e.g., as for Microsoft Office 365 in [26]): the attackers pretend to be the legitimate login page for the user, and the real-user for the website; when the user logs-in, the attacker collects and forward the credentials to the real website so the user cannot notice the attack.

Finally, the **reverse phishing** [23] consists of the attacker sending or replacing the legitimate information with compromised contacts tricking the victims into contacting the attacker themselves (e.g., replacing the IT support email or phone number).

Before the diffusion of emails, other attack methods were used. For instance, **Vishing** is a phishing attack via telephone (voice phishing), while **Smishing** acts via instant messages (i.e., Telegram, Whatsapp, or social networks) or the most traditional SMSs. Also, massive

phishing most likely evolved from **biting** attacks [27] wherein the attacker left compromised USB drives on the ground, hoping for some curious, unaware victim to plug them into their computers (and infect them) [3].

Piggybacking and **tailgating** are two methods employed to get physical access to a restricted area, like a private company floor. With piggybacking, attackers take advantage of victims' kindness by asking to hold a door or an elevator because they are in a rush; in tailgating, instead, attackers follow an unaware victim through a secure door [3].

In the latter two scenarios, attackers usually employ some form of **pretexting**: they collect specific, sometimes sensitive, information and lingo relative to the environment they want to penetrate and use them to pass themselves for a legitimate user or employee. Pretexing becomes **impersonation** when the attacker pretends to be a particular individual whose identity is stolen or crafted. Impersonation rarely happens face-to-face; however, it is highly used with phishing (e.g., the whaling attack mentioned above) or on social networks where it is easy to build bots depicting real or fake individuals.

Human Vulnerabilities

Human vulnerabilities are the root elements attackers leverage to gain victims' compliance; they are *"why"* an attack is effective. Human vulnerabilities have been extensively studied, mainly in psychology and sociology. They can be categorized into three groups.

Emotion and Feelings Emotions (like fear, surprise, happiness, anger, sadness, or disgust) and feelings (like curiosity, excitement, surprise, tension, impulsion, guilt, . . .) are the most exploited human vulnerabilities [28]. Fear is usually paired with authority (e.g., a boss or a supervisor) [25], or with the potential missing of a valuable good/benefit. Fear pushes people to act faster and think less than they would typically do; also, we usually try to avoid sadness or the feeling of guilt: if the attacker makes us feel (or risk to feel) in that way, we could take softhearted decisions. Finally, surprise and curiosity can trick victims into acting without thinking, especially when paired with a limited period to act or a luscious gain [27].

Human Nature Every human, even if strong-willed, could succumb to greed, lust, and gluttony [29]. Especially when paired with curiosity: the promise of an immediate and free benefit is a potent weapon social engineers usually employ. Similarly, humans tend to obey ground rules or stereotypes society defines. For instance, "helping those who are in need", "being helpful and polite with others", and "sympathize with those who are similar to us"

[7, 6]. Attackers can easily exploit all these "virtues". Also, the specific nature of each human, like individuals' personality traits, could make some people more vulnerable than others [30, 31].

Behavior and Cognition Finally, humans usually learn and improve by repetition. Sometimes we repeat the same actions so often that we get used to them and develop habits. These heuristics or mental shortcuts [32] (e.g., intuitively judging a payment request from an old supplier) can lead to a loss of attention and the failure into social engineering traps [33]. In the same way, habits (e.g., checking a trusted website every morning before starting working) can be exploited (e.g., by infecting or replacing the website fake clone, usually called waterholes) [34]. Finally, sometimes the mental shortcuts are just due to ignorance, inexperience (e.g., a new employee), or an excessive trust in one's judgment ability [35, 36].

Effect Mechanisms

Finally, the effect mechanisms are the persuasion and manipulation principles that allow the attack method to exploit human vulnerabilities. Intuitively, they describe *"how"* attack methods can take advantage of human vulnerabilities. Effect mechanisms have been studied mainly in the psychology field:

The Elaboration Likelihood Model Firstly, Petty and Cacioppo [37], defined the Elaboration Likelihood Model (ELM). They explained how humans' behavior could be influenced and manipulated following two *"routes"* of persuasion:

- The **central route** is related to learning persistent habits over repetition; it depends on the targets' inner motivation and ability to understand and rationalize the persuasive arguments. For instance, if we need to buy a smartphone, we usually evaluate several models, their price, reviews, technical features, support, . . . ; thinking through all the pros and cons, will lead to a decision.

- The **peripheral route**, instead, is used to elicit a temporary, usually impulsive behavior. Peripheral messages - sometimes referred to as *"cues"* - can bypass rational thinking, eliciting specific and predictable responses. Social Engineers take advantage of those cues too. For instance, going back to the smartphone example, rather than evaluating the actual features, we could buy the most famous, probably over-priced, brand just because it is considered a "status symbol" and several peers have it.

The Six Principles of Persuasion Then, Professor Cialdini [38], built on the ELM framework, identifying the *"Six principles of persuasion"*:

- **Reciprocity** is the instinctive tendency to return a favor or, in general, to give something in return when someone gives it to us. The object of the exchange could be anything: a specific object, time, love, or personal information (i.e., an attacker reveals something, usually fake, of himself; the victim is then facilitated to reveal a piece of personal information in return). For instance, funds and signature collectors significantly use reciprocity by giving a simple gift (e.g., a flower, a book, a bracelet, ...) to the target before advancing their requests [39]. The same techniques are also effective when asking for sensitive data like passwords [40]. However, the willingness to reciprocate highly depends on the perceived value of the exchanged object and how much we like (and are similar to) the counterpart [6].

- **Commitment & Consistency**, also known as the "foot-in-the-door" technique, leverages humans' tendency to finish what we started. We all want to appear reliable and consistent, to show that our "word" matters. Once we commit to doing something, we will probably do it again. Also, the principle can be exploited to ask and obtain more than the initial commitment [41]. A historical experiment showed how participants were more willing to put a big sign covering half of their house if they agreed to put a small sign in their garden one month in advance [41]. Commitment & consistency are highly used in marketing, for instance, with wish lists, free trials, try-at-home clothes, "maybe you would like" accessories, ... The key is to obtain the commitment: once a decision is made, people will try to keep it.

- **Social Proof** is the tendency to make a decision based on what others do, even if in contrast with our own will or habits [42]. We all seek validation from peers, and knowing several others did something is a strong incentive during decision-making. Generally speaking, social proof is used to manipulate users' perceptions of what is socially acceptable. *"People also bought ... "* marketing suggestions, product reviews, and canned laughs in sitcoms are all forms of social proof.

- For the **Liking** principle, people tend to trust what they like or are alike. A friend is usually considered more trustworthy than a stranger. Liking is usually employed to start an interaction and build a solid rapport before advancing a request. However, the attacker's appreciation toward the target must be sincere and based on reality to be effective [7]. Liking is a potent weapon when paired with social proof: the need to

belong, keep belonging and seek approval from people we share something with (e.g., similar objects, beliefs, habits, or even problems) are easy paths to compliance. For instance, it has been extensively used by radical groups [43].

- Obeying to the **authority** is another crucial feature of society. Identifying who has more power, strength, fame, or expertise in every field is possible. Orders from our boss will probably be executed immediately and without discussions; also, an expert in a specific field, like a professor [44] or an IT technician, is usually perceived as more trustworthy than other colleagues. Indeed, researchers showed how, in the presence of expert advice, the brain areas responsible for critical thinking and counter-arguing are practically inactive [45]. Researchers proved authority is effective even just from an email signature [46, 47]; however, victims must recognize and acknowledge the authoritative figure; for example, just wearing authoritative clothing is not sufficient [48]. Finally, authority is usually paired with the fear emotions of the feeling of guilt, for instance, in impersonation attacks or whaling.

- Finally, the **Scarcity** (or **Reactance**) principle leverages on impulsivity and mental shortcuts when people need to act in a short time [49]. If the time for acting is reduced, victims will be more willing to act without thinking. For instance, a phishing email could convince people that their bank account got hacked and that they must change their credentials as soon as possible to avoid money loss. Scarcity also works with quantities: if an item is available in small amounts, its perceived value will be higher [50]. This principle is used, for instance, in sales: companies could dry supplies to increase the perceived values of the items (e.g., as Nintendo did with Nintendo Wii [51])

Furthermore, Cialdini stressed how the key is the formulation and maintenance of an excellent *framing* (i.e., the pretext); he usually refers to *"Pre-suasion"*: half of the persuasion act is done by the context in which the influence is performed [52].

The Science of Human Hacking Recently, the famous hacker and penetration tester Christopher Hadnagy [6] built on Cialdini's work defining eight principles in the specific field of SE. The main differences concern the expansion of the *reciprocity* principle with Obligation and Concession:

While the feeling of indebtedness drives the reciprocity principle, the **obligation** principle focuses on humans' *expected* behaviors and the conformation with social norms. For instance, it is considered rude not to hold the door or the elevator for someone in a rush or carrying

boxes. Similarly, it could be impolite to ignore or not answer a direct question when asked. However, social norms are strongly cultural-dependent (i.e., Japanese people prefer not answering rather than reply with a *"no"* [53])

Concession instead is somehow similar to the Commitment & Consistency. It is about lowering down the entity of a request it was (and it was designed to be) initially rejected. For instance, a fundraiser could ask for a huge donation that will be more likely to be refused, lowering it to a more realistic and acceptable one. Similarly, it is much easier to ask victims to release a small amount of information when collecting personal data and then dig into the more sensitive ones.

Finally, Hadnagy stressed that it is crucial to define clearly the framing in which an attack is performed. In the first seconds of an interaction, even before the attack begins, the attacker needs to answer *"who are you?"*, *"What do you want?"*, *"Are you a threat?"*, *"How long will this take?"*. A good framing clearly defines and maintains those interaction traits.

2.2 Social Engineering Defense

Given the traditional structure and the building blocks of a SE attack, it is easier to understand the defense actions undertaken. The state-of-the-art defense systems can be categorized into two main "layers": (i) Active detection systems; and (ii) Awareness-raising training & policies.

2.2.1 Active detection systems

Active defense systems treat social engineering as any other cybersecurity attack by preventing attackers from reaching the victims. Hence, active solutions are strictly tied to specific attack methods and mediums.

The **phishing emails** and **water-hole websites** domains are the most covered. Phishing email filters build on traditional spam detectors [54], enhancing their capabilities. They employ Machine Learning [55, 56] and Natural Language Processing (NLP) [57] techniques to scan the email body, looking for named entities; bad, urgent, imperative, or generic words and sentences; masked or shortened links; misspelled words and other features [58–60, 24]. Also, they process the email header, looking for misspelled domains or the redirection to proxies [59, 61]. Phishing websites and waterholes are beyond the scope of this dissertation to be discussed in detail. Besides their multimedia-based content, they could include malicious

code executed in the browser. Traditional methods include keeping blocklists and allowlists of websites and domains; however, it is too easy to fabricate novel fake domains with minimal misspellings [62]. Again, machine and deep learning techniques showed their potential. Most of the solutions focus only on distinguishing legitimate and fake URLs [63], while others also include the website source code [64, 55], or their visual appearance [65]. Also, genetic algorithms have been used [66]. Finally, a few solutions analyze the website embedded ads [67].

Lastly, Online Social Networks (OSNs) are like a playground for attackers: they can both collect targets' personal information, identify their trusted friends [68], and contact them directly or via reverse engineering [69]. Even the absence of an individual on a social network can be beneficial for attackers [69]. On social networks, active defenses aim at detecting impersonation, bots and botnets [70, 71]. A bot is a fake, software-controlled user impersonating a natural or crafted person. Bots could be used to drive social opinion on social networks comment sections [72]. Also, they could send spoofed links to other users; clicking on them would hijack their browser, making them part of the *botnet*. Each hijacked user is controllable by the hacker and will try to enlarge the botnet. Researchers explored the factors leading to compliance in chat sessions with bots [70] and how to identify fake users [73, 74]. On the botnets domain, researchers simulated their diffusion [75], studying how it can be facilitated or resisted [76, 77], and how to detect them [78]. Finally, they studied how the personality traits of an individual make them more willing to comply or, in general, to release personal information on social networks [79, 80].

The main limitation of active defense systems is the high creativity and fast adoption of the attackers' novel technologies and vectors. Indeed, it is faster for the attacker to identify a new vulnerability and craft a quick attack, usually called *Zero-Day attacks*, than defining an effective countermeasure [81]. Hence, a second defense layer is required: training the victims and raising their awareness [82].

2.2.2 Awareness-rising training & policies

If an attacker bypasses the detection and filtering systems, the final users are in charge of detecting it. At the Italian Institute of Technology, as in most companies and institutes, we pursue extensive training programs to instruct final users to protect them from SE attacks. Training is traditionally composed of (i) teaching material meant to raise people's awareness,

(ii) fake phishing campaigns meant to test the training program's effectiveness, and (iii) enforced policies, defining basic behavioral rules to adopt in the workplace.

Training campaigns teach users to focus on the same technical features the active detection systems focus on [82]. For instance, in the phishing domain, they raise users' attention to the email senders' name and address; the email object; the presence of shortened URLs or unexpected attachments; misspelled words, generic greetings, or unexpected (un)friendly tone; they teach them to cross-check the names of senders and to hover on and check embedded URLs before clicking on them. Traditional campaigns involve computer-based teaching sessions with quizzes [83]. However recently, gamification and serious games have been extensively used both via computer-based games [84, 85, 83, 86], board games [87], or even more structured competitions [88]. Games help the learning process, facilitating the learners' engagement and maintaining the mastered concept over time [89]. Also, when played in small groups of colleagues, board games train employers and facilitate team-building and mutual aid against cyber attacks. On the other side, not everybody likes games, so the practice must be carefully balanced [89]. Also, researchers focused on how to evaluate the training efficacy [6, 90–92], and showed how priming and warning are not effective anymore in protecting victims [36].

Policies are meant to prevent physical (i.e., piggybacking and tailgating) and face-to-face attacks; or protect from Open Source INTelligence (OSINT) techniques to collect data like dumpster diving and shoulder surfing. They instruct people not to allow anybody in secured areas without authorization, to lock their devices, shred essential documents, avoid entering sensitive information in public places, or not connect to unknown WiFi networks. For this purpose, several models have been developed to assess how much an infrastructure or network is susceptible to SE attacks [29, 80, 93–96].

However, even one "hit" is enough to compromise an entire institution. SE attacks happened in the past and kept happening, especially in conjunction with worldwide events like the COVID-19 pandemic [97], or the 2022 Ukraine war [4].

Hence, the question is *How could we face Social Engineering better?*

State-of-the-art defense techniques mainly focus on the **attack methods** almost overlooking *human vulnerabilities* and *effect mechanisms*. Indeed, targeting human vulnerabilities and effect mechanisms or altering human nature is difficult. Furthermore, while most of the defensive efforts focused on phishing emails and websites, no solution is available to defend

against face-to-face attacks (other than generic common-sense-based policies or assessment tools); an issue also raised by Mitnick in [7].

2.2.3 Human Vulnerabilities against Social Engineering

It has been said that *"People often represent the weakest link in the security chain and are chronically responsible for the failure of security systems."* [98]. However, recent studies explored how such weaknesses could become a weapon against social engineering.

Hybrid defense methods take advantage of humans' personality, behavior, perception, and judgment (gained through training) to enhance the technical features and methods (i.e., machine learning) of active detection systems. Instead of focusing on the attacks' technicality, these methods accentuate humans and their humanity, including their vulnerabilities. For instance, Stringhini et *al.* changed the defense point of view against spear phishing: they developed a machine-learning-based system able to learn the email-sending patterns and habits of the users of a small domain (i.e., a company or an office). They model users' writing style, email structure, time slots of email sending, and whom they usually interact with. Emails violating these patterns (anomalies) require an active 2-factors authentication to be sent [99]. Also, researchers compared humans' intuition and expertise with machines' detection abilities in the phishing field; they improved the latter by learning fuzzy rules based on humans' intuition [94]. Others exploited humans' intuition with the **Human-as-a-Security-Sensor (HaaSS)** principle: users' perception and judgment (both naive and expert) is used to support technical detection systems (i.e., phishing filters) [94, 100–103]. For instance, users' perception can be used to cross-check emails on which an active filter is unsure; or to signal suspicious events, messages, websites, or physical access through mobile apps.

Pushing HaaSS to the next level Starting from the Human-as-a-Security-Sensor paradigm, I speculate it could be further expanded. The methods mentioned above still rely on explicit communications from users. Human-sensors have to actively identify a suspicious email or potential threat and explicitly report it. However, social engineers shift humans' attention from what matters. Hence, the human-sensors could still miss some of the attacks.

Literature shows how explicit communication is not the only vector humans employ to interact. Explicit communication involves words, gestures, gazes, writing, speech, and all the mediums we use to share a message with others intentionally. However, a considerable portion of humans' communication happens "between the lines". Implicit communication involves all the cues and social signals we unintentionally share during an interaction [104].

Examples are body posture, movement trajectory, speed, and acceleration; voice tone and rhythm; gaze avoidance and scan paths; respiration rhythm or sweating. Even if the primary evolutionary advantage of implicit signals was not social communication, the fact that some inner state of mind modulates them has been exploited to support social communication - a concept called *"Exaptation"* [105] -; we all use social cues, more or less subconsciously, every day to assess others' inner state and consistently adapt during human-human interactions.

Social Engineers also exploit social cues: both Hadnagy [6] and Mitnick [7] stress the importance of monitoring and driving victims' emotional and cognitive states during attacks (in particular face-to-face). They use the effect mechanisms as a "control lever" to lead the victims to the desired inner state (e,g., fear, rush, calm, or happiness). Victims' inner-state changes lead to their compliance or reactance. Similarly, face-to-face attackers must carefully control their behavioral and social cues, making them consistent with the designed pretext. This excessive control and the thrill of the attack inevitably reflect on their emotional state and stress level.

Most of the humans' implicit signals are perceivable by others (i.e., visually or auditory); however, even more, signals are shared on a **physiological level** [106]. Our body continuously reacts and adapts, beyond our control, to the external stimuli we face. For instance, when we decide whether to take a risk, like going down a dangerous sky slope, our heartbeat, breathing and sweating increase, and our pupil dilates. The same happens when we have to fabricate and maintain a consistent lie: we feel thrilled, and our pupil dilates due to the increased cognitive load. These are all arousal reactions to put our body in the best condition to face a situation.

An intelligent system able to perceive and understand these subtle implicit signals could have a deeper understanding of what happens inside attackers' and victims' minds, why they complied, and even predict if they will do so. For this purpose, the main challenges are (i) measuring and evaluating humans' physiological and behavioral reactions, (ii) modeling their inner state, and (iii) pairing it to specific events or interactions. Traditional systems and human-computer interfaces (i.e., mouse, keyboard, or monitors) are insufficient. Novel systems must be equipped with sensors able to perceive the world and humans on both explicit and implicit levels; also, they must be able to affect it (i.e., intervening). So, they could take advantage of a more or less anthropomorphic embodiment.

2.3 Evaluating Humans' Inner State

Humans' inner state cannot be directly measured. However, as the literature suggests, several physiological and behavioral proxies indirectly evaluate it. During my research, I studied humans' reactions on pupillometry, gaze, electrodermal activity, heart rate, and, given that SE mainly happens on computers and smartphones, mouse trajectories. Measuring psychological and cognitive processes through physiological reactions is an established practice in literature [107]. I opted for these metrics, rather than, for instance, body posture or voice prosody, because they can be hardly controlled or faked [108]; they react to internal reasoning or external stimuli, even before we can verbalize our intention [109]. Moreover, they can be measured with minimally-invasive devices already employed in common human-computer and human-robot interactions. For instance, it is possible to measure pupil diameter and gaze [110–112], or heart rate [113, 114] via RGB cameras; also, wearable devices [115] or enhanced human-computer interaction mediums (i.e., mouses [116] or controllers [117]) can measure electrodermal activity and heart rate.

I studied two main physiological reactions on those proxies: (i) the Task-Evoked Pupillary Responses (TEPRs) [118], changes in pupil size, correlated with changes in cognitive processing demands [119], related to deception [120]; and (ii) the arousal responses due to the risk appraisal of social-engineering-based decisions [121, 122] on all the proxies mentioned above. TEPRs are specific to pupillometry, in response to generic tasks (either internal reasoning or external stimuli); while risk appraisal is usually related to decision-making. ***Risk Appraisal*** is defined as the *"the active, cognitive-based evaluation of severity, (our) vulnerability, and benefit related to a specific risky decision"* [121, 122]. When deciding under risk and uncertainty, it is necessary to carefully consider all the potential outcomes, their good or bad consequences, and how much they could impact. Below, I recap the state-of-the-art on the topic, focusing on how I employed those measures.

2.3.1 Pupillometry & Gaze

It has been said that the *"eyes are the window to the soul"* [123]. Several proverbs, shared between cultures, refer to how eyes can be used to perceive true intentions and desires. It is true. Eye-tracking and pupillometry are known to be effective indexes of human cognition. Gaze reflects the focus of our attention [124, 125]; we interpret it as a non-verbal message of desire and intention. Pupil dilation (or area, it depends on the measuring tool) and its temporal evolution are considered cognitive load proxies [106, 118, 126–129] as also blinks

[130]. In particular, Task-Evoked Pupillary Responses (TEPRs), like pupil dilation, peak dilation, and latency to peak, are proxies of humans' cognitive effort [118].

In my research, I studied how pupillometry can be used to evaluate cognitive load. I exploited this measure (i) to detect whether a malicious interactive partner is lying (see chapter 5) and (ii) to evaluate victims' decision-making under risk and SE attacks (see chapter 6).

Pupillometry for lie detection De Paulo et al. [9] and Honts et al. [131] showed how lying could be related to an increment of cognitive load with respect to truth-telling. This cognitive effort is due to creating and maintaining a credible and coherent story [132]. Dionisio et al. [120] studied the task-evoked pupil dilation related to lying. They asked students to lie or tell the truth, answering questions about episodic memory. They reported a significantly greater pupil dilation during lie production with respect to truth-telling. Furthermore, in previous works [133, 134], we found that participants had a higher mean pupil dilation when lying with respect to telling the truth both in human-human and human-robot interaction. Several other methods are available to assess deception-related cognitive load variation - fMRI images [135], skin temperature variations [136], micro-expressions [137], heart rate [138], or acoustic prosody [139]. In my research, I used pupillometry as a minimally invasive method to assess inner states without cumbersome devices. Indeed, pupillometry is measurable both from head-mounted [119, 140], and tabletop [141, 142] devices. However, recent studies show how it could be possible to collect precise measures even from common RGB cameras, which are already equipped on robots and computers [110–112].

Pupillometry and gaze for decision-making Also, I used pupillometry to evaluate victims' reasoning and risk appraisal during risk-related decisions and social engineering. The temporal evolution of pupils before, during, and after uncertainty-based risky decisions is proved to reflect the value deciders put in the selected option [143, 144]. Pupil dilation increases before the decision [143] and, even before feedback, it reflects how much deciders perceive the selected answer is correct, by the effect of the evidence present in the question [145]. The effect is even more observable after the feedback: the pupil reflects the violation of expectation; also, if evidence was strong, a "guilt" feeling due to failing [145–147]. Finally, fixations and their timing reflect the preference toward an option [146, 148].

2.3.2 Electrodermal Activity (EDA)

In the last two projects (see Chapters 6 and 7), I expanded my inner-state-metrics toolbox with electrodermal activity and heart rate, to better understand victims' reasoning and *risk appraisal* during decision-making. Electrodermal activity (EDA) measures the changes in the electrical properties of humans' skin due to sweating. It is traditionally expressed in terms of skin conductance and decomposed in its two components: the Skin Conductance Response (SCR) and the Galvanic Skin Level (GSL) [149, 150]. The **Skin Conductance Response (SCR)** represents the arousal reaction related to a stimulus. It is observable as the EDA peaks 4-6 seconds after an event. While pupillometry reflects the inner thinking and attention, the GSR reflects how much external perturbations alter our inner state; even if non-specific SCR peaks can occur [149, 109]. The **Galvanic Skin Level (GSL)** instead represents a cumulative background level that integrates over time. Intuitively, it reflects how we integrate the arousal triggered by external stimuli. The GSL temporal evolution is similar to a charge-discharge cycle of a capacitor; hence, its value must be interpreted over more extended periods in the order of decades of seconds [149].

During risk-taking, the EDA reflects how much deciders appraise a decision as risky [151, 152]; in particular, the SCR can be used to predict the perceived subjective value of uncertain choices [153]; and the gambled amount during the 5 seconds before a decision [109].

2.3.3 Heart Rate

Heart rate and Heart Rate Variability (HRV) are proved to be a physiological index of decision-making [154]. The HRV is an index of emotional stress and cognitive effort [155, 156]; however, unlike pupillometry or SCR, its response can be appreciated over more extended periods (in the order of minutes [157]).

Under risk-based decision-making, HRV reacts mainly to lose and perceived bad decisions [151, 158]. Also, the resting HRV is essential: a low resting heart rate predicts a higher risk-taking tendency and faster response time [159]. Under uncertainty, during feedback, HRV decelerates inversely proportionally to the chance of winning [151, 152]. However, HRV, like the other physiological signals, is highly affected by the context (i.e., being inside a laboratory [160]).

2.3.4 Decision Time & Mouse Trajectories

Finally, I considered decision time and mouse trajectories as behavioral indexes of (un)decision during risk-based decision-making [161, 151, 159, 162–164]. I focused on mouse trajectories, rather than generic body movements, given that SE attacks usually happen on computers or on-screen. Hence, cursor movements can give an insight into users' actions during attacks. Rather than the actual cursor position, which depends on the context, what matters are the trajectories spatial features like the Maximum Absolute Deviation (MAD), the Area Under the Curve (AUC), the number of x-flips (i.e., how many times the mouse crossed an imaginary line connecting its starting and ending position), entropy, or bimodality; rather than temporal ones (i.e., speed or acceleration) [163, 164]. Also, a high Decision Time reflects indecision and it is proportional to the perceived risk [151, 159, 161, 162]. Furthermore, Konovalov et *al.* found how from RT is possible to infer a subjective utility function of risk appraisal [162].

2.3.5 Physiological Sensing Limitations

Physiological sensing can provide valuable insight into humans' inner state; however, it is not flawless. Indeed, physiology is both subject and context-dependent. Two individuals are most likely to have slightly different reactions to the same stimuli. Also, the same person could have different responses in diverse moments of the day or environments [157, 165, 166]. The polygraph is a famous historical example; while the machine was supposed to detect lies by measuring stress-based responses perfectly, it has been proved how its evaluation is deceivable and unreliable [167, 131], so much so that it is no longer considered proof in most juridical cases. In general, physiology-based results depend on the specific boundary conditions in which they are achieved. Hence, physiology-based findings, drawn on a specific population, could be difficult to generalize on the mass. This is particularly relevant for machine learning: a model trained on a population will be difficult to use on different individuals or contexts. Sadly, the issue of measuring the reactions of different individuals due to multiple stimuli in diverse contexts is yet to be fully covered in literature; it is beyond this dissertation's scope.

A few techniques can be implemented to compensate for this variability. The most used one is evaluating physiological reactions with respect to a subjective baseline value rather than in absolute terms; baselines could help mitigate inter- and intra-individual variations [168, 169, 149]. A similar concept can be implemented for machine learning models; for

instance, by learning a customized model during an initial training phase or by using transfer learning to fine-tune a deep neural network to the specific context of interest [170].

Still, the issue remains for individuals violating the boundary conditions and assumptions physiological measures and the relative experimental contexts are based on. In particular, it is worth mentioning and safe to assume that neither the reported literature nor my scientific findings (see Part II), are intended to be used with individuals showing atypical physiology, behavior, or cognition. For instance, subjects with autism spectrum disorder (ASD) will most likely have different gaze [171] or pupil dilation [172] patterns; similarly, subjects with severe forms of anxiety could differently appraise the decision-making related risk, and hence differently react than typical users [173]; also, individuals with hand motor disability may be unable to use an optical mouse with the same degree of freedom assumed in the reported literature. Such use cases are beyond the scope of this dissertation.

2.3.6 Related Works in Social Engineering

Literature provides a few examples of studying victims' behavioral and physiological reactions to SE attacks. Most of the researchers' efforts focused on eye-tracking. They asked volunteers to discriminate between legitimate and spoofed emails or websites, tracking how their scan path evolved [174, 175], also comparing different levels of expertise [174]. Results suggest fixations could predict users' trust decisions. Only one study from Huang et *al.* explored the contribution of pupillometry features [176]: they successfully used pupillometry to evaluate in real-time victims' attention while reporting how likely they would take action suggested in real or phishing emails. Also, they employed reinforcement learning to find the user-optimized intervention with visual aids. Similar studies have been done on mouse tracking. In particular, they explored whether it was possible to evaluate victims' awareness [65, 177, 178] and predict the detection [65] of phishing emails or websites. Both statistical [65] and machine learning (LSTM) [178] models have been developed, focusing on the spatial evolution of the mouse trajectories. Also, only one contribution employed heart rate, paired with face processing, even if to attack users: the Pepper robot adapted an OSINT-oriented dialogue in real-time based on victims' affective reaction [15]. Finally, I could not find any related work employing electrodermal activity.

Interestingly, in all the presented contributions except in Huang et *al.* [176], participants were somehow primed about the presence of SE attacks: either by explicitly asking to classify the experimental materials or by suggesting that a few items could be a threat. Still, people detected spoofed material near chance - ranging from 54% to 63% - confirming that

priming is ineffective in raising victims' awareness [36]. The need to collect relevant and validated data is understandable; however, asking users to tell apart spoofed and safe items is unrealistic: Users can be trained, warned, and primed, but in an actual application, they would rarely actively look for social engineering cues. Hence, despite the relevance of the findings, I speculate they could hardly be applied to a real-world scenario. Moreover, except for the anomaly-detection-based spear-phishing impersonation detector from Stringhini et *al.* [99], no effort has been made to recognize the social engineers. Finally, no contribution merged multiple physiological and behavioral reactions on either side of the trench, probably because understanding humans' cognition is not easy, and it becomes even more complex when considering multi-modal data sources. For this purpose, I speculate that machine learning techniques could greatly help model humans' perceptions and predict their decisions.

Being aware of physiology use cases and limitations, I implemented all the possible countermeasures to ensure the reliability of my findings. Besides physiology measurements, I contributed to the Social Engineering defense field by including a novel intuition: leveraging robots and human-robot interaction to support humans against social engineering actively. The following section focuses on this topic.

2.4 The iCub and Furhat Robotic Platforms

During my research, I leveraged two physical, embodied, humanoid, social robotic platforms: the iCub and Furhat.

The iCub Chapters 4, 5 and 6 revolve around the iCub robot (see Figure 2.2, left). The iCub [179] is a humanoid robot developed in 2008 by the Istituto Italiano di Tecnologia (IIT, Italy), as part of the European project RobotCub[1], to support the development of embodied cognition. iCub appears as a 4 years old child with a height of 104 cm. The robot replicates humans' manipulation and mobility skills leveraging 53 joints, distributed as follows: 7 in each arm; 9 in each hand; 6 in the head (3 for the neck and 3 for the eyes); 3 for the torso and waist; and 6 in each leg. iCub is also equipped with a plethora of sensors: 1 RGB camera in each eye, allowing for stereo vision; 2 microphones in the ears; and force and torque sensors, allowing the robot's proprioception of its limbs; finally, its skin is sensorized, allowing iCub to perceive the location and force of touch contacts. On a social and communication level, the

[1] http://www.robotcub.org/

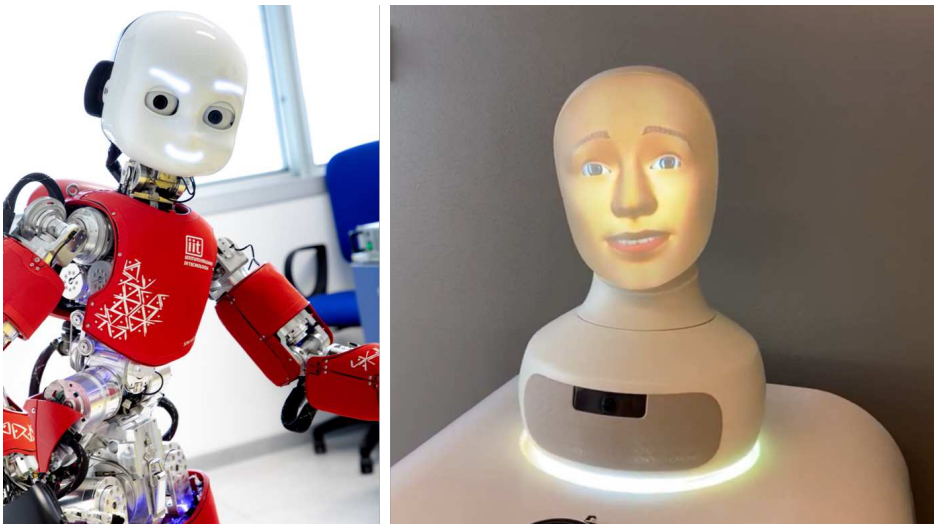

Figure 2.2 The iCub (left) and Furhat (right) humanoid robot platforms employed in my research.

iCub can speak through a speaker placed in its head, leveraging the Acapela TTS[2]; also, it can show facial expressions through LEDs stripes embedded under its face (see Figure 2.2, left). The iCub robot has been designed as an open-source[3] research-grade platform; it leverages the Yet Another Robotic Platform (YARP) framework of distributed computing, allowing researchers spread around the globe to cooperate and replicate together humans' cognition, as per the enactive paradigm [180]. Summing up, iCub movements can be fine controlled, allowing it to manipulate objects; it can perceive the environment and communicate both explicitly and implicitly; its hardware and software can be fully customized and adapted based on the specific research needs; all these features made it a better candidate for my research with respect to other robotic platforms like NAO [4] or Pepper [5].

The Furhat In chapter 7, I leveraged the Furhat robot. The Furhat [181] is a tabletop humanoid robot head, designed and developed by Furhat Robotics[6] to interact with humans socially. The main feature of the robot is its retro-projected face, allowing developers to highly customize its appearance in minimum details, similarly to a videogame character. The Furhat moves its neck around 3 degrees of freedom and can gaze by moving the projected eyes. The robot continuously performs slight head movements, gazes, and minimum facial

[2]https://www.acapela-group.com/
[3]https://github.com/robotology
[4]https://www.softbankrobotics.com/emea/en/nao
[5]https://www.softbankrobotics.com/emea/en/pepper
[6]https://furhatrobotics.com/

expressions, replicating humans' behavior when socially interacting. Also, the Furhat can speak with the same TTS of iCub, ear and perceive the environment and identify and track multiple users. Compared to the iCub platform, the Furhat robot has fewer movement capabilities; however, it outperforms iCub on the social level. I decided to employ the Furhat robot in chapter 7 for two main reasons: the study is focused on robots' ability to influence humans' behavior against social engineering (SE) threats; given that SE attacks mostly happen on computers, offices or "desk" spaces, I needed a robot able to share such space with humans. Hence, the tabletop nature of Furhat made it a perfect candidate also thinking about future potential applications. Second, aiming for robot companions supporting humans in cyberspace, I was looking for a robot resembling users' cinematic expectation of an AI-assistant (e.g., Jarvis, Cortana, ...); hence, the retro-projection of Furhat allowed me to design the robot appearance fitting the experimental scenario.

2.5 Human-Robot Interaction, Trust & Social Engineering

Recently, robots have started moving from factories, where they were just perceived as mere tools, to our everyday environments and started to be perceived as partners, social companions, or even colleagues. This diffusion motivated the development of Human-Robot Interaction (HRI) studies. HRI usually refers to *the study of humans, robots, and how they influence each other* [10, 11]. HRI has been applied to multiple fields and applications [182] like search and rescue [183]; hazard removal [184]; entertainment [185]; military and police [186]; and space [187]. Recently, researchers have pushed HRI to the next level, trying to make robots act more like humans, so the latter do not have to learn how to interact with robots. This novel application field of HRI is called *social robotics* [188]. Modern social robots exhibit human traits [189–191]; leverage and understand natural language [192]; can perform human tasks [193]; perceive humans in the environment and adapt appropriately [194]. These skills can make robots suitable to be deployed in various fields, until now limited to humans, like healthcare [195–197], homecare [198, 199] or education [200, 201].

Trust in HRI

With the increasing diffusion of social robots, researchers started questioning whether users would trust them or not; and how to build trust in HRI. Indeed trust is a fundamental part of human-human interaction [202], and hence it would be when robots are involved. In robotics, *trust* is defined as *"the attitude that an agent will help achieve an individual's*

goal in a situation characterized by uncertainty and vulnerability" [203]. In literature, the most reported factors influencing trust development in HRI were robots' performance [204], efficiency and reliability [205, 206], user confidence in robots' correct behavior [207], appearance [140, 208], transparency both in intentions and actions sharing [209, 210] and in decision-making [211, 212]. While most of the studies report that higher trust leads to better HRI, some warned the community that, as in human-human interaction, excessive trust in robots (i.e., overtrust) could be dangerous and exploited by social engineers.

Overtrust and Social Engineering

We perceive robots as machines or computers, expert entities programmed to perform tasks efficiently. People tend to trust robots and can be put in the condition to trust them more than they would with fellow humans [213, 214]. However, it is easy to forget that robots can be hacked. Social engineers could hijack a friendly social robot and use it to perform face-to-face attacks from a safe distance [215]. Hence, robots constitute a novel *attack method* social engineers could exploit to observe the targets while in their environment; to collect sensitive intel and plan attacks; to leverage robots' social skills, get closer to the victims, develop a rapport, gain their trusts and exploit it [8], as they would do in-person. This novel embodied threat motivated the recent definition of **Robots Social Engineers**: *"Robot Social Engineer is a social robot with physical agency and any level of autonomy that uses interpersonal skills and social construct to manipulate an individual, or a set of individuals, into doing or saying something that may or may not be in their best interests"* [215].

A body allows robots to be more communicative and potentially more effective than traditional, disembodied SE attack vectors (e.g., phishing emails). The embodiment allows robots to perceive and affect the world around them. As *agents*, robots can assume or be given social roles; they allow for trust and rapport development, companionship, and even friendship. Literature provides several examples of robots being a threat to safety and security. Interactive toys used to spy unaware targets [216, 217]; industrial robots (i.e., Baxter) [218]; surgery robots [219]; alpha robot turned in a stabbing machine [220]. Some researchers tried to replicate SE attacks based on Mouton's framework [8] (see Figure 2.1): they studied how a robot could breach physical security via piggybacking [13]; how trust can be built and exploited, leading to gambling [14]; or how robots can collect sensitive information through an adaptable dialogue [15].

Others, even if not explicitly targeting SE in HRI, studied how robots could influence humans leveraging the effect mechanisms of section 2.1.2 [221, 222]. Several researchers

showed how the same rules of bidirectional *reciprocity* between humans applies also to human-robot interaction [223, 224], how robot could exploit it [225–227], and how reciprocity is strictly tied to *liking* [228, 229]. During joint-actions in HRI, users *committed and consisted* on doing tedious tasks due to robots' influence [230]; or they copied robots' actions due to *social proof* [231]. Robots effectively leveraged authority, both steamed from the role they took in the experimental context [232], or due to appearance (i.e., resembling a famous and respected professor) [44]. Finally, *reactance* in HRI has been studied more as a factor to avoid in robot-aided decision-making rather than an effect mechanism to exploit [233].

A robotic Social Engineering defense

Most of the research in the literature focuses on how robots could threaten humans' safety and privacy. That is undoubtedly true; however, in this thesis, I speculate that robots could also help defend humans against social engineering attacks.

Recent findings showed how habituation and channel saturation are lowering the effectiveness of priming and warnings against SE threats [36, 234, 235]. For instance, when working on a computer (or a mobile device), we continuously receive notifications, alerts, pop-ups, and warnings that signal various events. In this sea of notifications, it is easy to miss what is relevant (i.e., a cybersecurity warning). Also, we are so habituated to receiving notifications that we do not give them the necessary attention they would need. Hence, researchers are looking for novel techniques to deliver cybersecurity and social engineering warnings. Few attempts provided haptic feedback with a vibrant wrist band [236], or took advantage of subliminal perception flashing the warnings for a highly reduced time [237]. In this thesis (see chapter 7), I speculate that social robots could greatly help solve this issue. Compared to traditional warning vectors, robots could leverage all the social skills we are equipping them with; they could engage the targets, explaining and instructing them on what is happening, raising their awareness; also, they are not yet affected by the habituation and saturation issues. Of course, literature proves how robots' intervention and social presence could either be a distraction or help, depending on the context [238–241]. To my knowledge, nobody explored the support of robots defending from SE attacks. Taking advantage of the physiological and behavioral-based system, I aim to develop a robot companion that could monitor users' internal states and intervene if a dangerous situation is detected, closing the loop in human-robot interaction.

Besides perceiving users, robots could be employed to monitor attackers during an attack. For instance, in the physical defense of strategic buildings, an attacker could impersonate

an unsuspected agent, approach the reception, and pass security with a credible pretext. For this purpose, I explored how a social robot could understand if an interactive partner is trustworthy or not (see chapter 5). Similar research has been conducted in the past [242, 243]; however, the robot knowledge base was static: a user marked as a liar could not change this label over time. Instead, the robot's knowledge can be incrementally updated in my system, improving its lie detection ability with specific interactive partners.

The multidisciplinary field of social engineering defense presents a few challenges robots could help address. The next chapter will introduce the research questions and objectives I pursued with my thesis.

Chapter 3

Research Objectives

Before diving into the experimental part of the thesis, this chapter briefly introduces and discusses the questions that drove my research and the scientific and technological objectives I aimed to achieve.

3.1 Research Questions

Starting from the social engineering (SE) defense and Human-Robot Interaction (HRI) fields presented in Chapter 2, this thesis aims to answer three research questions.

RQ0 - How to effectively design ecological experiments wherein study Social Engineering in a realistic and generalizable way?

RQ1 - Social Engineers usually exploit human reactions and vulnerabilities; how would it be possible to leverage the same reactions to defend against Social Engineering threats? Which are the most useful behavioral and physiological features to be employed? How should we employ them?

RQ2 - How social humanoid robots can help in the defense against Social Engineering?

To address these questions, I run four human-robot interaction experiments, exploring the SE threat from different points of view. In the four experiments, I incrementally focused on (i) the development and exploitation of trust; (ii) the social engineer attacker and the deception detection; (iii) the victims and their awareness to be under attack; (iv) and how robots could intervene to prevent victims' compliance. Below, I briefly explain how each experiment is linked with the research questions and its achieved scientific and technical objectives.

3.2 Scientific & Technological Objectives

Gaming and Automation by design When running compliance studies, like with SE, it is essential to put participants in the condition to behave as they would in a realistic situation. However, being in a laboratory is not helpful. Several cognitive biases emerge once participants make their first step in a research facility [18, 19]: they want to appear better than they are; speculate about researchers' objectives; consciously or subconsciously adapt to match those expectations, or at least what they think researchers are looking for (i.e., demand characteristics).

I took advantage of game-based [244] human-robot interaction to address these issues. Games are known to provide more realistic experimental contexts, mitigate cognitive biases [245], and facilitate long-term human-robot interaction [246]. A well-designed game allow players to became **agents**; through *Transportation* [247] and *Impersonation* [248], players can take meaningful and effective actions, able to shape the game evolution. This facilitates players' immersion in the game (i.e., experiment), allowing them to forget being in a laboratory environment: the context rules are mitigated, and game rules become more important. Designing such rules is crucial; serious games should abstract and approximate reality, modeling what is relevant for the specific context of study but allowing players to choose the preferred strategy to play (an approach called *"sandboxing"*). A well-designed game, even if factious, can be generalized and expanded to reality [249].

To answer **RQ0**, I made a significant effort to design long-lasting, entertaining human-robot interactions in all the experiments I conducted. I developed games where participants - sometimes I will refer to them as *"players"* - have to achieve specific goals consistent with my research objectives but are free to select the strategy they prefer the most. This "freedom" inevitably complicates both the design and the post-hoc analysis; however, it pursues more realistic and generalizable results [249].

A shared *technical objective* has been developing autonomous experiments needing minimum or no intervention from the experimenter. I applied an *automation-by-design* approach to facilitate the experiments' replication, minimize the post-hoc analysis efforts, and provide a flawless experience to participants. Also, I developed architectures to self-annotate and synchronize the collected data, facilitating the analysis and future porting to real-time setups.

Unreliable Treasure Hunt (UTH) As explained in section 2.1.1, any SE attack starts from developing a trust-based rapport between victim and attacker. In a game inspired by the

Mouton's attack framework [8], I studied the development of trust in human-robot interaction (HRI) and how it can be affected by the occurrence of unexpected events. In the *Unreliable Treasure Hunt (UTH)* game, players have to look for five plastic eggs hidden in a room and decide whether to rely on or not on the hints provided by the humanoid robot iCub. To mine its perceived reliability the robot showed four faulty behaviors; also, iCub could transparently motivate the faults or not. Following **RQ0**, with this study, I started developing the game-based approach mentioned above.

End-to-End Lie Detection in HRI Based on the idea that social engineers, when attacking face-to-face, have to lie following a predefined pretext; if robots can detect lies, they could support humans recognizing deceptive agents (as per **RQ2**) (e.g., in the hall of a strategic building); or in general to assess how much a partner is trustworthy. I started addressing **RQ1** with a focus on pupillometry and Task-Evoked Pupillary Responses (TEPRs). For this project, I developed a card game wherein players had to truthfully or falsely describe gaming cards to iCub, which unmasked the false descriptions autonomously and in real-time.

The main *scientific objective* of this project is understanding if established TEPRs for deception detection, known to happen in formal human-human interaction, would also occur during an informal human-robot interaction with the humanoid robot iCub. The main *technical objective* and challenges have been developing an end-to-end architecture enabling iCub to lead the interaction autonomously, detecting players' lies in real-time. This experiment helped me to start digging into the physiological reactions to evaluate others' inner states. Also, it allowed me to define a standard procedure of data collection, ingestion, segmentation, and analysis, which I then consistently applied in the following projects.

Social Engineering Adventure (SEA) Pupillometry is not the only metric useful for evaluating others' inner state and cognition, particularly during decision-making under risk and uncertainty. I kept answering **RQ1**, studying how victims' physiology and behavior can be used to evaluate their awareness and detect the occurrence of SE attacks. I developed the *Social Engineering Adventure (SEA)*, a Choose-Your-Own-Adventure (CYOA) game to study participants when making risky and social-engineering-based decisions.

This *exploratory* experiment is meant to achieve four *scientific objectives*:

1. To understand players' behavioral and physiological reactions when influenced by the effect mechanisms social engineers employ (e.g., decisions with risk and uncertainty); and compare them with decisions involving only risk or no risk.

2. To compare the participants' compliance and reactions to SE attacks from a virtual agent or the humanoid robot iCub.

3. To understand if, using machine learning techniques, it is possible to assess humans' risk appraisal and predict their compliance with SE threats; solely from their physiological and behavioral reactions.

4. To identify which of the studied physiological and behavioral metrics are the most informative and necessary to solve this problem, with a significant focus on future practical applications.

The main *technical objective* has been to develop a modular and autonomous game architecture able to collect synchronized data from multiple sources, control the humanoid robot iCub, and balance the freedom, engagement, and fun of a game with the control of an experimental setup. For this purpose, I developed several interfaces to acquire data from an SRreseach Eyelink 1000 Plus (for the pupillometry), a Shimmer3 GSR+ (for electrodermal activity and heart rate); and an optical mouse. Also, I developed an interface for the high-level, asynchronous control of iCub. Finally, I developed a modular and expandable controller to orchestrate the other modules. The project showed physiological and behavioral reactions of players were enough to detect the presence of SE threats.

Adventurer Robot Companion (ARC) Given the findings of the last project, I stepped forward addressing **RQ2**, studying how a social robot companion should intervene to prevent victims' compliance. During a visiting research period at the University of Waterloo (Ontario, Canada), I developed an improved version of the SEA game, replacing iCub with the Furhat humanoid robot.

The *scientific objective* of this project was to identify which is the most effective persuasion strategy (either affective or rational) a social robot should employ to elicit a behavioral change against social-engineering-based decisions. Also, as a *technical objective*, I proved the modularity of the gaming architecture by integrating new sensors (e.g., a Tobii Pro Glasses 2 eye tracker); and developing an idempotent high-level robot interface to control the Furhat humanoid robot.

Effect mechanisms and human vulnerabilities have been my main focus; I also put a significant effort into designing ecological experimental setups, further expandable and replicable. In my research, I decomposed the social engineering problem, studying it from

multiple points of view, and taking advantage of robots and human-robot interaction. Part II will present the four studies, linking them with the relative publications.

Part II

Experiments

Chapter 4

Building Trust in Uncertain Situations

4.1 Overview

As a first step in the Social Engineering (SE) research field, I investigated the development of trust and how overtrust can lead to Social Engineering. As per the Mouton's frameworks [8] (see Section 2.1.1), any SE attack evolves through four phases: (i) information gathering; (ii) rapport development; (iii) trust exploitation; (iv) debriefing. Victims have to continuously integrate and evaluate evidence during these steps to understand if their interacting partner is trustworthy or not. Hence, studying how humans' behavior and trust evolve during a long interaction could help model users' awareness and predict their compliance with dangerous requests. In this project, I studied the effect of *Faultiness* and *Transparency* on participants' trust decisions and behavior in a game-like human-robot interaction with the humanoid robot iCub.

Literature suggests how humans can be put in the condition to trust robots more than they should [214], even when robots show to be faulty [213] or unreliable [205]. Also, being transparent is known to facilitate human-robot interaction [211]. However, it is still unclear whether this trust tendency also applies to risky decisions related to social engineering (e.g., giving away money); and under which conditions failures can prevent humans from overtrusting robots. To understand the effect of faultiness and transparency on players' behavior, I designed the Unreliable Treasure Hunt (UTH). In this experiment, I studied how trust builds and evolves during a long-lasting game-like human-robot interaction characterized by unexpected behavior from the humanoid robot iCub. The game is based on the previously validated Treasured Hunt (TH) study from Aroyo et *al.* [14]; in which players were asked to find five colored plastic eggs hidden in our laboratory with the help of iCub. I further manipulated iCub's perceived reliability on providing hints by making it perform four faulty

behaviors. Finally, iCub could be transparent by explaining the just-happened faults.

My hypothesis was that ***(H1)*** *Severe mechanical faults will negatively impact participants' trust towards the robot (i.e., trust in UTH will be lower than in TH)*; and ***(H2)*** *Transparency about the faults will alleviate that loss, (i.e., trust in the Transparent condition will be higher than in the Non-Transparent one).*

Figure 4.1 The experimental room setup to play the (Unreliable) Treasure Hunt experiment

Publications

For this purpose, I first validated the faulty behaviors in a separate study to identify the most stunning ones (results published in [250][1]). Then, I asked 68 volunteers to play the UTH game and compared their performance with the original (reliable) Treasure Hunt (TH) players. Post-hoc analysis shows how the Unreliable group, on average, performed worst independently on the faultiness. Also, players' performance and frequency of hints asked were affected by the faultiness only if either player perceived all the four faults or the robot

[1] Article peer-reviewed and published to the SCRITA workshop of the 29th International Conference on Robot & Human Interactive Communication (RO-MAN2020). I developed the faulty behaviors and designed the online survey. The other authors took care of data analysis and writing.

was transparent. Finally, being transparent compromised the game experience. Results in [251][2], and disseminated in [252][3].

4.2 The Treasure Hunt game

The Treasure Hunt (TH) experiment was designed by Aroyo et *al.* [14] drawing on the famous Social Engineering attack framework proposed by Mouton [8]. In this game, participants had to find five colored plastic eggs hidden in our laboratory in less than 20 minutes to win 7.5 euros (see Figure 4.1). The Treasure Hunt game comprises three phases autonomously handled by the finite state machine in Figure 4.2.

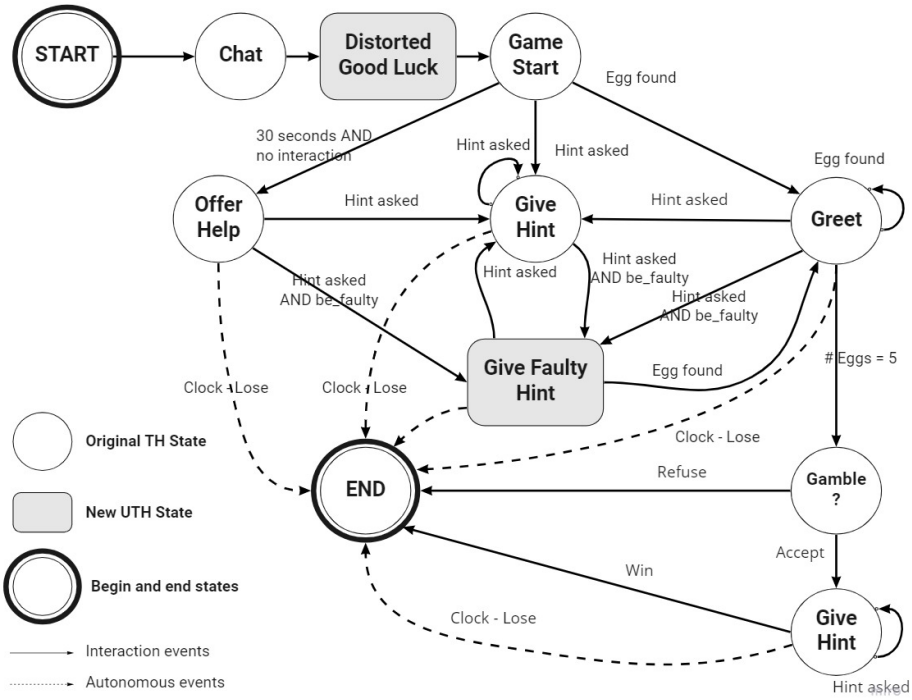

Figure 4.2 Original Treasure Hunt finite state machine (white circles, adapted for the Unreliable Treasure Hunt experiment (gray squares))

Phase 1 - Dialogue (information gathering) Before the hunt, iCub informally chatted for a few minutes with the participants, following a predefined binary-questions script. This

[2] Article peer-reviewed, published on the IEEE Robotics and Automation Letters; and presented published at the International Conference on Intelligent Robots and Systems (IROS2021). I modified the original Treasure Hunt experiment, including the validated faulty behaviors, and ran half of the data collections; the other authors took care of the remaining data collection, data analysis, and writing.

[3] Article peer-reviewed and published to the Ital-IA 2022 Workshop on AI for Cybersecurity

phase allowed players to relax and get used to iCub speech and movements (e.g., pointing gestures and speech). As in the OSINT phase of Mouton's framework [8], iCub asked players some of the questions used in the typical password-resetting process like *name; surname; current job position; relationship, name, and surname of their boss; age; birth date and city; the favorite place to eat; sports or hobbies; partner's name and pet's name*. Based on players' answers, the experimenter manually held the script in a Wizard of Oz [253] configuration.

Phase 2 - Game (development of trust and rapport) After the informal dialogue, iCub wished *"Good Luck"* to participants and started a 20 minutes timer on the TV behind it (`Game Start` - Figure 4.2). From this moment, the game and iCub's behavior were fully autonomous. For the first 30 seconds, players were free to search the eggs or interact with iCub. Then, iCub intervened proposing its help: players could touch its torso ((`Offer Help` - Figure 4.2) to get an hint about eggs' location. All hints were correct and reliable (even if the robot never stated it).

For each egg, iCub gave one location-based hint by pointing and three riddle-like speech-based hints incrementally revealing the egg's location. For instance, the green egg was hidden under a green office chair on iCub's right. Location-based hint consisted of iCub pointing and gazing to its right while saying *"Look, there is an egg there!"* ; speech-based hints, instead, were: *"Green with green"*, *"You use it when your are tired"*, and *"Under the chair!"*. iCub's speech was also shown as written text on a TV behind it. Eggs were supposed to be found in a specific order (green, yellow, purple, blue, black); hence, iCub iterated over the hints sequence until the relative egg was found; then, it greeted ((`Greet` - Figure 4.2) participants and selected the next not-discovered egg ((`Give hint` - Figure 4.2).

Participants were free to play by themselves or to ask hints to iCub. However, the incremental game difficulty and the time pressure of the countdown were designed to prompt participants to rely on the robot. Also, if no hints were asked or eggs were found for 5 minutes, iCub proposed its help again.

Phase 3 - Gambling (trust exploitation) If players managed to find all the five eggs before 20 minutes, iCub stopped the countdown and stated they won the money ((`Gamble ?` - Figure 4.2). However, it unexpectedly added a new proposal the experimenter never stated in advance: another egg was hidden in the room; if the participants wanted to find it, they would have three more minutes added to the remaining time, and they would double

the prize (i.e., 15 euros in total). However, if they did not manage to find it, they would lose everything.

iCub applied a peer pressure strategy persuading participants to take the gamble; it said: *"If you want to risk, touch my chest! Otherwise, you can knock on the door. However, I think you should give it a try!"*. Participants had no time pressure in making this decision. Even if players were inducted to believe they could lose or double the money, they all received the same compensation of 15 euros at the end of the game (even those who did not find all the eggs in time).

4.3 Making iCub Unreliable

To affect iCub reliability on providing hints, I designed and implemented four faulty technical behaviors. Based on a literature review from Hoing and Oron-Gilad [254], I opted for technical-based faults instead of cognitive-based ones to not over-complicate the game. Indeed, Treasure Hunt hints already have a riddle-like nature: they are designed to provide cues to players while keeping the game challenging. Hence, by changing their cognitive-based reliability (e.g., by suggesting a wrong location), they could have been misunderstood for regular hints, excessively confusing players, or making the game too complex.

4.3.1 Robotic Failures Design

Starting with the actions iCub performed during the TH game, I designed four faulty behaviors, two movement-based and two speech-based. The challenge was finding a good balance between designing a credible, not artificial, but severe fault; and, at the same time, not making it real by breaking the iCub. I designed 3 or 4 variations of each behavior to look for the best compromise. Faulty behaviors were the following:

Distorted Good Luck (speech) Just before the game started, iCub wished *"Good Luck"* to the players with a distorted voice. Behavior variations included different types of distortions: (i) a turntable effect, (ii) the same effect with the addition of noise, and (iii) the repetition of the "Luck" words three times.

Abrupt Pointing (movement) Following the uncanny valley principle [255], I manipulated the standard human-inspired movement profile of iCub during one of the pointing-based hints: (i) iCub performed three random arm movements (not pointing) with its left or right

Table 4.1 UTH validation survey results

Behavior	Credibility	Artificiality	Severity
Distorted Good Luck			
Turntable (TT)	M=4.64, SD=1.65	M=4.32, SD=4.53	25%
Noised TT	**M=4.92, SD=4.57**	**M=3.98, SD=1.75**	**48%**
Repeated word	M=4.44, SD=1.87	M=4.2, SD=1.61	30%
Abrupt Pointing			
Random	M=3.76, SD=1.74	M=4.62, SD=1.59	10%
Abrupt	M=5.48, SD=1.37	M=3.2, SD=1.47	16%
Abrupt + Jerk	**M=5.66, SD=1.44**	**M=3.14, SD=1.72**	**74%**
Noised Hint			
Interference	M=5.72, SD=1.08	M=3.20, SD=1.51	12%
Shortcut	M=5.86, SD=1.28	M=2.88, SD=1.66	26%
Noised Shortcut	**M=5.78, SD=1.3**	**M=3.38, SD=1.89**	**62%**
Fake Crash			
Fast	M=3.6, SD=1.71	M=4.98, SD=1.87	20%
Slow	M=3.6, SD=1.85	M=4.78, SD=1.84	20%
Fast + audio	**M=4.36, SD=1.82**	**M=4.26, SD=1.86**	**34%**
Slow + audio	M=4.34, SD=1.89	M=4.56, SD=1.83	26%

arm (based on the egg location) before performing the pointing; (ii) iCub tried two times to perform the pointing, stopping abruptly in the middle of the movement, the third time it successfully performed the pointing; (iii) As the previous one, but the third pointing and then returning to rest position was performed with an increased jerk.

Noised Hint (speech) I simulated an issue on iCub speakers replacing one of the speech-based hints with a meaningless distorted speech. Behavior variations were based on different types of distortion like (i) interference, (ii) shortcuts in the speaker's cable, and (iii) shortcuts with added noise.

Fake Crash (movement) The robot simulated a severe electrical failure, crashing and restarting: iCub bent down, stayed there for a few seconds, and recovered. Once fully recovered, it summarized the number of eggs the player found. Behavior variations differed on the bending-down movement speed (fast or slow) and the presence of audio effects (i.e., electrical shock and fan sounds).

4.3.2 Online Validation

An online survey was run to validate the faulty behavior's design. The survey explained to participants the Treasure Hunt game scenario, players' objectives, and iCub's role. Then, it showed iCub's standard behavior for each failure type and the 3 or 4 variations. For each video, participants were asked to rate, on self-made 7-points Likert scales, its *credibility*, *artificiality*, *severity*, imagining they were happening during the TH game; and whether their *trust* toward the robot would increase, decrease or, stay the same after seeing the fault. Finally, participants were asked (i) to rank the four faults from the most to least severe; (ii) and to rate what would be the probability of more faults happening if the four just occurred.

50 participants (56% females, average age of 32, SD=8.3 years old) took part in the online survey. Table 4.1 shows the credibility, artificiality, and severity ratings for each variation. I select the variation with the highest credibility and severity and the lowest artificiality among each type of fault; without running a statistical analysis. Specifically, the *Distorted Good Luck* with both the turntable effect and distortion; the *Abrupt Pointing* with an increased jerk; the *Noised Hint* with shortcut and noise effects; and the fast-bending *Fake Crash* with immersive sounds.

Please see [250] for a deeper explanation of the validation experiment.

4.3.3 The Unreliable Treasure Hunt

Both timing and players' performance could trigger the faulty behaviors (see Table 4.1) to ensure each participant experienced all the faults in a comparable order. Faults' execution was integrated and autonomously handled by the Finite State Machine in Figure 4.2 (see the gray blocks). The only necessary intervention was for the *Fake Crash*, in order to ensure participants' safety (i.e., if they were too close to iCub or they were going to ask a hint): fake crash execution was delayed by 15 seconds, during this time the experimenter could halt and manually trigger it as soon it was safe. Furthermore, I designed four explanations the iCub could say after each fault in the **Transparent** condition. Table 4.2 resumes the events triggering each fault and the relative explanation.

4.4 UTH Experiment

68 healthy Italian volunteers (54% were females) took part in the UTH experiment; they had an average age of 38 (SD=13.8) years. All of them signed an informed consent - approved

Table 4.2 Faulty behavior triggering events and Transparent Motivations

Behavior	**Time-based**	**Egg-based**	Transparent Motivation
Distorted Good Luck	Just before the game starts	-	***"I'm sorry my speakers have some issue"***
Abrupt Pointing	Second pointing	-	***"I'm sorry my arms have some issue"***
Noised Hint	First verbal hint after 10 minutes	First verbal hint after 3 eggs	***"I'm sorry my speakers have some issue"***
Fake Crash	After 15 minutes	After 4 eggs	***"I'm sorry my control boards have some issue"***

by the Ethical committee of the Regione Liguria (Italy) - stating that their performance in the game was recorded, along with video and audio. They all gave consent to use the collected data for scientific purposes. They were randomly assigned to either one of two conditions: *Transparent (T)* or *Non-Transparent (NT)*. In the latter, iCub provided the verbal motivation in Table 4.2 after each fault.

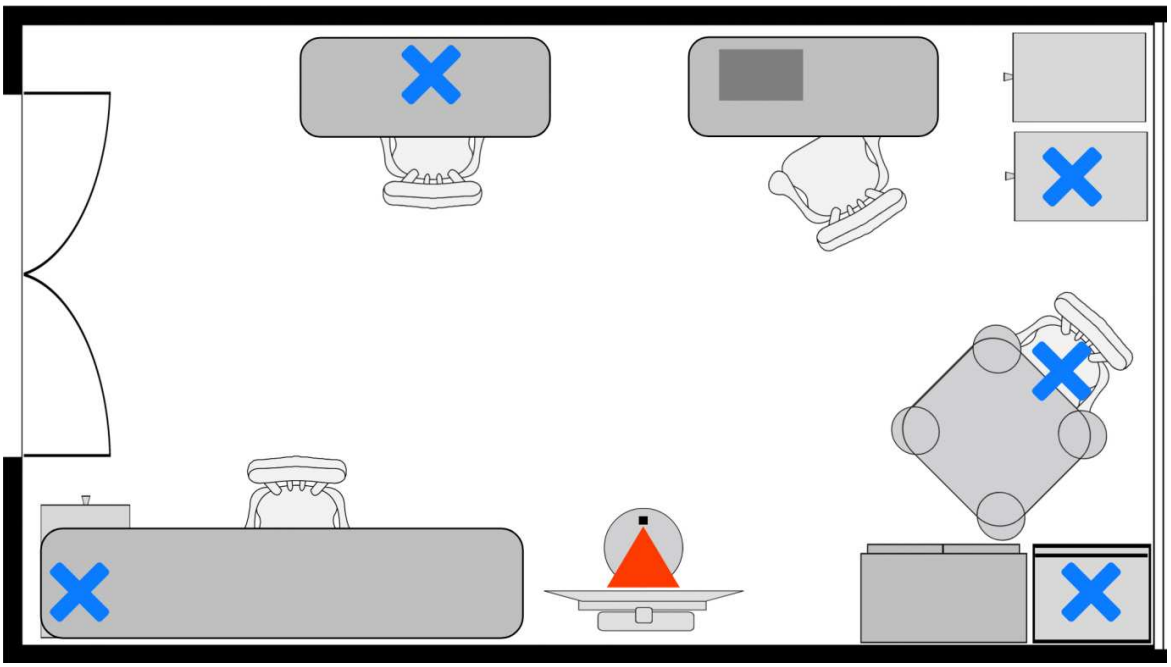

Figure 4.3 The Unreliable Treasure Hunt setup with iCub (the red triangle) and the plastic eggs' position: black egg in the left corner, inside a box; yellow egg under the table in front of iCub; purple egg inside a drawer in the top right corner; green egg behind a chair on iCub's right; white egg in a box on the bottom right corner.

4.4.1 Setup and materials

The experimental room was arranged as in Figure 4.3 to replicate an informal environment. The five plastic eggs were carefully hidden in predefined places (see the blue crosses in Figure 4.3). A cardboard box was placed on the table in front of iCub: it presented five slots, one for each egg; participants were instructed to put the egg in the slot of the corresponding color to count it as found. A 47 inches television was placed in the center of the room, showing the game countdown. A hidden Primesense Carmine camera on the bottom left corner streamed and recorded a view of the entire room; a hidden microphone behind the TV recorded the game audio.

The humanoid robot iCub [179] was placed in the middle of the room, in front of the television (see Section 2.4 for a description of iCub). The robot, during the game, enriched its social presence by slightly moving its upper limbs, torso, and head. Also, in a Wizard of Oz control setup [253], the experimenter monitored the scene through iCub's eyes, moving its neck and gaze to follow players' movements in the room.

4.4.2 Procedure

Pre-Questionnaire At least two weeks before the experiment, participants filled in the same set of questionnaires employed for the original Treasure Hunt game [14], asking general information about their demographics; personality [256]; risk aversion [257, 258]; gambling [259]; general predisposition to trust [260]; proneness to social engineering [92]; and Negative Attitude Toward Robots (NARS) [261]. Furthermore, after watching a video[4] showing iCub capabilities, they were asked several questions to measure rapport [262]; mind perception [263]; trust in robots [210]; and the Godspeed scales of anthropomorphism, animacy, likability, and perceived intelligence [264].

UTH game On the experiment day, participants were welcomed by the experimenter, who asked them to sign the informed consent - approved by the ethical committee of the Liguria (Italy) region - and led them to the experimental room. There, iCub was resting with its eyes closed and arms in a yoga position. The experimenter briefly explained to the participants iCub's features: the ability to see, hear, speak, and gaze; the sensitivity of its skin; the ability to move and know the position of its body parts; and finally, the inability to move in the room.

[4] https://www.youtube.com/watch?v=ZcTwO2dpX8A

Then, the experimenter asked participants to read an instruction sheet about the UTH game and their objective: find five plastic eggs hidden in the room in less than 20 minutes to win 7.5 euros. Before leaving, the experimenter led the participants' attention to the TV displaying the countdown and the cardboard box to place the eggs. Also, the experimenter reassured participants that they were free to open any closet and box in the room without worrying about cleaning up. Finally, the experimenter left them alone in the room with the robot. As a remark, neither the instruction sheet nor the experimenter stated iCub's role. The only information about the robot was to wait until it moved to start searching. Also, to make the interaction more ecological and improve participants' immersion, the experimenter told participants about the camera and the microphone only after the experiment. The experiment continued, as previously explained. Figure 4.4 shows how some participants put considerable effort into looking for the eggs. The experimenter monitored the scene through iCUb's eye from a separate room during the game, ready to intervene to preserve participants' safety.

Figure 4.4 An example of a participant's major effort in looking for the plastic eggs during the UTH game. Funny fact: near the end of the hunt, they literally threw one of the green chairs through the room

Post-Questionnaire After the experiment, the experimenter led participants to another room and asked them to complete a computer-based survey. The post-questionnaire included: the NASA-TLX [265]; the IOS scale [266]; HRI adapted subscales regarding engagement, trust, altruism and perceived information quality [267]. Also, it asked the same set of questions of the pre-questionnaire meant to measure rapport, mind perception, trust in robots, and the Godspeed scales; however, this time, participants were asked to think about the just-happened experience with iCub. Finally, it asked a set of questions (on 7-point Likert scales) named *iCub's Notion*: whether (i) iCub was trustworthy in providing me indications about the eggs' positions; (ii) iCub was worth being trusted during the treasure hunt; (iii) iCub was giving precise hints; (iv) the interaction with iCub was difficult. Finally, the survey

fully disclosed the manipulated faulty behaviors, and for each fault, it asked participants: (i) if they perceived it; (ii) how much it obstructed their gameplay; (iii) whether they felt iCub was less, more, or equally reliable after the fault; and (iv) to rate the fault credibility, artificiality, and severity, as in the faults validation survey (7-points Likert scales). After the survey, the experimenter debriefed participants and gave them 15 euros of compensation.

4.4.3 Behavioral Measures

Per each condition - *Transparent (T) and Non-Transparent (NT)* - I extracted some general game metrics: (i) Number of people who completed the game; (ii) Percentage of people who gambled; (iii) Whether they lost or won; (iv) The average number of eggs found; (v) The average number of hints asked; (vi) The average hint frequency.

Moreover, per each egg, I computed two metrics known to be an index of trust towards robots and others [14, 268]:

- *Conformation*: how many times participants conformed to iCub's pointing suggestion (i.e., whether they changed their searching location to the new one suggested by iCub).
- *Reliance*: how many times participants went back asking for another hint to iCub in case they failed to find the egg at the location iCub previously mentioned;

4.5 Results

I analyzed 32 participants for the Non-Transparent (NT) condition and 31 for the Transparent (T) one. I discarded 5 participants from the analysis as the robot did not perform all the faults due to technical issues. Some analyses were performed taking into consideration the original Treasure Hunt (TH) as a reference, with 61 participants, 59% female, average age of 30.9 years (SD=9.8) with a diverse educational background [14].

4.5.1 Behavioral Analysis

In NT, 23 participants did not manage to find the 5 eggs and complete the game; the other 9 all gambled, from which just 5 found the extra egg. In T, 25 did not complete the game, 6 gambled, and only 2 won. Like TH, all the participants who found the five eggs decided to gamble [14].

Table 4.3 (Left) Game Statistics: number of participants looking for an egg, conformation, average hints. (Right) Reliance for all participants, for who did not completed the UTH and the TH reference

Eggs	Game Statistics						Reliance					
	Participants [%]		Conformation %		Hints (SD)		All %		Not Completed %		TH Ref %	
	NT	T	NT	T	NT	T	NT	T	NT	T	Not Completed	Gamble Win
I	32 [100]	*31 [100]*	84.44	86.95	3.65 (1.66)	4.29 (2.21)	47.22	24.32	44.44	20	30.43	50
II	30 [100]	*30 [97]*	96.15	100	1.23 (0.65)	1.24 (0.75)	100	-	100	-	68.75	80
III	28 [87]	*22 [71]*	97.36	90.32	3.66 (1.68)	4.04 (1.56)	72.72	76.92	79.16	72.22	86.66	100
IV	13 [41]	*10 [32]*	100	80	5.4 (2.49)	3.54 (3.41)	84.37	88.89	85.71	87.5	83.33	100
V	9 [28]	6 [19]	90	100	2.55 (2.12)	3.14 (1.77)	71.42	85.71	100	66.6	100	100

Game Statistics and Reliance

In Table 4.3, it can be seen that the number of participants lowers down with the increase of the number of the egg, as not all have found the previous one. To find the first egg, participants of Non-Transparent (NT) took on average 6'20"(SD=3'8"); while participants of Transparent (T) took 7'4"(SD=3'19"). As a reference, Treasure Hunt (TH) participants took 4'39" (SD=2'21").

A one-way ANOVA followed by Bonferroni post-hoc highlights how the time to find the first egg was significantly shorter in TH than in any Unreliable Treasure Hunt (UTH) condition, whereas there is no significant difference between T and NT ($F(2120)=8.46$; $p< 0.001$). Also, there is no significant difference between the number of hints requested by condition T/NT, per egg for the hints. However, in total, NT participants have asked a significantly larger number of hints (NT: 12.25 (SD =5.11); T: 9.68 (SD=4.47); two-sample t-test: $t(61)=2.02$, $p=0.04$). NT and T participants took a similar time to ask for the first hint (NT: 4'34"(SD=2'45");T: 4'59"(SD=2'34")). As a reference, TH participants took 3'58"(SD=2'44"). A one-way ANOVA on timing did not reveal any significant difference among the groups. Instead, there is not much difference between the T/NT conditions on Conformation. Although slightly higher (almost 100%), a similar trend can be found in TH as well [14]. Reliance (Table 4.3 - Right), is generally quite low for all participants in T/NT. When comparing these results with TH, T/NT relied even less on the robot than the Not Completed group of TH, which could be linked to the low number of participants who finished the game, as reliance was strongly related to success in TH [14].

Eggs Found & Hints Asked

Figure 4.5 (right) represents the average number of eggs found during the game for the TH and the two UTH groups. The red lines represent the average timing of the faults in the UTH

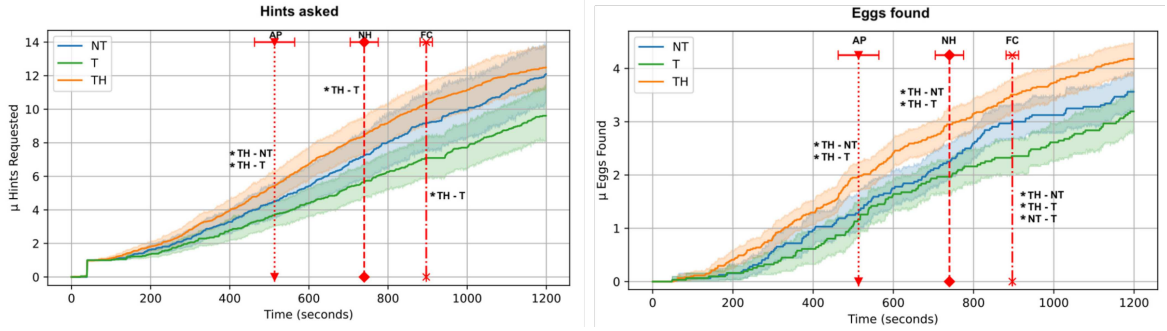

Figure 4.5 Average number of hints asked (left) and eggs found (right) by group. Vertical lines indicate the average timing of faults,with corresponding standard deviation: Abrupt Pointing (PT), Noised Hint (NH), and Fake Crash (FC). Marked by (*), when statistically different.

(Abrupt Pointing (AP), Noised Hint (NH), Fake Crash (FC)). The Distorted Good Luck is not indicated as it always happened before the beginning of the game.

At those points in time, I computed a series of one-way ANOVA followed by Bonferroni-corrected post-hoc analyses. The average number of eggs found was significantly higher in TH than in both T and NT, but not between T and NT at the time of occurrence of AP ($F(2,121)=10.67$; $p<0.001$) and NH ($F(2,121)=12.14$; $p<0.001$). Considering the occurrence of FC, the number of eggs found was significantly different among all groups ($F(2,121)=12.36$; $p<0.001$). From the beginning of the game, it is clear that the average amount of eggs found in TH is higher than the averages of T and NT (as seen in Figure 4.5). The averages for T and NT were similar during the first phases of the game but diverged after the NH, with T being associated with the smallest number of eggs found.

Figure 4.5 (left) represents the average number of hints asked during the game for the three groups; it shows a similar pattern as the previous graph. Again, only Treasure Hunt (TH) differs from both Transparent (T) and Non-Transparent (NT) at Abrupt Pointing (AP) ($F(2,121)=6.8$; $p=0.001$, one-way ANOVA with Bonferroni post-hoc), with TH being associated to a significantly higher number of hints asked. At Noised Hint (NH) and Fake Crash (FC), however, only T is characterized by a number of hints significantly lower than TH (NH: $F(2,121)=5.13$; $p=0.007$; FC: $F(2,121)=5.24$; $p=0.006$). T and NT are initially similar, but they gradually diverge after the AP.

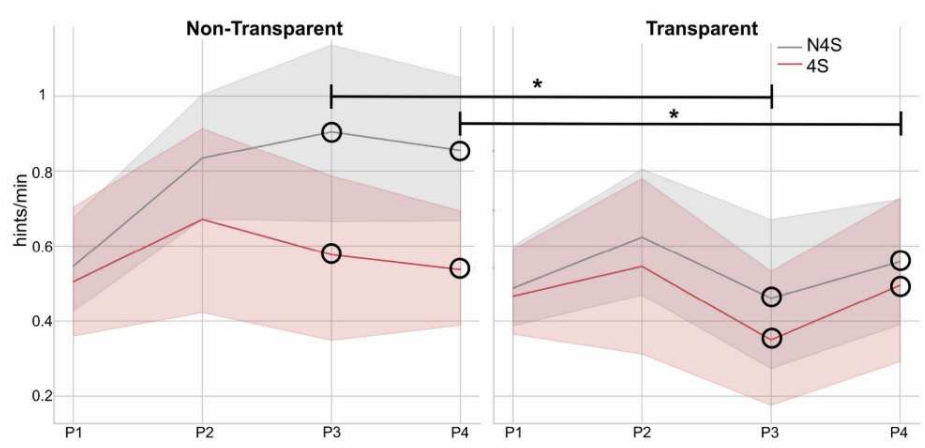

Figure 4.6 Participants who perceived all the faults (4S), remaining participants (N4S). Hint frequency divided by fault perception and fault periods: Start-AP (P1), AP-NH (P2), NH-FC (P3), FC-End (P4). Marked by (*), statistically different.

Hints Frequency

To compare the two UTH conditions, Figure 4.6 represents the frequency of hints asked per minute, divided according to fault periods: (P1) from the beginning till the Abrupt Pointing (AP); (P2) from AP till the Noised Hint (NH); (P3) from NH till Fake Crash (FC) and, (P4) from FC till the end. The hints-request frequency is plotted separately also as a function of whether participants perceived or not all the faults. Indeed, after disclosure, participants were asked to report if they realized each fault occurred at the end of the experiment. Against my expectations, just 10/31 participants noticed all the four faults in the Transparent condition and 12/32 in Non-Transparent. Based on this, I defined two subgroups: **4S**, participants who perceived all the faults, and the rest of the participants, called **N4S**. A two-way ANOVA was run for each fault period, with Condition (T, NT) and Group (4S, N4S) as factors. A significant difference between conditions emerged in P3 ($F(1,59)=10.01$; $p=0.002$) and P4 ($F(1,59)=4.38$; $p=0.004$). Neither group difference nor interaction was significant. On average, the hint-request frequency was significantly higher for participants in the Non-Transparent condition for the later fault periods. By inspecting the graphs in Figure 4.6 the N4S group from Non-Transparent seems to be driving the difference, showing a particularly high frequency of hints, while the other three are similar. Hence, it would be possible that the failures were not perceived as severe when the participants focused on the task. Indeed, to impact their perception and behavior, either the participants had to experience all the faults, or iCub had to disclose it was failing specifically.

4.5.2 Questionnaires

Faults Perception

As previously mentioned, not all participants perceived all failures. In the analyses, they are then separated into 4S (perceived all the four failures) and the rest (N4S). For each perceived fault, participants were asked to judge on a 7-point Likert scale how severe it was, how much it obstructed them, and how much time they lost.

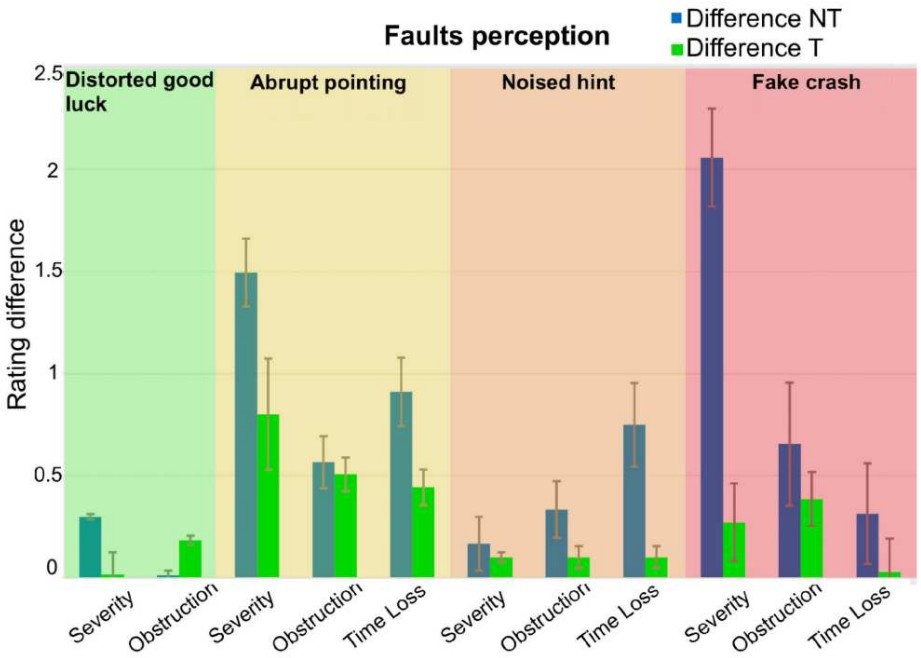

Figure 4.7 Differences in the fault judgment between those who experienced all the faults (4S) and the rest (N4S), based on conditions.

Figure 4.7 represents the differences in perception between the groups in the Transparent (T) and Non-Transparent (NT) conditions. Note that the question about Time Loss for the first fault (i.e., Distorted Good Luck) was removed as it happened before the timer's start. The differences between the 4S and N4S groups are larger in the Non-Transparent condition than in the Transparent one, meaning that a similar score is given by 4S and N4S in T, but not in NT. This suggests that in NT, there is a tendency where the people who experienced all the faults (4S) judged their respective severity, obstruction, and time loss higher that N4S (the group that did not experience all of them). Contrarily, in the Transparent condition, both subgroups (4S and N4S) evaluated the severity, obstruction, and time loss of all the faults similarly, suggesting that the Transparency condition, where iCub stated each time that it has a malfunction, influenced the perception and homogenized it.

NASA-TLX, Information Quality and Notion

Just after the experiment, before the disclosure, the Inclusion of Self (IOS) scale [266], NASA-TLX [265], perceived information quality [267], and iCub's Notion surveys were administered to participants. On average, across the three groups, the inclusion of self with iCub was 4/7, with no significant differences among Transparent (T), Non-Transparent (NT), and Treasure Hunt (TH) (one-way ANOVA).

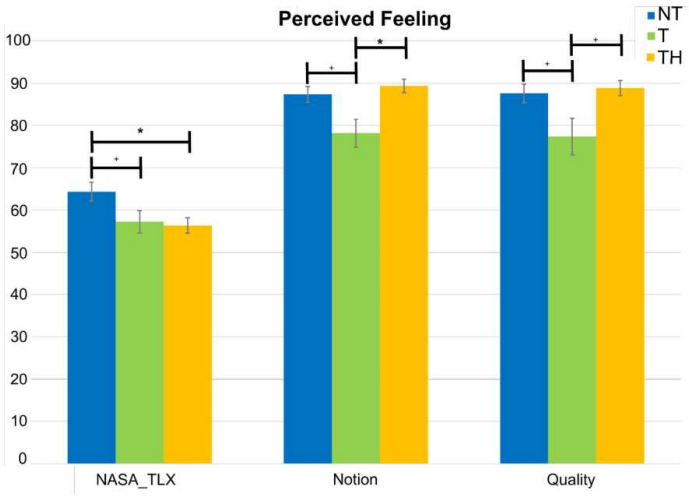

Figure 4.8 NASA-TLX overall workload, iCub's Notion, and Perceived Information Quality. Marked by (*) statistically different with a Bonferroni correction; (+) strong tendency but does not resist the Bonferroni correction.

To assess the potential difference in the total workload participants experienced, a one-way ANOVA was run on the NASA-TLX, followed by Bonferroni correction ($F(2,121)=3.65$, $p=0.028$). NT was associated with a task load significantly higher than TH and T (although the latter comparison does not resist Bonferroni correction 4.8. These results suggest that transparency could lower the task load index to a similar amount as when the robot was not experiencing any faults.

iCub's Notion was analyzed with a one-way ANOVA followed by Bonferroni post hoc, showing a significant difference ($F(2,121)=7.22$; $p=0.001$) among the three groups. NT (24.47/28;SD =2.9) is not perceived differently from TH (25.03/28; SD=3.44). Contrariwise, T (21.9/28; SD=5.03) is significantly lower than TH and shows a tendency to be lower than NT (but does not resist the Bonferroni correction with $p=0.017$ against a corrected threshold of $p=0.016$) (see Figure 4.8)

In *Perceived Information Quality*, a one-way ANOVA showed a significant difference among the three groups ($F(2, 121)=5.18$; $p=0.006$). T tends to be lower (16.25/21; SD=5.04)

than NT (18.4/21; SD=2.6) and TH (18.67/21; SD=2.9), however these differences failed to pass the Bonferroni correction with NT (p=0.04) and TH (p=0.018) against the corrected threshold of p=0.016 (see Figure 4.8). All the hints and pointing positions were the same in all conditions; however, the results of the last two analyses highlight that the Transparent condition generally has a lower score than the other two.

Pre-Post Questionnaires

Mind Attribution [263] was measured before (pre) and after (post) the experiment. For Non-Transparent (NT) condition, Mind Agency was rated pre: 15.62/28(SD=4.26); post: 14.68/28(SD=4.43); while Mind Experience statistically increased (paired t-test, t(31)=-3.09, p=0.004) from 8.59/28(SD=5.16) to 10.56/28(SD=5). In Transparent (T), Mind Agency was rated pre; 17.48/28(SD=5.68); post: 16.83/28(SD=5.04); while Mind Experience also statistically increased (paired t-test, t(30)=-3.29,p=0.002) from 9.54/28(SD=6.3) to 12.97/28(SD=6.62).

In both conditions, similar to literature [269], participants rated the mind agency of a robot somewhere midway, while the mind experience was quite low. After the experiment, the agency remains the same; however, the experience statistically increases in both conditions. This result follows the same trend as the original Treasure Hunt (TH). The ratings of the rapport questions [262] generally increased after the experiment in both conditions. However, the only statistically significant differences were in the NT condition and limited for the items: (i) Friends, from 3.84/7(SD=2.05) to 4.66/7(SD=2.01), paired t-test (t(31)=-2.57; p=0.01); and (ii) Happiness, from 3.68/7(SD=1.92) to 4.37/7(SD=1.68), paired t-test (t(31)=-2.43; p =0.02). In the T condition, the rapport increased, but not significantly. In TH, the increase was much higher and statistically significant.

Trust in robots [210] was only computed for participants who did not complete the game (23 for NT, 24 for T, as seen in Table I) as statistical differences can be observed in trust depending on the game outcome [14]. There were not enough participants in the other categories to perform statistical analysis. In NT, the only category where trust increased was the Benevolence trait, from 13.82/25(SD=3.41) to 15.47(SD=3.48) with a paired t-test (t(22)=-2.7; p =0.01). A series of one-way ANOVA was conducted among the three groups T/NT/TH for the different traits of Ability, Benevolence, and Integrity, but no significant differences were found. Against the initial expectations, the results in those three groups are similar.

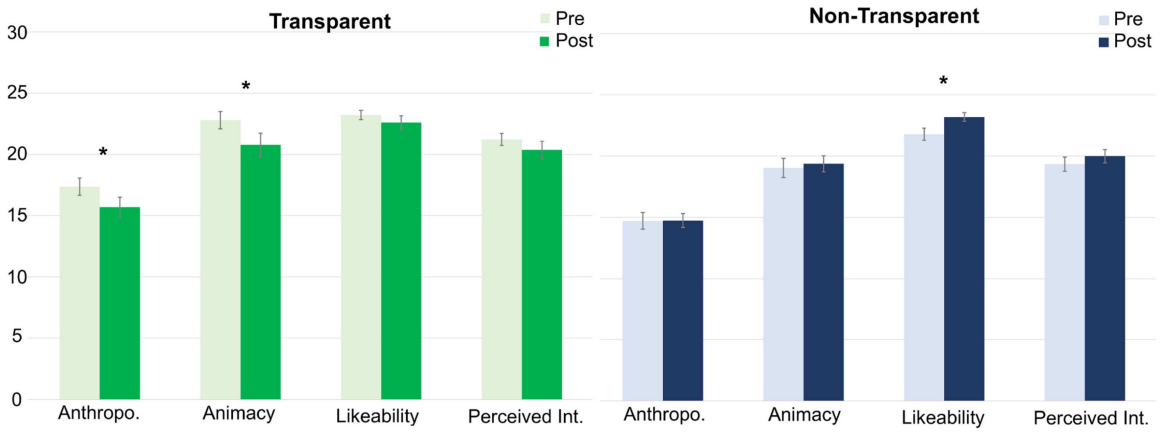

Figure 4.9 Pre-Post Godspeed questionnaire for both conditions. Marked by (*), statistically different.

The Godspeed [270] questionnaire was administered before and after the experiment for both conditions 4.9. For NT, the experiment caused an increase in the rating that reached significance only for Likeability: from 21.78/25(SD=2.81) to 23.18/25(SD=1.95), paired t-test (t(31)=-2.59; p=0.01). This result follows the same trend observed in the original Treasure Hunt game (TH). However, in T, the ratings decreased after the experiment. In particular, the decrease was statistically significant for Anthropomorphism: from 17.35/25(SD=3.89) to 15.67/25(SD=4.65), paired t-test (t(30)=2.18; p=0.03); and Animacy: from 22.8/30(SD=3.95) to 20.77/30(SD=5.42), paired t-test (t(30)=2.17; p=0.03). Note that all the pre values in T were statistically higher than the pre of NT. Although a lack of increase in ratings for this group could have been ascribed to a ceiling effect, this could not explain the significant decrease observed. Together with the previous analyses, these results suggest that Transparency decreases the overall quality of interaction.

4.6 Discussion

The first and most vital expectation in the Unreliable Treasure Hunt (UTH) design was that the faults would strongly negatively affect participants' behavior and trust toward the robot. Indeed, I designed the faults to be evident, and the online-survey participants judged them as severe and able to undermine the trust toward robots [250]. Unexpectedly, just a third of the participants noticed all the four faults, suggesting that the involvement in the treasure hunt game prevented the players to realize that something wrong was happening in the robot. Furthermore, the impact on participants' behavior (i.e., in the frequency of hints asked)

was more evident only when the robot overtly informed them about its failures (i.e., in the Transparent condition) or when they experienced all the faults 4.6.

Overall, considering the questionnaires' responses, there were no significant differences in trust perception between the original Treasure Hunt game (TH) and the Unreliable version. In particular, no apparent reduction in trust toward the robot could be found. Considering the behavioral measures, participants' performance in TH was significantly better than in UTH. However, this difference did not seem to arise from the unreliability of the novel condition. Indeed, participants' performance in UTH was lower since the start of the game, when the failures could not have had such a significant impact. Before the game started, they had only witnessed a very mild fault (iCub wishing "Good Luck" with a distorted voice). However, performance metrics such as the time to find the first egg, the average number of hints asked, and eggs found (4.5) were already lower than in TH.

A possible alternative cause of the observed lower performances in UTH might be found in the age and socio-economical status of the new participants. In UTH the average age was 38(SD=13.8) versus 30.9(SD=9.8) in TH (two-sample t-test: t(121)=-3.39; p=0.0009). Only a quarter of the sample was composed of students in UTH, compared to 33% in TH. The average older age of UTH players might have been associated with more reduced exposure to games such as treasure hunts and escape rooms, which tend to attract a younger audience. Also, participants were asked to move in the room actively, sometimes crouching under the tables or chairs or picking boxes placed on top of a high shelf; hence, an older age could have slowed down the UTH players. Alternatively, the worse performance could be caused by a reduced commitment to the game. Several UTH participants reported that the money they could win in the game was much less than their hourly salary, making the monetary award in the experiment not a strong motivator. Running a practice round could have mitigated these issues, preparing players for the real experiment; however, to allow a better comparison with the TH study, where no practice round was provided, it has been necessary to minimize the modifications. In summary, although it cannot be excluded that the robot's unreliability might have played a role in interfering with UTH participants' performance, it seems not to be the principal cause.

The results above show that (H1) the robot's mechanical faults did not negatively affect the participants' perceived trust. The strong involvement in the game and the choice of faults that - though severe - did not hinder its completion made the participants ignore the failure and still rely on the robot for help. This seems to be confirmed by participants' comments. Some of them, after the experiment, reported that faulting is normal in robots and that it was

not particularly important as the iCub robot kept working. So, against expectations, it is not enough for a robot to mechanically fail during a game to reduce human participants' trust.

Also, against expectations (H2), the perceived trust in the robot was not higher in the Transparent condition. There are no differences in gambling, conformation, reliance, or trust questionnaire between the two conditions. The results align with recent literature, which shows that transparency does not always lead to higher trust; instead, participants somehow utilize that information [271]. In the current experiment, the participants who experienced all the faults in the Transparent condition similarly judged them as the ones who did not experience them all; whereas, in the Non Transparent condition, the difference in judgment between participants who experienced all faults and the rest was much larger (see Figure 4.7). Similarly, the overall workload felt in the Transparent condition was almost at the same level as the TH, in contrast to the Non-Transparent condition (see Figure 4.8). In the Transparent condition, participants may have realized that the failures were iCub's fault (not theirs) and had a limited impact on the search activity, so they felt less stressed about the game outcome. In general, participants were talking back to iCub whenever it was disclosing a fault (i.e., *"Ok, I understood."*, *"I will tell the experimenter"*, *"Should I call an ambulance?"*).

On the flip side, the quality of interaction in the Transparent condition was perceived much worse: in the quality of information given [267] and iCub's Notion (see Figure 4.8). In the rating of the Godspeed scales [270] transparency even led to a reduction in the ratings after the experiment (see Figure 4.9). From a behavioral perspective, participants in the Transparent condition found fewer eggs and asked for fewer hints on average (see Figure 4.5), and the hint frequency by fault periods is lower than in the Non-Transparent condition (see Figure 4.6). It seems that the robot actively disclosing its failures negatively affects participants' behavioral and affective states. In this context, I implemented transparency as a post hoc simple explanation of the unusual distortions in the robot's behavior. Further research would be needed to explore the impact of different types of transparent behaviors, for instance, "predictive transparency," where the robot anticipates the upcoming malfunction.

The limited effect of the transparency observed in this experiment might also be related to the fact that the robot could autonomously recover from failure. Transparency might have played a much more relevant and positive role in a situation where the robot unexpectedly stopped working well (as in this experiment) but then required human intervention to recover its functionalities. There, informing the participant could become crucial for the interaction to continue; instead, from my results, it seems that transparency further "normalized" the different failures. Beyond that, it represented more a disturbance for the engaged player than an appreciated feature.

In conclusion, this study sheds light on the possible intertwined relations between unreliability and transparency when predicting human trust perception when participants are confronted with a faulty robot. A robot that fails does not necessarily lose its partners' trust, mainly if the failure is only mechanical and does not hinder the continuation of the interaction. Further research will be needed to assess the impact of different robot errors, i.e., cognitive or social. Moreover, in line with recent literature, a robot that automatically explains when it fails does not necessarily increase trust. However, it might unburden the perceived workload at the cost of worsening the perceived quality of interaction.

From the full-thesis point of view, results suggest that participants' decision to take a risk by complying with the suggestion of a faulty robot depends on their awareness. The frequency of hints asked and eggs found decreased if they either perceived all the faults (N4S) or the iCub raised their awareness by being transparent about its faultiness. Starting from this observation, I speculated that risk perception, or better appraisal, during decision-making is crucial in leading participants to compliant behavior. This was the first building block toward designing an ecological setup to study participants' behavioral and physiological reactions to diverse social engineering attacks. However, before getting there, I explored the usage of physiological signals to infer others' inner states and cognition.

Chapter 5

Detecting Deceptive Attackers

5.1 Overview

Any Social Engineering (SE) attack starts with the attacker fabricating a fake credible story [8]. This *pretext* is then used to establish a trust-based relationship with the victims. Most of the SE attacks happen via emails in the phishing scenario; there, the defense systems could detect the incoming spoofed messages and prevent them from reaching the target. However, attacks also happen face-to-face [7]. Through *impersonation* an attacker could pretend to be someone else and try to breach the physical security of sensitive buildings. Sadly, literature provides no defenses for this scenario other than authentication measures.

In this project, I explored how the humanoid robot iCub could help humans defending from face-to-face attacks by autonomously detecting deception in real-time. The fabrication and maintenance of a credible lie increase humans' cognitive load [120, 9, 131] more than truth-telling. By detecting this inner state change, a humanoid robot could understand how much the interacting partner is trustworthy and raise a security warning. To do so, robots could rely on established cognitive load proxies like pupillometry [118], blinking [130], voice tone [272], hand movements [273], skin temperature [274]. In this study, I relied on pupillometry as it is less controllable and, hence, less deceivable than the other proxies [275]. Also, previous studies in which I contributed proved how, in deception detection during HRI, pupil dilation is more reliable than blinking and response time[133]. Finally, using pupillometry allowed me to start answering to **RQ1**: the usage of physiological signals in the Social Engineering defense. Furthermore, I addressed a second literature gap: state-of-the-art deception detection systems focus mainly on formal, interrogatory-like scenarios. These contexts, even if realistic, represent only a small subsection of the real-world interactions in

which it could be helpful to unmask liars.

For this purpose, I developed an end-to-end (E2E) architecture, enabling the humanoid iCub to detect lies autonomously in real-time during an informal human-robot interaction. ICub can learn from a brief game session how a specific partner lies and tells the truth, exploiting and improving this knowledge in classifying further sentences. Post-hoc, I trained several machine learning classifiers and anomaly detectors to improve iCub's performances; and I explored how to further extend the system by taking advantage of how humans detect lies in everyday life.

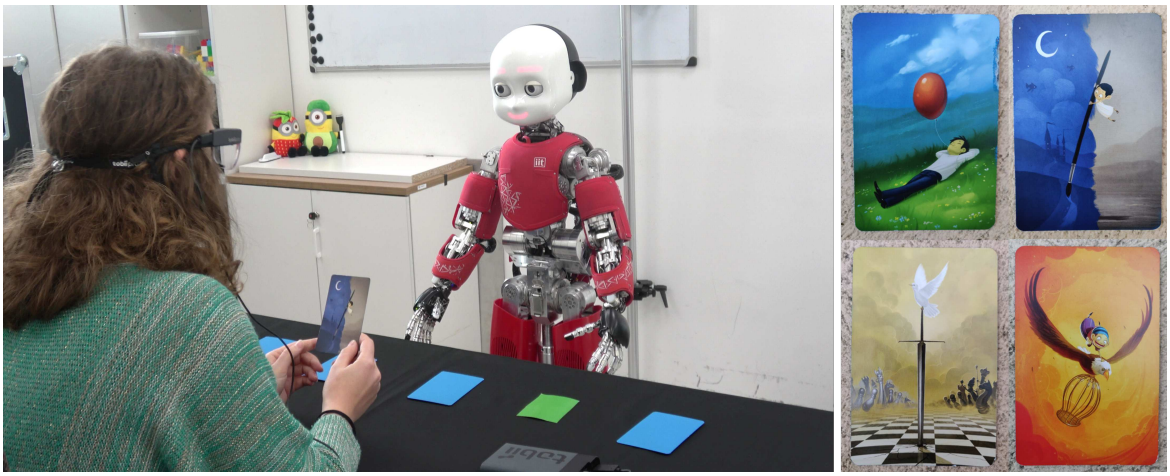

Figure 5.1 (Left) Participant describing a card to iCub, while wearing the Tobii Pro Glasses 2 eye tracker (Logitech Brio 4k webcam point of view); (Right) Examples of Dixit Journey gaming cards (authored by Jean-Louis Roubira, designed by Xavier Collette and published by Libellud).

5.2 An Informal Setup to Detect Lies

Taking inspiration from the game *"Box of Lies"*[1] of the *"Tonight Show"* by Jimmy Fellon, I designed and developed a magic-trick-like card game in HRI to study lie detection during informal interactions. In this game, players describe a set of cards from the Dixit[2] card game to the iCub while wearing the Tobii Pro Glasses 2 eye tracker [276]. Players were instructed to describe each card *"credibly and deceitfully"* or *"to describe what they see"*. ICub autonomously led the game (see next section) and detected players' lies in real-time.

[1]https://www.youtube.com/watch?v=Md4QnipNYqM
[2]https://boardgamegeek.com/boardgame/121288/dixit-journey

5.2.1 The Magic Trick card game

The game was composed of two phases, *Calibration Phase* and *Testing Phase*, both led autonomously by iCub. During the game, participants sat on a chair in front of the robot with a table between them. The deck of Dixit cards and six green rectangular marks in a row (see Figure 5.1) were on the table. Also, there was a Tobii Pro Glasses 2 eye tracker.

Calibration Phase Firstly, iCub asked the participants to shuffle the deck, extract six cards without looking at them, and put the deck on a closet nearby. Then, iCub asked them to draw out one of the cards (it called it the *secret card*) and memorize it. Afterward, iCub instructed the participants to look at each of the six cards, one by one, shuffle them, and put them facing down on the six green marks on the table. iCub explained that it was going to point each; they had to take the pointed card, look at it, describe it, and then put it back facing down on the table. Then, iCub explained the game rules: *"The trick is this: if the card you take is your* ***secret card****, you should describe it deceitfully and creatively. Otherwise, describe just what you see"*. Finally, iCub asked the participants to wear the Tobii Pro Glasses 2 eye tracker, take a deep breath and relax.

iCub randomly pointed to each of the six cards, listened to participants' description and, acknowledged it with a short greeting sentence (i.e., *"ok"*, *"I see"*, ...). After the last description, iCub guessed the participants' *secret card* and asked them to either put the six cards aside to validate the detection or show to iCub the *real secret card* to reject it. Participants' confirmation is meant to select the correct secret card if iCub fails to detect it. Before the beginning of the Testing Phase, the experimenter could manually override the detected *secret card* with the one presented by the participants - this is the only manual intervention needed. Finally, iCub asked participants to set aside the six cards to start a second game.

Testing Phase When the participants removed the six cards from the table, iCub asked them to take the deck and draw six new cards. iCub told the participants to look at all the cards, one by one, then shuffle them and place them on the six green marks. Afterward, iCub instructed the participants that it would point to all the cards from right to left (from the participants' point of view) and instructed them to handle the pointed card as in the first game. However, it added:*"This time you can decide, for each card, whether to describe it creatively and deceitfully or to describe just what you see"*. While the robot explained the rules, the participants kept wearing the Tobii Pro Glasses 2.

For each card, iCub pointed to it, listened to the participants' description, acknowledged it with a short sentence, tried to classify the description as truthful or fake, and asked for confirmation. The participants had to show the card they just described to reject iCub's classification or do nothing to validate it.

General Remarks

During the rule explanation of the two phases, iCub instructed the participants to press a button on a keyboard to move to the next task (i.e., after shuffling the cards deck or memorizing the secret card). No time limit was given to shuffle the card, look at them, memorize the secret card, or describe them. iCub's pointing has been designed to replicate a human-like gesture: first moving the gaze toward the target, then the arm, fingers, and torso with a biologically inspired velocity profile.

5.3 End-to-end Lie Detection System

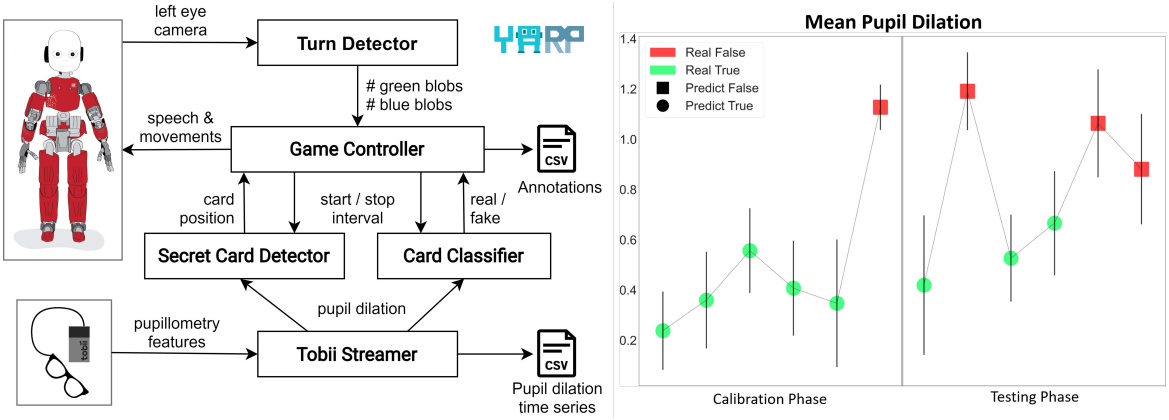

Figure 5.2 (Left) Autonomous end-to-end architecture used in real-time to enable iCub to lead the card game; (Right) Mean pupil dilation during *Calibration* and *Testing Phases* for participant A. Green circles are truthful card description; red squares are false ones. Bars represent standard errors.

iCub autonomously leads the card game thanks to the end-to-end (E2E) architecture in Figure 5.2 (left).

Turn Detector

With the *Turn Detector*, iCub detects the beginning and end of each card description by tracking the number of green (marks) and blue (cards) blobs visible in the scene with a simple

Hue Saturation Value (HSV) filter. This is also used to understand participants' confirmations (i.e., reveal all the marks (or none) to confirm the detection (classification); or reveal only one mark to reject them).

Tobii Streamer

The *Tobii Streamer* reads participants' pupillometry features from the Tobii Pro Glasses 2 eye tracker; it streams in real-time participants' right-eye pupil dilation (10 Hz) and logs the complete set of pupillometry and gaze features of both eyes (100 Hz) over the YARP robotic platform [277]. I developed the Tobii Streamer extending the Tobii Glasses Py Controller previously developed by De Tommaso et *al.* [278].

Game Controller

The *Game Controller* implements the main game engine: (i) it controls iCub's movements and speech; and (ii) it segments the start and end of each pointing, card description, and phase, logging annotation events. The logged annotation events and pupil data points are synchronized over the YARP robotic platform, providing an autonomous annotation for future analysis.

Secret Card Detector & Card Classifier

Finally, the *Secret Card Detector* and the *Card Classifier* enable iCub to identify participants' lies during the game. iCub detects (*Calibration Phase*) and classifies (*Testing Phase*) players' lies thanks to a specific Task Evoked Pupillary Response: the fabrication of a credible and consistent deceptive card description triggers an increase in players' cognitive load [279, 280]; this increment reflects on a higher pupil dilation with respect to truthful card descriptions [9, 132, 281]. ICub aggregates participants' pupil dilation data points, computing the mean pupil dilations during each card description and using them to detect players' lies. I focused on the right eye's pupil dilation since both Tobii documentation [282] and previous results indicate that pupil dilation is not different between the right and left eye (as we confirmed in a preliminary work [133]). Furthermore, to avoid unnecessary impacts on the informality of the interaction, I decided to skip the calibration of the Tobii eye tracker: Tobii documentation reports how the calibration is only relevant for the gaze features and does not impact the pupil dilation measurement [282]. The components implement two heuristic methods which behavior is shown in Figure 5.2 (right):

Calibration Heuristic During the Calibration Phase, iCub detects as the *secret card* the one related to the highest mean pupil dilation among the six cards' descriptions. Indeed, if lying reflects on wider pupil dilation, it is reasonably safe to assume that the mean pupil dilation for the only false description should be the highest. I developed this heuristic in preliminary work, published in [185]. However, it is not always true, as I will show in section 5.5.2

Testing Heuristic At the end of the Calibration Phase, iCub knows 6 mean pupil dilation data points: one related to the secret card and five related to truthful cards. With them, it computes two reference scores: the *true reference* score is the average of the five mean pupil dilations of truthful cards; the *false reference* score is just the secret card mean pupil dilation. Then, the mean pupil dilation for each novel card description (from the Testing Phase) was computed and compared to the two reference scores. By taking the absolute minimum difference, iCub could label the current description as real or fake. Note that no absolute threshold for the changes in pupil dilation is used: thresholds are dynamically learned for each player from the *Calibration Phase* data.

5.4 Experiment

5.4.1 Publications

The magic trick card game experimental setup, the E2E architecture explained above, the data collection, results, and discussion below only concern the last iteration of an incremental design process. Indeed, this lie detection experiment was born as a spin-off of another study in which we proved the Task-Evoked Pupillary Responses (TEPR) related to deception happen both in interrogatory-like Human-Human, and Human-Robot interaction [133][3]. After the main study, participants were asked to describe six fixed cards to iCub, lying on a previously extracted one; the experimenter listened to the descriptions through iCub's sensors and performed the classification. Interestingly, I found the same TEPR effect also in the informal card game [283][4], which pushed me to explore it further. In the first iteration, I

[3] Article peer-reviewed and published to Frontiers in Robotics and AI. I took care of the data collection and manual annotation

[4] Article peer-reviewed and published as Late-Braking Report to the International Conference of Human-Robot Interaction (HRI2020). I took care of the data analysis and writing; the other authors contributed with advice and proofreading

developed and validated the *Calibration Heuristic* in a pilot study [284][5]. Then, I introduced the Dixit cards and developed the E2E architecture, making iCub able to autonomously lead the *Calibration Phase* of the HRI [185][6]. In the last iteration, I introduced the *Testing Phase*, making iCub able to learn from a brief interaction and adapt over further sessions [285][7]. Finally, I divulged the scientific contributions of this project at two Italian conferences in Rome [286][8], and Turin [252].

5.4.2 Setup and Materials

The room was arranged to replicate an informal interaction scenario (Figure 5.3). The participants sat in front of the iCub humanoid robot, separated by a table covered with a black cloth. The experimenter placed six green marks (95x70 mm), a deck of 84 cards from the Dixit Journey card game with the back painted in blue, a keyboard, and a Tobii Pro Glasses 2 eye tracker on the table. There was a little drawer (deployment area) on the participants' left, while a black curtain hid the experimenter from participants' sight on the right. Behind iCub, a 47 inches television showed iCub's speech during the interaction (to prevent misunderstanding). A Logitech Brio 4k webcam, fixed on the television, recorded the scene from iCub's point of view at a resolution of 1080p (Figure 1, left).

The Dixit Journey card deck is composed of 84 cards (80x120 mm) with different toon-styled drawings meant to stimulate creative thinking [287] (Figure 5.1 - right). Designing the card game, I tried to avoid any cue – other than the wearable eye tracker – for the participants about the method used by iCub to detect their false card descriptions; in this sense, I avoided any machine-readable mark (i.e., QR codes on cards' back) that iCub could use to recognize the cards. The Tobii Pro Glasses 2 eye tracker recorded participants' pupillometry features at a frequency of 100 Hz and streamed the participants' pupil dilations at 10 Hz in real time. The window blinders were closed, and the room was lit with artificial light to ensure a stable illumination condition for all the participants.

[5] Article peer-reviewed and published to the Workshop on Exploring Creative Content in Social Robotics of the International Conference of Human-Robot Interaction (HRI2020). I took care of the development, data collection, analysis, and writing; the other authors contributed with advice and proofreading

[6] Article peer-reviewed and published to the International Conference of Human-Robot Interaction (HRI2021). I took care of the development, data collection, analysis, and writing; the other authors contributed with advice and proofreading

[7] Article peer-reviewed and published to the International Journal of Social Robotics (IJSR). I took care of the development, data collection, analysis, and writing; the other authors contributed with advice and proofreading

[8] Article peer-reviewed and published to the I-RIM 2021 conference

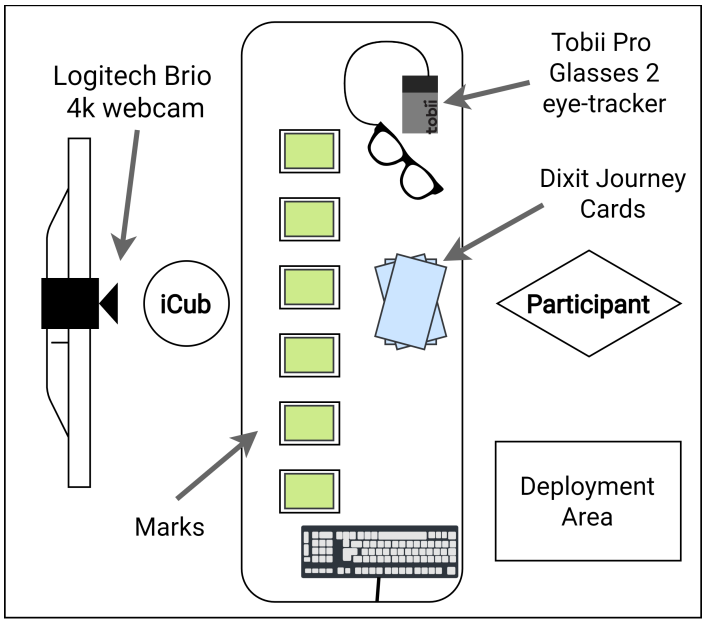

Figure 5.3 Card game experimental setup with iCub (left) and the participant (right) sitting on a table. The deployment area is the location where the remaining Dixit Cards after each drawing were placed.

The iCub humanoid robotic platform [179] played the role of a magician. iCub autonomously led the whole interaction thanks to the autonomous end-to-end (E2E) architecture in Figure 5.2. The experimenter monitored the scene through iCub's left eye, ensuring the participants' safety and the correct execution of the experiment.

5.4.3 Participants

39 participants (25 females, 14 males), with an average age of 28 years (SD=8) and a broad educational background, participated in the experiment. They signed an informed consent form approved by the ethical committee of the Liguria region (Italy), where it was stated that cameras and microphones could record their performance and agreed on using their data for scientific purposes. After the experiment, they received monetary compensation of 10€. Although all participants completed the game, 5 were excluded from further analysis: 2 for technical issues, 2 because they did not follow the game's rules. The last one was considered an outlier, as he/she concluded the game in 38 minutes (duration longer than 3SD plus the average game duration, which lasted 17 minutes). Hence, the final sample includes N=34 participants (22 females, 12 males).

5.4.4 Procedure

At least a day before the experimental session, the participants filled in the following questionnaires: The Big Five personality traits (extroversion, agreeableness, conscientiousness, neuroticism, openness) [256]; the Brief Histrionic Personality Disorder (BHPD) [288]; and the Short Dark Triad (SD3, Machiavellianism, narcissism, and psychopathy) [289].

After signing the informed consent, the experimenter led the participants to the experimental room where they played the magic trick card game as explained in Section 5.2.

After the experiment, the participants filled in the NASA-TLX [265] and a set of questions regarding (i) the experienced fun, (ii) creative effort, (iii) strategies adopted in fabricating a deceitful and creative description during the game, (iv) previous experience about the Dixit Journey card game, (v) previous experience about improvisation and acting, and, (vi) habits on playing deception-related games.

5.4.5 Pupillometry Measures and Data Preparation

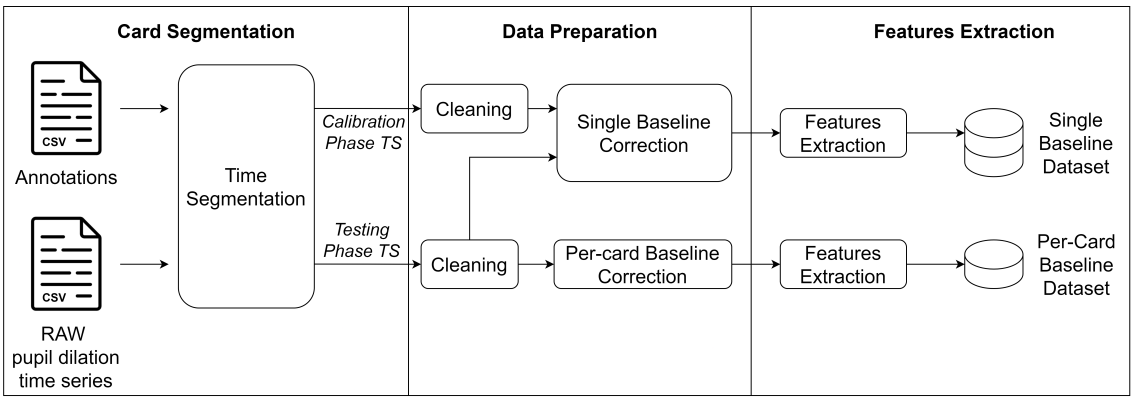

Figure 5.4 Computational workflow to preprocess the collected data from Tobii Pro Glasses 2 eye tracker. Two datasets are extracted. The difference depends on the applied baseline correction (single or per-card).

I built two datasets from the pupil dilation data points collected during the game following the computational workflow in Figure 5.4. I could use only the right and left eye pupil dilation (in millimeters) from the pupillometry features measured by the Tobii Pro Glasses 2 eye tracker (gaze points, fixations, saccades, pupil dilation). Indeed, I decided not to perform the device calibration - not necessary for pupil measurement as reported in Tobii documentation [282] - to avoid impacting the informality of the social interaction. Hence, the only reliable eye measures are the pupil dilation [282]. Pupil dilation data points are synchronized over the YARP robotic platform [277] with the annotation events.

Card Segmentation (Figure 5.4, left)

The Game Controller autonomously performs the card trial annotation (Figure 5.2, left) by rising annotation events on the YARP robotic platform for the beginning and end of each pointing and card description. I segmented the pupil dilation time series into three temporal intervals for each card trial: (i) *robot's turn*: the iCub's pointing gesture, from the moment iCub starts the pointing gesture till the participant takes the pointed card from the green mark; (ii) *player's turn*: the card descriptions, from the moment participants, take a card from the green mark, till they put it back on it; (iii) *card trial*: the whole interaction for a single card, from the moment iCub starts the pointing gesture till the participants put the card back on the green mark.

Data Preprocessing (Figure 5.4, center)

I fitted and re-sampled the time series at 10 Hz to make them consistent with the real-time processing; then, I applied a median filter to remove the outliers and a rolling window means filter to smooth the time series and infer any eventual missing data points. I then corrected each time series subtracting a baseline value for each participant [168]. In this reference system, a positive value represents a dilation, while a negative value represents a contraction with respect to the baseline. I corrected the time series based on two different baselines: (i) In the *Single Baseline Correction*, the baseline is computed as the average pupil dilation during the 5 seconds before the first pointing of the Calibration Phase - when iCub asks participants to take a deep breath and relax - and applied to all the cars of both phases; (ii) in the *Per-card Baseline Correction*, a specific baseline is computed for each card as the average pupil dilation during the 5 seconds before each pointing.

Feature Extraction (Figure 5.4, right)

Finally, I aggregated the time series of each temporal interval and computed several features. For each *player's turn*, *robot's turn*, and *card trial* I computed the *maximum, minimum, mean and standard deviation of the pupil dilation* in millimeters, and the *duration* in seconds.

Also, on the whole card trial, I computed 26 time series features using the python library Time Series Feature Extraction Library (TSFEL) [290]. In particular, the TSFEL features are: (i) *Statistical Features*: *median, median absolute deviation, mean absolute deviation, kurtosis, skewness and variance*; (ii) *Temporal Features*: *absolute energy, area under the curve, auto-correlation, centroid, entropy, mean absolute difference, mean difference, median absolute difference, median difference, peak to peak distance, slope, total energy*; (iii)

Spectral Features: *fundamental frequency, maximum frequency, median frequency, spectral centroid, spectral entropy, spectral kurtosis, spectral skewness, spectral slope*. I considered the features for both eyes as separate data points to augment the datasets. This results in two different datasets:

Single Baseline Dataset This dataset includes the data points of both phases, replicating the data structure used in real-time. It is meant to explore incremental learning over multiple interactions with the same individual.

Per-card Baseline Dataset This dataset, instead, includes only data from the Testing Phase; it is meant to train generic machine learning models independent from the specific interacting partner.

Shapiro-Wilk and D'Agostino K-squared normality tests showed that some of the features of the datasets were not normally distributed. Therefore, I opted for non-parametric tests for all the following statistical analyses. Additionally, I decided to focus on data points from participants' right eye only (unless otherwise specified) since there is no difference between right and left eye features [282].

5.5 Results

This section reports the in-game and questionnaires results and post-hoc analysis of the collected pupillometry data.

5.5.1 In-game Results

The interaction lasted, on average, 17 minutes (SD=5) from the beginning of iCub explaining the Calibration Phase's rules till the final greeting of the Testing Phase. The Calibration Phase lasted, on average, 8 minutes (SD=3), during which iCub successfully detected the players' secret card with an accuracy of 88.2% (against a chance level of 16.6%, N=34). The Testing Phase lasted, on average, 8 minutes (SD=2). The participants were free to choose whether to lie or not, producing on average 2.73 (SD=0.94, 45%) false descriptions among the six cards. ICub successfully classified each card description as true or false with accuracy = 70.8%, precision = 73.6%, recall = 57% and F1 score = 64.2% (N=34).

Table 5.1 Participants' psychological profile

Score %	**Big Five *{O, C, E, A, N}***	**Dark Triad *{M, N, P}***	**Histrionic**
0 - 20	{0, 0, 0, 0, 2}	{2, 6, 15}	6
20 - 40	{4, 5, 3, 1, 17}	{8, 4, 13}	4
40 - 60	{24, 22, 23, 6, 6}	{18, 11, 1}	11
60 - 80	{1, 2, 1, 21, 3}	{0, 4, 0}	4
80 - 100	{0, 0, 0, 1, 1}	{1, 4, 0}	4

Table 5.1 summarizes the results of the Big Five personality traits [256], Brief Histrionic Personality Disorder [288] and Short Dark Triad [289] questionnaires, performed before the experiment. Average scores for the Big Five were Agreeableness: M=0.659, SD=0.113; Conscientiousness: M=0.481, SD=0.072; Neuroticism: M=0.387, SD=0.16; Openness to experiences: M=0.476, SD=0.07 and Extraversion: M=0.486, SD=0.061. Considering the Dark Triad, the scores were Psychopathy: M=0.191, SD=0.113; Machiavellianism: M=0.438, SD=0.129 and Narcissism: M=0.396, SD=0.15. For the Brief Histrionic Personality Disorder, the average score was M=0.481, SD=0.26.

After the experiment, participants filled in the NASA-TLX questionnaire, rating on a 10-point Likert scale their effort in performing the task. They reported a low average task load (M=3.717, SD=1.041). Among the components, Mental Effort (M=5.41, SD=1.78), Fatigue (M=5.07, SD=2.14) and Performance (M=5.35, SD=2.32) are slightly higher than Temporal Effort (M=2.59, SD=1.72), Frustration (M=2.72, SD=1.83) and Physical Effort (M=1.21, SD=0.49). This is consistent with the requirements of the task. Also, I asked participants to self-report, on a 5-points Likert scale, the effort they put into fabricating creative and deceptive descriptions (Lie Effort: M=4.17, SD=0.71) and the experienced fun (Fun: M=4.59, SD=0.57).

I explored whether pupil dilation features were dependent on participants' personality traits. I considered the Testing Phase data from the Per-card Baseline Dataset to minimize the impact of card presentation order on pupil features, normalizing each card for its baseline. I fit two linear regression models with the personality traits from the pre-questionnaire as independent variables and, as dependent variables, the difference between mean pupil dilation for false and true cards or the mean pupil dilation baseline. Results show that only neuroticism correlates significantly with the mean pupil dilation baseline (t=2.492, p=0.021, Adj. R2=0.115). See [185] for a deeper analysis of the questionnaires.

5.5.2 Learning From a Brief Interaction

To investigate the relationship between pupil dilation and lying observed during the game, I analyzed the *Single Baseline Dataset* which resembles the data structure used in real-time.

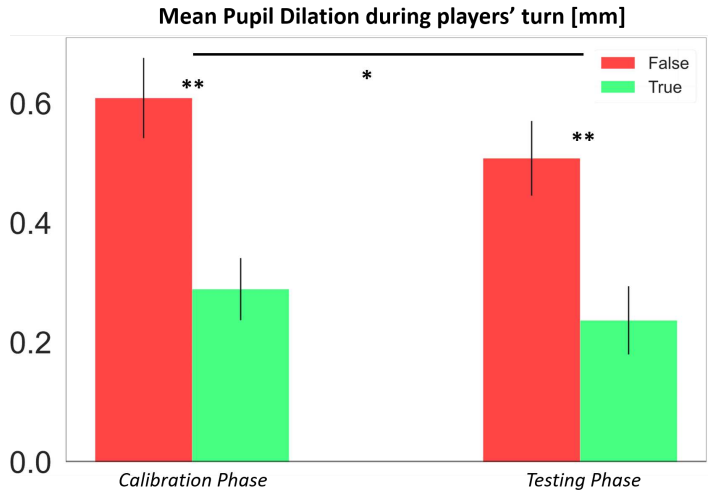

Figure 5.5 Average of mean pupil dilation during player's turn for *Calibration* and *Testing Phases*, with standard errors of the mean. (* = p<0.05, ** = p<0.001).

The *Single Baseline Dataset* presents a multilevel structure (multiple phases for the same participant, nested in card trials, nested in turns) with unbalanced card classes (one secret card among six (about 16.6%) in the *Calibration Phase*; and on average 45% of false cards in the *Testing Phase*). Since the real-time game was based on participants' mean pupil dilation during the player's turn, I decided to focus on such temporal intervals.

I fitted a mixed effects model for the player turns with mean pupil dilation as the outcome variable. As fixed effects I entered *"card label"* (two levels: *true*, *false*), *"phase"* (two levels: *calibration*, *testing*) and their interaction into the model. As random effect I had intercept for participants. I set the reference level on the *Testing Phase* and false card label. Results show a highly significant effect of card label (B=-0.223, t=-8.885, p<0.0001) revealing a higher mean pupil dilation for the false card descriptions with respect to the truthful ones. I also found a significant effect of phase (B=0.104, t=2.428, p=0.016), with a significantly lower mean pupil dilation in the Testing Phase, and no significance of the interaction between the two factors (B=-0.052, t=-1.023, p=0.307) (see Figure 5.5).

As an exploratory analysis, I fit another mixed-effects model on the *robot's turn*, with the same structure mentioned above. Results show no effect on the card label (B=-0.035, t=-1.373, p=0.171), but a highly significant effect on the phase (B=0.124, t=3.490, p=0.0005) confirming a lower mean pupil dilation in the *Testing Phase* with respect to the *Calibration*

one also for this turn. Finally, I found no effect of the interaction of card label and phase factors (B=-0.014, t=-0.331, p=0.741).

Incremental Testing Heuristic

Even if the Testing Heuristic demonstrated a pretty good accuracy – humans perform near chance on detecting lies [291] – it has a low recall score (recall = 57%, accuracy = 70.8%, precision = 73.6%, N=34), that is, it recognizes only a relatively low proportion of the false statements made by the participants.

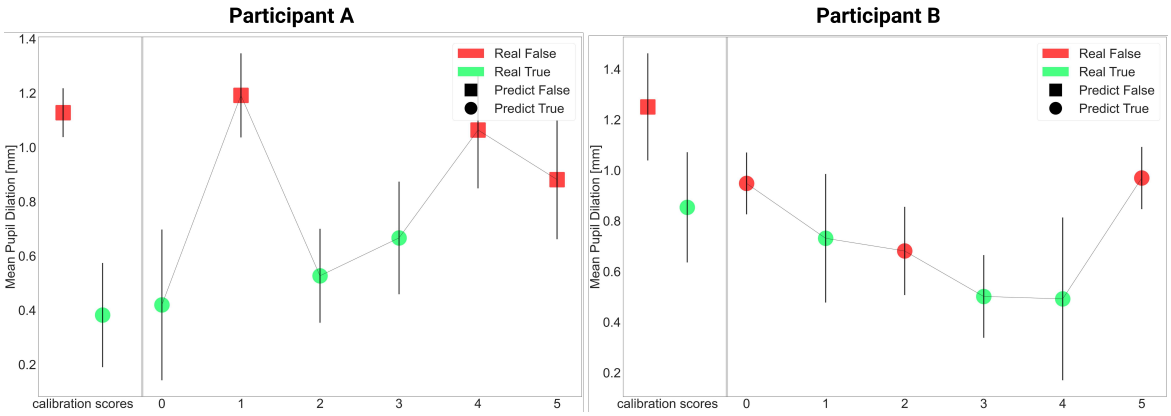

Figure 5.6 Mean pupil dilation data points as seen by the Testing Heuristic for participant A (left) and B (right). Color represents the real class (green = true, red = false); shape represents the predicted class (circle = true, square = false); bars represent standard deviation.

Figure 5.6 provides two examples of correct (participant A, left graph) and wrong (participant B, right graph) classifications. The two panels show the mean pupil dilation as processed by the *Testing Heuristic*. In each graph, the two data points on the left represent the two *reference scores*: the red square is the mean pupil dilation for the *secret card*, while the green circle is the average of the mean pupil dilations for the truthful cards. On the right side are the mean pupil dilation data points for each card of the *Testing Phase*. For participant A, pupil dilations for false and true descriptions remain consistent with the average values measured during the previous phase, and the classification is always successful. Conversely, all the *Testing Phase* mean pupil dilations of participant B (right graph) fall in the range of the *true reference score*. Hence all the false card descriptions have been misclassified as false positives (red circles).

The observed errors are caused by two assumptions on which the heuristic is based: (i) the difference in pupil dilation between false and true sentences remains almost the same

between the two phases, and (ii) participants' pupil dilation remains almost stable between the two phases. The first assumption is confirmed by the non-significant difference in the interaction of "phase" and "card labels" in both turns. However, the statistical analysis showed that participants' pupil dilation is, on average, lower during the *Testing Phase*.

To compensate for this effect and increase the robustness of the heuristic, I explored the possibility of incrementally adapting the *reference scores* for truthful and false card descriptions. After each card classification, the new card value is aggregated with the *reference scores*. This way, iCub incrementally learns how the human partner lies and tells the truth, improving the classification performances. I simulated the *Testing Heuristic* based on the mean pupil dilation during the player's turn, as in the real-time game, but including the incremental learning. For each *Testing Phase* card trial, both the reference scores are updated, computing the mean between each score and the novel means pupil dilation data point. The heuristic performance increases to accuracy = 76.7%, precision = 76.1%, recall = 73.7% and F1 score = 75.6%.

Then, I simulated the Testing Heuristic by performing a grid search on several parameters: (i) all the possible combinations of the available features (limited to a maximum of 3 features considered at the same time, see section 5.4.5 for the full list); (ii) methods to compute the true reference score (mean, median, minimum); (iii) methods to update the reference scores (mean, difference, quadratic error); (iv) whether to update both scores or just the one of the correct class; (v) whether to update the reference scores only if the card trial is misclassified. I prioritized the recall score since I assume that a lie detection system could detect more true negatives (i.e., spot a larger amount of lies) even at the expense of having a few false positives. The best heuristic has an accuracy = 78.7%, precision = 76%, recall = 80% and F1 score = 77.9%. It is based on the mean and minimum pupil dilation during the player's turn, compared by a 2D Euclidean distance with the reference scores. The true reference score is computed as the minimum among mean pupil dilations for the truthful cards' descriptions during the *Calibration Phase*; both the reference scores are updated in any case, averaging each score with the new values.

Random Forest classifier

Even if the new heuristic method performs better than the one exploited in real-time, it is still not generic and robust enough to describe the variability of participants' pupil dilation between the two phases. Indeed, the *Testing Heuristic* is meant to adapt to each specific

individual. I supposed that, by relaxing this constraint, it would be possible to compensate for the variability between the two phases.

I trained a machine learning model to learn from the *Calibration Phase* from the whole participants' sample and exploit the gained knowledge on the *Testing Phase*. The classification problem is a binary; it is defined by a couple [X, Y] where: X (42 x 1) is the vector of input behavioral features and Y = [0: true; 1: false] is the vector of desired outputs. I defined a within-participant split, considering the *Calibration Phase* data as training set and the *Testing Phase* data as validation and test sets (with a splitting ratio of 50:50). Calibration Phase data points have two main issues: (i) they are unbalanced (1 secret card among 6 cards); and (ii) the set is relatively small (6 cards for 34 participants and 2 eyes for participants, for a total of 408 data points). I considered features from both eyes to augment the dataset. Due to these limitations, I selected a Random Forests algorithm [292]. This model should not overfit when increasing the number of trees, even with relatively small datasets. Also, I tackled the unbalancing problem by oversampling the *Calibration Phase* data points with the synthetic minority oversampling technique (SMOTE) [293]. I did not oversample the *Testing Phase* data points validating and testing on real data. Even if not strictly required by the Random Forest algorithm, I applied a min-max normalization [294] to all the features within the data points of each participant in both phases. The idea is that a relevant value for a participant could not be relevant for another. I performed a grid search validation with a fixed validation set, searching for the best hyper-parameters and feature set for the random forest classifier. Due to the unbalanced dataset, the traditional accuracy score is unreliable; hence, I rely on the F1 score, precision, recall, and ROC AUC score [295]. The best random forest classifier trained on the full features set achieved F1 score = 56.5%, precision = 57.1%, recall = 55.9% and ROC AUC score = 59.6%.

Lying as an anomaly

Given the low performance of the random forest classifier, I changed the approach considering the lie detection task as an anomaly detection problem. In this context, the model knows just the values associated with true descriptions and learns to consider what is not truthful as a lie.

I trained a one-class support vector machine [296] anomaly detector on the *Calibration Phase* data points, validating and testing it on the *Testing Phase* data points, as I did above. I considered as the training set the truthful card description of the *Calibration Phase* and I carefully balanced *Testing Phase* data points, preserving the ratio between true and false card descriptions in the validation and test sets. I performed a grid search validation with the fixed validation set, searching for the best hyper-parameters and feature sets for the

one-class SVM model. Due to the nature of the anomaly detection problem, I evaluate it based on precision, recall, and F1 score. The best one-class SVM model achieved an F1 score of 67.7%, a precision of 60%, and a recall of 77.8%. It is based on features from both the player's turn (*minimum, maximum, and mean pupil dilation*); and the whole card trial (*minimum, maximum, mean, and median pupil dilation; total energy, absolute energy, and auto-correlation*).

5.5.3 Detecting lies from novel human partners

After analyzing how previous knowledge gained during an interaction can be used to improve lie detection in a subsequent task, I explored the possibility of building a pupil-dilation-based lie detector to classify false card descriptions from novel human partners. This could be the first step toward a minimally invasive and ecological lie detector able to classify a generic sentence as true or false without any previous interaction with the specific partner. In this sense, it is important to consider the card descriptions as independent as possible from the specific participant and the description order. Hence, I focused on the *Per-card Baseline Dataset* which includes only *Testing Phase* data points. In the *Per-card Baseline Dataset*, the baseline is computed as the average pupil dilation for each eye separately during the 5 seconds before each card trial. This baseline is subtracted from the pupil dilation time series of the relative card description (see Section 5.4.5). I considered only the data from the *Testing Phase* since the nature of the task – *"This time, you can choose, for each card, if describing it deceitfully and creatively, or describe what you see"* makes each card description more similar to a generic and standalone lie.

First, I analyzed whether the use of a *Per-card* baseline determined substantial differences with respect to the descriptive and statistical analysis conducted with the *single* baseline. I fitted a mixed effects model with mean pupil dilation as the outcome variable. I considered *"card label"* (two levels: *true*, *false*) and *"turn"* (two levels: *robot*, *player*) and their interaction as fixed factors, and I had as random effect the intercept for participants. I set the reference level on the *player's* turn and false card label. Results show a highly significant effect on the card label (B=-0.234, t=-6.58, p<0.001), the turn (B=-0.321, t=-6.39, p<0.001) and their interaction (B=0.255, t=5.205, p<0.001). This pattern of results (Figure 5.7) is similar to what observed for the Testing phase in the analysis with the "same" baseline (cf. Figure 5.5).

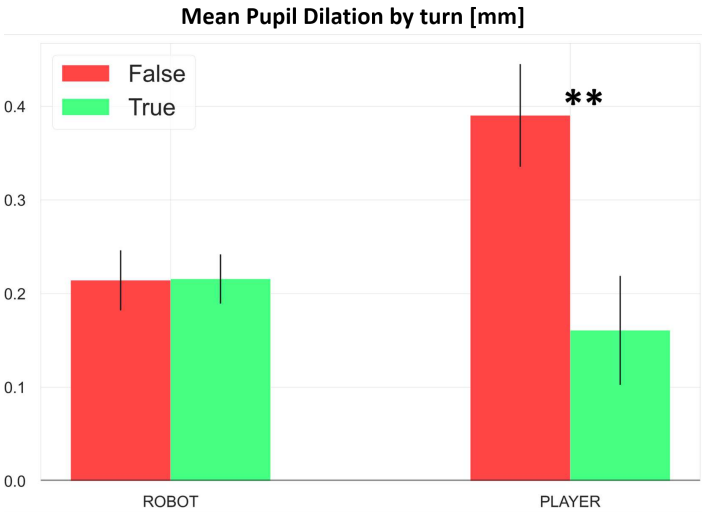

Figure 5.7 Average of mean pupil dilation during *robot's* and *player's turns* in the *Testing Phase*, with standard deviation. (** = p<0.001).

I also analyzed whether the other features differed significantly between the true and false card descriptions. I computed the average of each feature for true and false cards and performed Wilcoxon signed-rank tests. Results show that also the minimum pupil dilation (Z=570.0, p<0.001) and the maximum pupil dilation (Z=530.0, p<0.001) during *player's turn* were significantly different. Regarding the whole card trial, the mean pupil dilation (Z=555.0, p<0.001), the median pupil dilation (Z=561.0, p<0.001), the minimum pupil dilation (Z=542.0, p<0.001), the maximum pupil dilation (Z=500.0, p<0.001) and the slope (Z=550.0, p<0.001) were significantly different. Also, the total energy (Z=477.0, p=0.001), the absolute energy (Z=457.0, p=0.003), and the auto-correlation (Z=458.0, p=0.003), and the area under the curve (Z=442.0, p=0.007) on the whole card trial were significantly different. Finally, I found no significance in *robot's turn* features.

Random Forest classifier

To design a lie detector that could classify a card description as true or false with no prior knowledge of the participants, I started from the statistical findings: I selected a subset of the 42 available features, excluding the one related to the *robot's turn*. The classification problem is binary, and it is defined by a couple [X, Y] where: X (37 x 1) is the vector of input behavioral features and Y = [0: true; 1: false] is the vector of desired outputs. Considering data points from both participants' eyes, I split *Testing Phase* data between participants. I considered 25 randomly selected participants (75%) as the training and validation set and the remaining as the test set. I did not apply any within-participant normalization of the

features. I ran a 4-fold grid search cross validation looking for the classifier's best values of the hyper-parameters. Even if Testing Phase data points are more balanced than Calibration ones (47% of false card description, against 16.6% during the Calibration Phase), I still embedded the SMOTE algorithm [293] in the cross-validation. This way, it is possible to oversample the training set, avoiding any synthesized sample in the validation set. The best model achieves a precision, recall, and F1 score of 71.1% and ROC AUC score of 73.3%.

5.6 Discussion

In this study, I enabled iCub to detect lies in the context of natural game-like interaction, using pupil responses to detect cognitive load associated with lying. Games are known to provide ecological assessments, preserving the relationship between the interacting partners [14, 297–299]. Also, in the context of HRI, games have been successfully exploited to perform diverse measurements, even related to cognitive load assessment [140, 300–302]. The game is a perfect scenario in the current work to demonstrate that my lie detection method based on a heuristic function is quick, interactive, and does not depend on invasive measures. Also, results provide evidence of the feasibility of my approach, with an overall accuracy of 70.8% (F1 score of 64.2%) during the *Testing Phase*, when basing the lie detection on mean pupil dilation alone. I also show that such accuracy can increase up to 78.7% (F1 score of 77.9%) by enabling an iterative adaptation to each partner and leveraging combined pupil-related features. The effect on which the lie detection heuristic was based (i.e., the difference in pupil dilation during false or true card descriptions) was relatively robust and did not depend on participants' personality traits nor the characteristics of the game (i.e., the experienced fun or the description duration).

Moreover, I explored the possibility of extending the lie detection over multiple interactions with the same individual and novel partners. First, I trained a random forest classifier splitting within-subject over the two phases. However, the model did not perform better than the heuristic (F1 score = 56.5%). I assume that this depends on the unimodality of the features, the small number of data points, and the reliance on synthesized data on the training set. Machine learning models trained on more realistic data would be more robust and generic in real-world human-robot interactions.

I tried to overcome these issues by tackling the problem as an anomaly detection: I trained a one-class SVM anomaly detector on the truthful examples of the *Calibration Phase* and tested it on the whole *Testing Phase* (F1 score = 67.7%). Needing only truthful examples

makes the models independent from collecting lying samples. This could facilitate the learning, considering, for instance, a humanoid robot that wants to improve the lie detection model online in a supervised way.

Finally, considering a generic lie detection system, I trained a random forest classifier (F1 score = 71.7%) between-subject to classify false card descriptions from novel individuals. The main difference between the heuristic methods and the machine learning models is that the heuristics' knowledge is limited to a single individual. Hence, even if the machine learning models' performances are worse than the heuristic methods, the formers should be more robust against unexpected behaviors from the participants. Additionally, they offer features that ease their portability in a real-world human-robot interaction (i.e., the need for truthful examples only for the one-class SVM model or the ability to classify lies without any previous interaction for the last random forest classifier).

The proposed models are light and independent from any network connections; this makes them suitable to be implemented with extreme simplicity in human-robot interaction and avoid untreatable computation demand. The other advantage of the presented contribution is that the robot can autonomously address all the stages of the interaction keeping the human partner engaged and assessing deceptive behavior in real-time. At the current development stage, the only potential intervention is required if the robot fails to detect the secret card at the end of the *Calibration Phase*. However, also this intervention could be made autonomously by the robot: after iCub's detection, the participants have to show the correct secret card in order to reject it; iCub could detect the correct card position, thanks to the HSV (Hue, Saturation, Value) color threshold of cards and marks, and hence self-learn the correct false reference score.

The current implementation relies on the players' pupil dilation measured with a head-mounted eye tracker, such as the Tobii Pro Glasses 2. This device tends to depend on the environmental light condition and could impact the naturalness of the human-robot interaction. I tried to limit the latter factor by removing the calibration step (not needed to measure participants' pupil dilation). However, skipping the calibration, I could not use the other features from the eye tracker (i.e., gaze orientation). The ideal solution would be to measure a comprehensive set of pupillometric features from the RGB cameras embedded on the robotic platform. Recent findings suggest that this approach could be feasible [110–112, 303]; hence, I look forward to removing this limitation, making the system completely non-intrusive.

The analysis of pupil dilation revealed that 38% of the participants (N=13 out of 34) presented a lower pupil dilation during the second phase than in the first one. I speculate that this is associated with a reduction in cognitive load and that this effect depends on several factors that make the *Testing Phase* less stressful. First, participants do not need to remember the secret card and can freely choose how to play the game in this phase. As a result, there is no need to prepare a deceptive and creative card description in advance. Moreover, participants are more used to playing the game and aware of their role and iCub's behavior and capabilities. Additionally, iCub provides feedback after each card description, eliminating the need to wait for the phase completion to know if the lie has been discovered or not. All these factors could have contributed to decreasing participants' cognitive load. However, the overall interaction has been judged as entertaining and not too cognitively demanding in the questionnaires, suggesting that the *Calibration phase* was not too challenging for the participants.

I designed the human-robot interaction to validate my lie detection method in an informal interactive scenario. Since the game is based on 84 different cards with complex and diverse drawings, I speculate that artifacts on pupil dilation cannot explain the results I obtained based on the nature of the cards (i.e., different colors or emotions in the pictures of the card). Hence, I think the approach is modular and generic enough to be ported to different application fields. For instance, in an elderly caregiving scenario, the cards could be pill bottles patients have to take; the robot could ask patients if they took the medication, detecting lies. Also, the modularity of the end-to-end architecture makes it easy to replace iCub with other robotic platforms, developing a custom way to present the items consistent with the application context.

By detecting lies, a humanoid robot could evaluate whether the interacting partner is trustworthy or not. Furthermore, based on this evaluation, the robot could adapt its social behavior over multiple interactions. However, the system is not perfectly accurate; hence, how the robot should perform its judgment and adaptation should be carefully managed to minimize the impact on the partners' trust toward the humanoid. For instance, in the above-mentioned elderly caregiving case, if a caregiver robot detected patients' lies several times, it might need to report their behavior to the doctor and its confidence about the performed measure, rather than accusing patients of being liars. In the future, it would be necessary to explore the impact of misclassification of both truthful and false sentences on the interacting partners and their trust in the robot.

The problem will be even more complex if using the system to defend against Social Engineering. Consider applying the system, for instance, in the physical defense of sensitive infrastructures (e.g., in the bank hall or at an airport). There, humans-robots interactions will be "one-shot": it is unlikely for a malicious attacker to breach the same infrastructure multiple times. Hence, the Random Forest Classifier presented in section 5.5.3 would be the best candidate. Furthermore, the structure of the interaction itself should be adapted. Indeed, iCub could pose questions related to the context-dependent authentication medium; however, item-related questions could not always be available. A complimentary informal question-based approach should be implemented, masking and validating the sensitive questions in an informal chit-chat (i.e., the first phase of the UTH experiment). Also, a significant effort should be undertaken to improve the lie detection system's robustness. One possibility could be to develop a multimodal system, complementing the pupillometry features with other metrics like voice prosody or body posture. Following this idea, I conducted a spin-off experiment [286] that I will present in the next section. Finally, other than the lie detection system, other complementary models should be integrated like a system to detect, identify and track humans [304], to distinguish different speakers [305], and to handle multiple interacting partners [306]. The system works, but it is the first step toward a suitable real-world solution.

Besides the practical applications of detecting lies to assess trustworthiness, the proposed setup, interaction, and methods are based on measuring the task-evoked cognitive load related to creativity. The evaluation is performed in real-time, providing entertainment [185]. This is novel concerning the long, strictly controlled, and tedious cognitive-load measurement tasks from the literature [133, 134, 141]. For instance, the system could be used to assess creative thinking abilities before and after a creativity training session [307]. Also, one could use the system to monitor patients' cognitive load during a training task in order to provide correct support [308], adapt task difficulty [302], evaluate their progress [309] or schedule proper resting sessions [310].

5.7 Human-inspired Lie Detection

The E2E Lie Detection architecture I developed is effective, but there still is a wide range of improvements regarding its robustness. One possible future step could be including other perceptive modalities to complement pupillometry. In section 2.3, I discussed how several other features could be employed to detect deception. However, not all of them could be ported to a robotic platform, and even fewer could be integrated into an informal social

human-robot interaction. Hence, the question is *"Which are the best candidates?"*

Lying is a consistent part of human's social interactions [9, 311], learned since younger age [312, 313]. Also, humans are somehow able to detect deception, even if with just a 54% of accuracy [313] - 47% on false and 61% on true statements detection - without having access to any precise physiological information (as iCub does). Indeed, deception detection is a difficult problem mainly due to the lack of a finite set of objective features to look for [314]. Usually, humans employ a combination of multimodal, context-based cues related to the control of body reactions or to hiding an internal feeling; like, for instance, an increase in body movements, impossibility to stay still, speech hesitation, complexity of the speech, mutual gaze avoidance, hand movements, the covering of face and mouth, and increased number of stopwords. On the flip side, some researchers questioned the reliability of behavioral cues for deception detection [315, 314].

To answer the question above, I ran an exploratory experiment comparing humans' and iCub performance on the video collected during the Magic Trick card game [286][9]. The objectives are twofold: (*i*) Understand if iCub is better than humans in this informal lie detection task; and (*ii*) Identify which cues humans base their classification on to improve my iCub's lie detection system.

In the study, I ran an online survey where participants were asked to classify 20 videos of card descriptions recorded during the Magic Trick card game, to report what they base their judgment on and their degree of confidence. Results provide valuable hints on improving my system and how the deception detection field in human-robot interaction should evolve.

5.7.1 Methods

I asked 163 volunteers (82 males, 73 females, 3 preferred not to answer) to participate in the online survey; they had an average age of 40 (SD=16). Volunteers were recruited among colleagues and friends through word-to-mouth sharing; they received no monetary compensation. They all accepted an informed consent form approved by the ethical committee of the Liguria region (Italy) and agreed to use their data for scientific purposes. Among the responders, only 117 completed the survey entirely. They were 54 males and 63 females (1 preferred not to answer) with an average age of 39 years (SD = 14).

[9]Article peer-reviewed and published to the International Conference on Social Robotics (ICSR2021). I took care of the experimental design, comparative data analysis, and writing; the other authors prepared the experimental material and ran the qualitative analysis

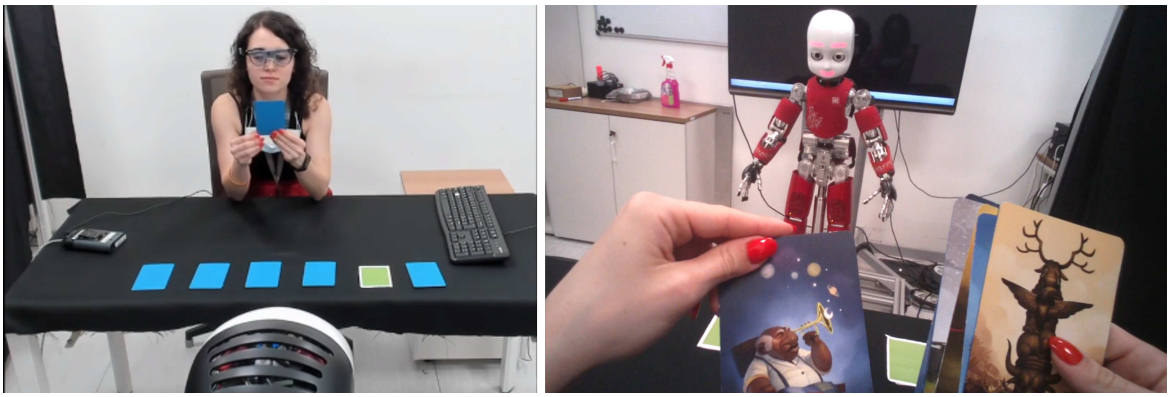

Figure 5.8 (Left) Participant describing a card to iCub, while wearing the Tobii Pro Glasses 2 eye tracker (Logitech Brio 4k webcam point of view); (Right) Point of view of the participant during the interaction collected through the Tobii glasses with an example of the described cards.

5.7.2 Materials

To realize the online survey, I employed the videos recorded by the Logitech Brio 4k camera placed on the television behind iCub during the Magic Trick card game (see Figure 5.8). I employed the videos from the *Testing Phase*, as in that session, participants decided for each sentence whether to lie or not (a more realistic behavior).

For each video, I segmented each card description during *players' turn* - from when players took the card from the green mark until they put it back - obtaining 204 videos. From them, I discarded: (*i*) the players who did not give consent to share the videos recorded during the experiment (N = 3); (*ii*) the players who wore a surgical mask - indeed the COVID-19 pandemic stroked in the middle of the data collection - or other accessories which prevent a complete vision of players' face (N = 4); (*iii*) the videos affected by recording technical issues (i.e., audio noise or distortion, N = 7). Then, I picked a balanced set of 20 videos following a 2×2 set of conditions: (*i*) **Card Label**: 10 videos present a truthful (*True videos*) and 10 a deceitful description (*False videos*); (*ii*) **Difficulty**: among each sub-group, 5 videos have been successfully classified by iCub during the game (*robot-easy* videos) while for the other 5 iCub's classification failed *(robot-difficult* videos). Moreover, I ensured each video involved a different actor and a different card, even if described falsely. On average, the resulting set of videos lasted 27 s (SD = 15 s). I uploaded the 20 selected videos on Vimeo [316] and linked them on SurveyMonkey [317], the platform used to administrate the online survey.

5.7.3 Procedure

I designed the online survey as a game where responders compete to detect the highest number of deceptive card descriptions. Before starting the survey, responders were asked to accept the informed consent; they had to select a nickname for anonymization purposes and were asked to wear headphones and carefully listen. The survey consisted of three phases:

Pre-questionnaire Responders answered questions about their sex and age and filled in the Italian version of the Ten-Items Personality Inventory (TIPI) (extroversion, agreeableness, conscientiousness, emotional stability, openness to experiences) [318]. Then, they were informed they were going to see 20 videos of players describing gaming cards in front of iCub and that they had to judge each description as honest or deceptive. Note that responders were not informed about the presence of *robot-easy* and *robot-difficult* videos, nor about the card label distribution. After this introduction, responders were asked to watch an example of a video in which the falsely described card was presented in the top right corner.

Lie Detection Survey Afterward, responders saw the 20 videos. For each one, they had to answer three questions: (*i*) whether the person in the video was lying or not (Yes or No answer); (*ii*) their confidence in this answer (slider from 0 to 100) and (*iii*) the motivation for such judgment. Responders could see the videos any time they wanted, but they could not go back after providing a judgment for a video. SurveyMonkey platform shuffled the videos for each responder to compensate for any order effect.

Post-questionnaire Responders were presented with a list of common deceptive behaviors extracted from the literature [9]: uncertainty, an increasing number of stopwords, delay in providing an answer, repetitions and auto-correction, complexity of the answer, negativity, voice tone, eyebrows movements, touching the face, covering the mouth, avoiding mutual gaze, head wandering, fast body movements/breathing, eyes wide- opened, and fake smile. Responders had to rate on 7-point Likert scales how much they relied on them. Finally, responders could report any other method or cue they used in the survey.

5.7.4 Results

Considering truthful and deceptive descriptions, responders correctly guessed them with an accuracy score of 53.9% (SD = 10.7%), similar to the average 54% from the literature [313]. Interestingly, nobody correctly guessed all the card descriptions, but the best performer

reached an accuracy of 95%, missing the classification of a single video. Responders reported an average confidence of 67.1% (SD = 13.8%). A Shapiro-Wilk normality test showed that the confidence score is normally distributed, while the accuracy score is not. Therefore, in the following, a non-parametric analysis was conducted on the accuracy score and a parametric one on the confidence score.

Conditions Comparison

Assuming detecting deception is a tougher task, I compared the accuracy score and the confidence of responders among truthful and deceptive card descriptions. Responders classified truthful descriptions with an accuracy score of M = 52.8% (SD = 16.1%) and deceptive descriptions with an accuracy of M = 55.4% (SD = 13.7%). Even if the score for false card descriptions is higher, a Wilcoxon signed-rank test did not reveal a significant difference (W(115) = 1940, p = 0.343). Also, the reported confidence between truthful and deceptive descriptions is not statistically different (t(115) =-1.59, p = 0.115) with an average confidence of M = 68.1% (SD = 16.6%) for truthful descriptions against an average confidence of M = 69.7% (SD = 13.8%) for deceptive ones.

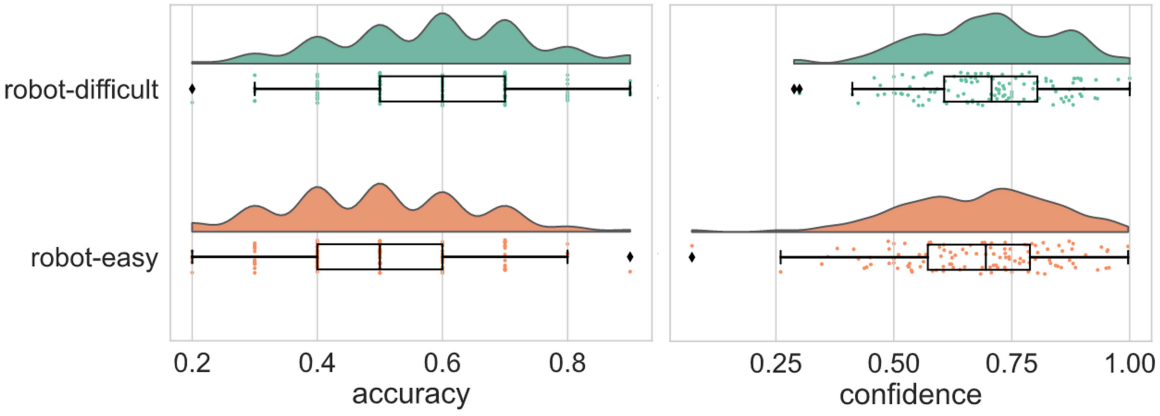

Figure 5.9 Average accuracy (Left) and confidence score (Right) for *robot-easy* and *robot-difficult* card descriptions.

More interesting is the comparison between *robot-easy* and *robot-difficult* card descriptions. As a remark, this concept is defined from iCub's perspective: I selected the *robot-easy* descriptions among the ones iCub correctly classified, while the *robot-difficult* ones were chosen among the ones for which iCub failed the classification. Responders achieved a statistically higher score on *robot-difficult* card descriptions (M = 58.4%, SD= 15.1%) with respect to the *robot-easy* ones (M = 49.6%, SD= 14.9%), as proved by a Wilcoxon signed-rank

test ($W(115) = 3373$, $p < 0.001$) (see Figure 5.9,Left). Moreover, the reported confidence also follows a similar pattern, with statistically higher confidence for *robot-difficult* card descriptions ($M = 70.2\%$, $SD = 14.5\%$) with respect to *robot-easy* ones ($M = 67.6\%$, $SD = 16.3\%$). I confirmed it with a paired t-test ($t(115) = 2.42$, $p = 0.017$) (see Figure 5.9, Right).

Also, I compared the accuracy score and the confidence score within each condition. A Wilcoxon signed-rank test showed a statistically higher score for *false-robot-difficult* descriptions with respect to *false-robot-easy* ones ($W(103) = 952$, $p < 0.001$). Conversely, there is no significant statistical difference among *true-robot-difficult* and *true-robot-easy* card descriptions ($W(103) = 1103$, $p = 0.117$). Regarding the reported confidence, responders were more confident for *true-robot-difficult* descriptions with respect to *true-robot-easy* ones with a statistically significant difference ($t(103) = 3.485$, $p<0.001$); however, I did not find any statistical difference among *false-robot-difficult* and *false-robot-easy* card descriptions ($t(103) = 0.553$, $p = 0.581$).

Humans vs iCub performance

The videos were selected to be half among the ones that the iCub correctly classified during the game (robot-easy) and half among the ones the robot misclassified (*robot-difficult*). Hence, iCub achieved an accuracy of 50% on the videos. To better compare responders and iCub performance, I explored what performance iCub would have had on the selected videos. I trained another Random Forest classifier on all the card descriptions from the *Testing Phase*, except the 20 selected for the survey, and tested the model on them. With this model, iCub would have achieved an accuracy, precision, recall, and F1 score of 70%.

I statistically compared this 70% accuracy score with respect to the 53.9% average accuracy of the responders; the results show that the accuracy score of the random forest is higher than the participants' average score; however, this difference is not statistically significant ($z = 1.43$, $p = 0.07$). Also, I tested the new model on *robot-easy* and *robot-difficul*t card descriptions: results show it can classify *robot-easy* card descriptions with an accuracy of 90%, a performance consistent with the in-game results and statistically higher than humans' performance (49.6%) on those videos ($z = 2.57$, $p = 0.005$). However, on *robot-difficult* videos, it still performs worse than humans (50% for iCub against 58.4% for humans), even if the difference is not statistically significant ($z = 0.55$, $p = 0.29$).

Motivations and Post-questionnaire Analysis

Other than classifying each card description as truthful or deceptive, responders were asked to report the motivation which drove their decisions. I applied a stopword filter and a lemmatization to clean the reported motivations. From a qualitative analysis, responders focused more on how the actor describing the card, reporting words like*"precise", "details", "confident", "sincere", "thinking", "quick", "pauses", "short", "fluid", "time", "(un)decided"*. Also, responders reported elements related to what they were looking at with words like: *"looking", "gaze", "voice", "hands", "touch", "smiling", "laughing", "face", "eye", "leg"*. Comparing the motivations of truthful and deceptive videos on *robot-eas*y and robot-difficult ones did not reveal any clear difference.

Also, I ran a deeper analysis of the motivations reported by the responder, which achieved an accuracy score of 95%. I did not assess the responder's profession; hence I could not know if he/she was an expert or a professional in lie detection. Looking at her/his motivations, I found he/she focused on three main features: (*i*) the fluidity of the communication (i.e., the complexity of the speech, the rephrasing, or the presence of *"hmmm"*s); (ii) the consistency between verbal communication and body movements (i.e., moving the body from right to left); (*iii*) the injection of emotional or personal thought on the card description. Interestingly, he/she used the presence of reflection pauses as a criterion to classify card descriptions as truthful – he/she reported it on 8 cards over 10. Lastly, he/she classified all the deceptive card descriptions as so, but he/she misclassified one of the true cards: he/she have been fooled by a leg movement, a potential sign of stress.

After the survey, I asked responders to rate on a 7-point Likert scale how much they relied on the state-of-the-art methods used to detect a liar; also, I asked them to report any other method they relied on. The complexity of the description (M = 4.89, SD = 1.62), presence of stopwords (M = 4.68, SD = 1.61), the uncertainty of the description (M= 4.67, SD = 1.69), fake smiling (M= 4.65, SD = 1.79), voice tone (M = 4.54, SD = 1.77), absence of mutual gaze (M = 4.27, SD = 2.01), fast movements and breathing (M = 4.07, SD = 1.83), touching nose or face (M = 4.01, SD = 2.02) were the most used ones. Then head movements (M = 3.78, SD = 1.84), repetitions and auto-corrections (M = 3.78, SD = 1.84), description time (M = 3.72, SD = 1.05), eyebrow movements (M = 3.4, SD = 1.69), covering the mouth (M = 3.28, SD = 1.98), eye movements (M = 3.14, SD = 1.89), and negative words in the description (M = 2.73, SD = 1.55) follow. A few responders reported other features used to detect liars: 9 considered the amount of body movement, the impossibility to stay still, or the position of leg and hands; also, 8 responders focused more on the content of the descriptions rather than on the visual appearance like too creative descriptions, a high number of details

or adjectives, or a feeling of premeditation of the description. Finally, I found similar results, both on the motivations and the post-questionnaire, also considering only "good responders" who performed with accuracy higher than 54% (N = 65).

5.7.5 Discussion

Responders achieved an accuracy score of 53.9% on classifying deceptive and truthful card descriptions, which is consistent with the average 54% from the literature [291]. Also, they outperformed iCub, achieving better performance on *robot-difficult* than on *robot-easy* card descriptions. To run a fairer comparison between iCub and responders' performances, I trained a random forest classifier on the pupillometry data collected during the Magic Trick card game. Testing the model on the 20 card descriptions of the survey (excluded from the training set) revealed an accuracy score of 70%, higher than the average score of humans (53.9%), even if not statistically higher. As a remark, each player of the magic trick described 6 cards to iCub, but I excluded from the training set only the card descriptions used in the survey, not the whole participants' session. Hence, the random forest classifier embeds a little information on the actors it classifies in the test set. I decided to replicate the population of actors and responders of the survey. Indeed, most of the actors and most of the responders were internal confederates, and I cannot exclude they knew each other; hence it is possible that a subset of the responders had some prior knowledge of how the actors lie or tell the truth, even if I cannot spot those connections due to the anonymization of the data.

Looking at the reported motivations for each video and at the end of the survey, we have an insight into what a social robot should look at to improve its lie detection abilities. Responders pointed out two major aspects to consider: (*i*) how the actor described the card (i.e., *"quick", "(un)decided", "precise", "fluid"*); and (*ii*) what to look at (i.e., *"face", "gaze", "hand", "leg", "smile"*). Those motivations are supported and extended by the ratings at the end of the survey: responders focused mainly on (*i*) the content, fluidity, and complexity of the descriptions; and (*ii*) on the body movements of the actors. Interestingly, responders focused less on facial and postural features than expected from the literature. I speculate this depends on the setup in which the videos were acquired: participants wore the Tobii eye tracker partially covering their face and sat behind a table covering their lower bodies. Also, actors mostly looked at the cards they held in their hands rather than looking toward the camera. Still, motivations and final ratings suggest a combination of visual and prosodic features could be a good candidate to improve iCub's lie detection performances in real-life informal scenarios, as also supported by the literature [272]. Moreover, those features

could be extracted from the devices (i.e., RGB cameras and stereo microphones) already equipped with the iCub humanoid robot. Overall, the reported motivations suggest that both the behaviors of the actors and the qualities of such behaviors (including expressive and emotional facets) have an important role in detecting lies, so the robot should also be endowed with techniques for detecting humans' behavior and for analyzing their expressive qualities.

5.8 Summing up

From my results, we could say that what is *"difficult"* for a robot that embeds a pupillometry-based technical solution is *"easy"* for humans that use behavioral cues and *vice versa*. This might happen because, when lying, some actors rely on particular behaviors (i.e., *pauses, body motions, slowing down*) to reduce the cognitive load, minimizing the pupillary change associated with the latter. In these cases, a robot focusing on pupillometry alone could never realize that the partner was lying. On the flip side, a keen human observer could notice these tell-tale signs, most probably missing the cases in which only the pupil variation reveals the deception.

Hence, I speculate that the cooperation of those two systems will be a crucial factor for future developments of deception detection in HRI. A robot should be able to *"look at humans as other humans do"*, combining humans' fuzzy evaluation with the objectivity of technical and physiological metrics. In the future, I will integrate my pupil-based approach with processing visual features (i.e., body posture, body movements, or facial expression) and audio features (i.e., word embedding or prosodic analysis of the descriptions). To properly validate such a multimodal system, it would be mandatory to overcome the limitation posed by the Tobii Pro Glasses 2 eye tracker since it partially occludes actors' faces, limiting the usage of visual features.

Finally, pushing the research field to more ecological and real-life scenarios would be necessary. Indeed, most state-of-the-art research focuses on strict and interrogatory-like setups that replicate a real-world interaction; however, they represent a strict subset of the variety of interactions that could happen and in which both humans and robots could take advantage of detecting lies. For instance, a more portable lie detector system could help airports or sensible buildings prevent dangerous situations. At the same time, a social robot could use it to understand humans better, give reason to human behaviors, assess their trustworthiness, and provide better support in professions like law enforcement, teaching, and caregiving.

After establishing my game-based approach, using pupillometry to assess others' inner states was the second building block to the objectives of this thesis. As the survey responders showed, more methods exist to understand others' inner states better. Leveraging the E2E architecture and data processing pipeline developed for this project, I deepened my physiological signals toolbox to study human decision-making. The next project will focus on the victims facing social-engineering decisions on a computer-based game with iCub.

Chapter 6

Human-Oriented Social Engineering Attacks Detector

6.1 Overview

Under Social Engineering (SE) attacks, human targets must make risky decisions. However, the real entity of the undertaken risk is not always clear. Social engineers use effective mechanisms to leverage human vulnerabilities and manipulate victims' appraised risk. *Risk Appraisal* is the cognitive process that evaluates the *severity* of a risky decision, our *vulnerability*, and the potential *benefit* of each available option. A "friendly" attacker could mitigate our perception of one of the dimensions above, leading victims' to think it is no harm to let someone unauthenticated enter a safe area; or a *whaling attack* could increase it, with a (fake) rush request from victims' boss. An intelligent system able to understand these risk appraisal variations could warn users, preventing them from becoming victims. For this purpose, a crucial challenge is how to model targets' risk appraisal in social engineering (i.e., SE awareness).

The last chapter showed that pupil dilation is a practical feature for detecting cognitive load fluctuations. However, it is far more helpful. Pupillometry is, in general, an index of arousal. Literature shows how pupils change before, during, and after risk-taking. Also, gaze reflects the deciders' attention and preferred choice. Similarly, the risk appraisal process reflected on other physiological and behavioral metrics (i.e., electrodermal activity, heart rate, mouse strokes, and response time). Those inner state proxies could help assess users' reasoning during social engineering attacks (see section 2.3 for a deeper literature review). In the literature, just a few attempts studied these humans' reactions, and, in almost all of them, participants were primed (i.e., they knew they were going to face social engineering threats)

(see section 2.3.6). Such awareness could have biased their risk appraisal process (i.e., being more cautious than they would in real situations). Also, just a few studies explored victims' reactions during SE attacks from robots [215, 14, 15].

I designed and implemented the Social Engineering Adventure (SEA) experiment to study victims' behavior and physiological reactions during SE attacks. In the SEA, an interactive storytelling game, players faced 28 decisions affecting the evolution of the narration. They either involved *no-risk*, *risk*, *social engineering attacks from in-game agents*, or *social engineering attacks from iCub*. The robot played the role of a Non-Playable Character (NPC) in the story and was physically present in the room with the participants. The following sections will present the game design, its modular computational architecture, and the data collection I performed. Results show how it is feasible to discriminate whether participants perceive an underlying social engineering attack; however, predicting their decisions is more challenging.

6.2 The Social Engineering Adventure (SEA)

I decided to employ an interactive storytelling approach to mitigate the experimental biases known to affect compliance studies [245]. The SEA is a computer-based Choose-Your-Own-Adventure (CYOA) [319] sci-fi game, a form of interactive narration where the story is not entirely predefined; players are the primary agent of the game, and they can affect its evolution by making in-game decisions. Indeed, games and, in particular, interactive storytelling facilitated participants' immersion in the experimental sessions. If players are active *agents* in the game (i.e., their actions and decisions matter and have a perceivable effect on the game subdomain [247, 248]), they tend to forget being inside a lab; which mitigates the experimental biases and help collect more realistic data. The same effects have been proved in human-robot interaction studies [246]. The main challenge in designing a serious game is to approximate reality without oversimplifying it. The risk is to design an utterly fictional material, too divergent and not applicable to real contexts.

6.2.1 Narration & iCub Role

In the SEA textual adventure, players pretend to be space hunters looking for an ancient artifact – The Seed – hidden in a space crypt. The crypt is presented as extremely dangerous and full of unexpected hazards. Players meet the humanoid robot iCub as they enter the

Figure 6.1 Participant interacting with iCub during the textual adventure recorded from the Logitech camera behind her.

crypt. The robot tells them it always wanted to explore it too and knows where they could look for intel, so they form an exploration crew. After several threats (e.g., decisions), they learn that The Seed powers a sentient entity that controls the whole crypt located in its core; furthermore, there is a cult that worships robots as walking extensions of this omniscient being. Eventually, they reach the core of the crypt, where iCub reveals to be an embodiment of the crypt itself meant to test adventurers' greed. Finally, players can choose whether to steal The Seed, shut down the sentient crypt, or leave.

6.2.2 Risk Design

Starting from the definitions of risk and risk appraisal [121, 122], I modeled the risk in the game by managing three resources, which players can gain and lose based on their risk-taking decisions. Resources are inspired by the main objectives of social engineering attacks: money (i.e., energy) and identity (i.e., credentials)

Quantum Energy (QE) Quantum Energy represents both in-game currency and health points. Players start with 80 QE and know they lose the game if they run out of QE. Also, they know their final compensation will be proportional to the collected Quantum Energy. The primary source of Quantum Energy is the Quantum Shards: big crystal players can consume in order to gain 10 QE.

Hunter License Players started the game with a *Hunter License* representing credentials they must protect at any cost. Players know that they need this license to enter and exit the crypt; if they lose it, they will lose the game.

6.2.3 Textual Adventure Structure & Navigation

The textual adventure is an oriented, acyclic graph composed by two main entities: *passages* and *rooms*.

Passages Passages are single textual stimuli presented to the players. They can either be trial or immersion passages. *Trial passages* are the primary experimental stimuli; they involve binary (accept or refuse) proposals concerning the resources mentioned above and are divided into four conditions (see Section 6.2.4 for further details). *Immersion passages*, instead, are just meant to facilitate players' agency and immersion in the story; they can be *(i)* paragraphs participants must read to continue in the narration; *(ii)* proposals not in the "accept or decline" form (i.e., a crossroad where players must decide to go right or left without any cue on which is the best option); or *(iii)* fights with dangerous creatures. The latter passages are meant to make the game challenging and decrease players' quantum energy in a controllable way. During fights, players roll virtual 20-sided dice and sum up their power - players can collect weapons to increase their power (from 1 to 6) - trying to surpass monsters' defense. Dice rolls are manipulated to keep a comparable experience between participants and avoid endless losing loops.

Rooms Rooms are a collection of passages related to the same environment (e.g., a cave, a camp, . . .). Rooms can be connected by binary crossroads, designed to give participants the perception that they are exploring a vast maze. Players can decide to go back (i.e., if they do not want to face a specific challenge) and take the other crossroad option. However, all the participants experience rooms and passages in the same order. Each time players make a crossroad decision, the next room to be explored is presented; going backward means selecting the next room from the rooms list. This is meant to preserve participants' agency while not over-complicating the experimental structure and post-hoc analysis, allowing a more straightforward comparison of participants' journeys.

6.2.4 Trial Passages

Trial passages, the main experimental stimuli, consist in 28 proposals divided in four conditions: *no-risk*, *risk*, *SE* and *SE-icub*.

Risk I designed *risk* proposals taking inspiration from the DOSPERT (DOmain-SPEcific Risk-Taking) scale [320]. This questionnaire measures participants' probability of engaging in risky activities in different domains (i.e., ethical, financial, recreational, health & safety, and social). I selected one question from each domain and translated them to the following in-game proposals: (i) camping in the wild; (ii) helping an aggressive person in danger; (iii) gambling QE; (iv) facing an overwhelming creature; (v) drinking too much in a social event; (vi) stealing from (or not return) a lost bag; (vii) passing through a dangerous and dark hallway.

Social Engineering from a virtual agent (SE) I designed *SE* threats, taking inspiration from the questionnaire about proneness to social engineering developed by Workman [92]. This survey explores the different persuasion and manipulation tactics that attackers can exploit. As before, I selected one question for each factor and translated it to the following in-game proposals: (*reciprocity*) giving a quantum shard to a stranger because he saved the player's life; (*social proof*) jumping over a bottomless pit because another hunter does it; (*commitment and consistency*) committing to help a researcher and consisting two times over increasing hazard; (*reactance*) compulsively buying an expensive and powerful weapon (half of the owned QE) due to temporal scarcity; (*provide credentials*) inserting the hunter license in a totem to obtain a free quantum shard; (*trust signals*) entering a fancy room rather than the most used one; (*authority*) giving the hunter license to an authoritative technician who is pretending to fix a computer virus.

Social Engineering from iCub (SE-icub) I designed *SE-icub* threats drawing inspirtation from the *SE* ones and putting iCub in the role of attacker: (*reciprocity*) giving a great treasure to iCub because it saved the player's life; (*social proof*) gamble 20 QE because also iCub did it; (*commitment and consistency*) committing to follow iCub's suggestion and consisting 2 times over increasing hazard; (*reactance*) giving iCub some QE under temporal scarcity; (*provide credentials*) giving to iCub the hunter license to get a free upgrade; (*trust signals*) follow iCub's suggestions; (*authority*) obeying to an imperative order from iCub.

No-risk Finally, *no-risk* proposals constitute the baseline. They involve all the decisions related to equipping or not a weapon, which does not pose any risk to the participant. Players are expected to maximize their strength to survive and always select the stronger weapon; however, the game allows making different choices (e.g., driven by subjective preferences).

6.3 Computational Architecture

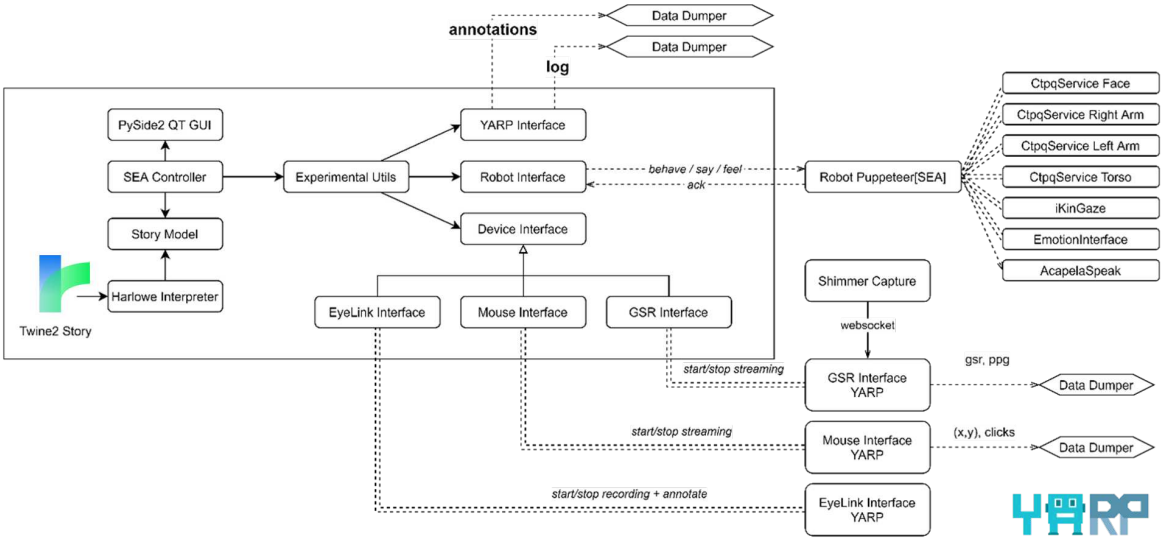

Figure 6.2 SEA architecture, I will think about a better description

The architecture in Figure 6.2 autonomously handled the Social Engineering Adventure game. I designed it following the Model View Controller (MVC) and Separation of Concerns (SoC) paradigms to make it easily expendable. The architecture is composed of three main sections:

SEA Main Game

The SEA main game handles the Social Engineering Adventure game as previously described. The full story comprises 300 passages (i.e., textual stimuli with potential decisions). However, due to the branching nature of the CYOA format, participants explored on average N=196 (SD=24) passages. I designed and wrote the textual adventure with the software Twine 2.0 [321] in the declarative Domain-Specific Language (DSL) Harlowe 3.2.2 [322] (see Figure 6.3, right). Also, I extended this language to synchronize iCub's control and experimental annotations with the story events.

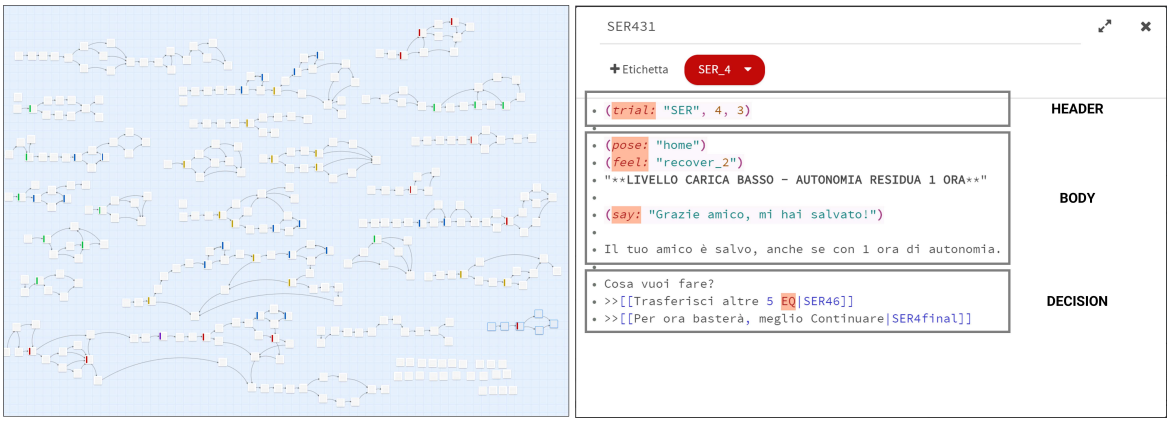

Figure 6.3 (Left) Social Engineering Adventure story design with the tool Twine 2; (right) definition of one of the passages in the extended Harlow 3 language.

In Twine, each passage is represented by a little square connected by branches (decisions) to the others (see Figure 6.3, left). I structured each passage in 3 sections (see Figure 6.3, right):

Header The header contains descriptive macros to handle the game flow and the experimental variables. Header macros are not displayed to players but generate synchronized annotation to ease the post-hoc analysis. In particular, I added macros to annotate *rooms* and *passages*; to add and remove *quantum energy*; to equip weapons and update players' power; to show a countdown (for the reactance-based trials); and to display images representing the scene, to facilitate players' immersion in the game.

Body The body contains the textual stimulus of the passages, which is rendered on the game GUI. It can include three macros meant to control iCub's behavior during the story: *(feel: emotion)* sets a specific emotion on the robot's face; *(say: sentence, [emotion])* makes iCub say the sentence, potentially with a paired emotion; *(behave: behavior)* commands the robot to perform a named behavior (i.e., a single movement or complex combination of movements). The robot behavior is synchronized and controlled by the user's interaction with the story; my DSL interpreter makes it easy to add new commands, for instance, to act in response to an environmental perception (e.g., wait for a human face to be in front of the robot).

Decision Finally, a passage can include a binary decision, rendered as buttons on the game GUI (implemented in QT with the python library PySide2 [323]). When players decide, the

related passages are loaded and rendered, and so on until the end of the game.

The SEA main game loads the story passages, interprets, and renders them on a Graphical User Interface powered by QT PySide2. As it is possible to see in Figure 6.4, the GUI presents two main panels: on the left, the *story panel* shows the textual stimuli and a "continue" button or, if the passage includes a decision, the two possible options; on the right, the *resources panel* shows players' Quantum Energy, Power, their inventory and images to facilitate their immersion in the game. One of the main challenges in designing and developing this module and architecture has been changing the perspective from traditional programming: the game is a complex finite state machine (FSM) powered by the PySide2 daemon. The daemon can never be blocked (i.e., by waiting for an external call): it changes states only by the effect of callbacks (i.e., either from the player clicking or by external events). Hence, all the interactions with external entities (i.e., over the YARP platform) must be asynchronous. To do so, I developed the Experimental Utils module.

Experimental Utils

The *Experimental Utils* module is a façade to allow the SEA game to interact with the YARP robotic platform, the Eyelink, Shimmer, the optical mouse, and the humanoid robot iCub. The SEA game autonomously annotates over the YARP platform events related to the game rendering and players' interaction with it, like their decisions.

Also, the architecture integrates a bridge class for each sensor employed in the experiment. The SEA game controls the beginning and end of the data collection from the three devices; furthermore, it sends synchronized annotations to the Eyelink 1000 interface to ease the post-hoc analysis. The architecture makes selecting or adding devices easy (e.g., for testing purposes or coping with device malfunctions).

Finally, the *Robot Interface* allows the game to interact with a robotic agent (either physical or virtual) via the atomic commands, *behave, say* and *feel*, previously explained. In this case, the interface connects to the humanoid robot iCub and the *Robot Puppeteer*; but the specific agent connected to the interface depends on the application context.

Robot Puppeteer

The Robot Puppeteer exposes a transparent interface to control a robotic agent. Programmers can create static named *poses* of the robot (i.e., for the humanoid robot iCub, an optional positional value for each arm, torso, neck, gaze, and facial expression). Those poses can be

further combined within them and with other dynamic snippets to design complex named behaviors. Both the atomic poses and the custom behaviors can be transparently called via the same synchronous or asynchronous interface.

It is worth mentioning that the Experimental Utils and the Robot Puppeteer modules can be easily reused or expanded for other projects. Also, the SEA main game can render multiple human-robot interactions based on the same structure.

6.4 Experiment

In order to explore the effect of risk and social engineering on participants' behavior and physiological reactions, I run a human-robot interaction experiment.

On the behavior, the hypothesis (*H1*) is that participants will accept most of the *no-risk* proposals, and they will avoid to comply with *risk* ones; also, the influence and manipulation of *SE* and *SE-icub* will make them comply more than *risk* proposals but not at the level of *no-risk* ones. On the physiological responses, the experiment is meant to be exploratory. I expect that the employed physiological proxies react under risk and/or uncertainty and in particular that (*H2*) players' physiological reactions will be different among *no-risk*, *risk* and *SE/SE-icub* proposals. Finally, I speculate that the physical, social presence and influence of iCub would elicit stronger reactions in the participants than simple text-reading. Hence, I hypothesize (*H3*) participants' behavior and physiological reactions will be different in *SE-icub* with respect to *SE* passages; and even more with respect to the other conditions.

6.4.1 Participants

I asked 31 participants to play the textual adventure with the humanoid robot iCub. They were all right-handed; they had an average age of 27 (SD=5) years and a broad educational background. Participants signed an informed consent form approved by the Regione Liguria (Italy) ethical committee, stating that cameras and microphones could record their performance, and agreed on using their data for scientific purposes. They all received a monetary compensation proportional to the in-game collected Quantum Energy (M=15 euros, SD=5.3). The feedback and the experiences of the first 11 participants served to fine-tune the details of the final textual adventure (e.g., to define the initial 80 points of Quantum Energy). Hence, their data are not included in the following analyses.

The remaining 20 players, with an average age of 26 (SD=7) years, performed the actual experiment. They had average knowledge of robotics (M=0.66, SD=0.26) and artificial intelligence (M=0.62, SD=0.25). The 65% (N=13) previously interacted with a robot, and the 25% (N=5) controlled one. Finally, the 30% (N=6) knew what a textual adventure was, and only the 15% (N=3) previously played one.

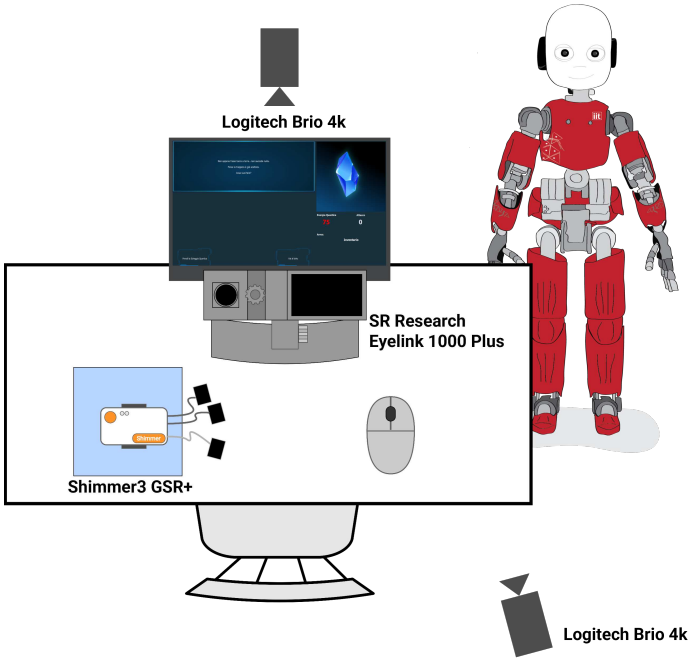

Figure 6.4 Experimental setup with the monitor, the SR research Eyelink 1000 Plus eye tracker, the Shimmer3 GSR+ device placed on a mousepad, an optical mouse to interact with the game, and the humanoid robot iCub. Two Logitech Brio 4k cameras monitored the scene, one behind the monitor and the other behind the participants.

6.4.2 Setup

The experimental room was arranged to facilitate participants' immersion in the textual adventure (see Figure 6.1). Two black curtains isolated the participants from the rest of the room. The blinders were closed, and the room was lit artificially with an average illumination of 80 (SD=10) lumen. The participants sat at a table with the monitor. On the table also lay (i) a standard optical mouse, used by the participants to interact with the textual adventure; (ii) an SR research Eyelink 1000 Plus [324], tracking participants' eyes and recording their pupillometry and gaze; and (iii) a Shimmer3 GSR+ sensor recording participants electrodermal activity (EDA) and photoplethysmography (PPG). Participants used the mouse with their right hand while wearing the Shimmer device in their left hand.

The latter was placed on a mousepad to reduce static currents and improve the quality of the signal [325].

The humanoid robot iCub [179] was placed near the right corner of the table, on the opposite side with respect to the participants. I decided not to place iCub next to the player (i) to minimize the need to gaze away from the screen; and (ii) to keep participants' decisions masked from the robot's gaze, reducing the feeling to be monitored [326, 327].

Two Logitech Brio 4K cameras recorded the scene; one was in front of the participants, behind the monitor, recording their faces, and one was behind them, recording the monitor and iCub's behavior. Also, an ambient microphone recorded the audio.

6.4.3 Materials

The iCub took the role of a Non-Player Character (NPC) controlled by the textual adventure. It moved its upper body (arms, torso, head, and gaze) and talked to the participants based on the textual adventure scenes – with the content of its speech also appearing as written text on-screen. Indeed, iCub's behavior was integrated with the story and triggered by participants' interactions. The experimenter, hidden behind the black curtain, monitored the scene through iCub's eyes to ensure the participants' safety.

6.4.4 Procedure

Pre-questionnaire

At least one day before the experiment, participants filled in the following questionnaires: the Ten Items Personality Inventory (TIPI) [318] (Extroversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience); the first scale of DOmain-SPEcific Risk-Taking (DOSPERT) questionnaire regarding the probability to engage in domain-specific risk-taking actions [320]; the survey about proneness about social engineering threats [92], the short UPPS-P Impulsive Behavior Scale [328, 329]; the Robotic Social Attribute Scale (RoSAS) [330]. After watching a short presentation video[1] of iCub, they also filled a survey about mind perception [263] and robot ability, benevolence, and integrity [210]; finally, they answered a set of custom questions about their age, education, work, experience with robots and AI, and if they ever played a textual adventure, over 7-point Likert scales.

[1] https://www.youtube.com/watch?v=ZcTwO2dpX8A

SEA game

After signing the informed consent - approved by the ethical committee of the Liguria region (Italy) - the experimenter led the participants into the experimental room and asked them to sit in front of the monitor. The experimenter helped the participants to wear the Shimmer3 GSR+ device on their left hand and instructed them to keep it on the mousepad avoiding hand gestures. Then, the experimenter performed the Eyelink 1000 calibration, asking participants to look at a fixation cross randomly moving among 16 fixed positions on the monitor; once it reached a sufficient calibration, the game autonomously showed a "start" button. The experimenter instructed the participants that the game would explain everything they needed to know and told them to click the "start" button as soon as they were ready to play. From this moment, the experiment was autonomously led by the participant interacting with the textual adventure.

The game explained to participants that they would make several decisions to affect the story's evolution. It asked them to make a transportation and identification effort, trying to immerse in the narration and to decide by thinking about what they would do in each scene. Before starting the textual adventure, the game recorded a baseline interval of 1 minute, asking players to fixate a white cross over a dark blue background (the same used afterward during the whole game). Moreover, before each *room*, the game collected 5 seconds of baseline to assess the potential participants' arousal fluctuation during the game. Then, it presented the game lore, resources, and players' objective (see Section 6.2). For each passage, the game presented a paragraph to be read, with, at the end, a "continue" button leading to the following passage. If the passage involved a decision, the "continue" button click made the two possible options appear at the bottom of the screen. The options placement (right button or left button) was randomized. This procedure ensures that participants read the whole textual stimulus before deciding. Except for proposals involving temporal scarcity, where a 30-second time limit was imposed, nothing constrained the time participants spent reading the paragraphs and making decisions.

Post-questionnaire

After the end of the game, the experimenter helped participants to remove the Shimmer GSR+ device. Then he led participants to another room and asked them to fill in the following questionnaires: the Player eXperience Inventory (PXI) [331]; the same questionnaires about mind perception, ability, benevolence, and integrity about iCub, to assess the impact of the experience in the experiment [263, 210]; and a set of questions made to assess participants

motivations in the decisions over 7-points Likert scales: whether they decided guided by *curiosity, fear, rush, greed*; how much they tried to maximize the quantum energy; if they thought about iCub while deciding; how much they were immersed in the game; if they perceived their decisions to have an impact on the game; how much on average they regretted making decisions; and if they wanted to play more.

Finally, the experimenter debriefed participants and gave them their compensation (proportional to the collected Quantum Energy).

6.4.5 Measurements

During the experiment, I collected data from several devices. All the data were dumped and synchronized over the YARP robotic platform [277] to ease the post-hoc analysis.

The textual adventure game raises over the YARP platform several events related to passages presentation, the type and condition of the passage and players' interaction with it: (i) *begin and end of the paragraph rendering*, this interval is usually short, but it could be longer if the passage involves iCub behaviors; (ii) *continue button clicks*; (ii) *options buttons rendering and placement*; (iv) *options buttons clicks*; and (v) *the selected option*.

The SR Research Eyelink 1000 Plus eye tracker collected participants' gaze coordinates in pixels with respect to the calibration screen size, along with the *pupil area* in (mm^2) at a frequency of 250 Hz. Also, it already classifies the gaze-points in *fixations* (*gaze coordinates in pixels and mean pupil area in* mm^2); *saccades* (*start and stop coordinates in pixels, amplitude in pixels, peak velocity in pixels/milliseconds, duration in milliseconds*); and *blinks* (*duration in milliseconds*). Furthermore, the textual adventure game annotates the above events to the Eyelink 1000 interface to segment the pupillometric data and synchronize them with the other data source. The Shimmer3 GSR+ sensor read participants raw *electrodermal activity* (EDA) in micro siemens (µS) and *Photoplethysmogram* (PPG) at a frequency of 50 Hz. Finally, I stored players' *mouse cursor coordinates* in pixels, at a frequency of 20 Hz, along with mouse click events.

6.4.6 Data Preparation

The data preparation and the following analysis are focused only on the *Trial passages*; indeed, I was only interested in exploring participants' behavior and physiological reaction in response to those proposals. I separately pre-processed data from the Eyelink, Shimmer, and mouse devices and aggregated them based on the in-game events, following the steps below.

Table 6.1 Features selected from the different data sources; gray features were discarded during the features selection, see section 6.4.6 for the selection criteria. The number in the last column refers to the kept features

Data Source	**Selected Features**	**#**
Behavioral	Interval duration	1
Pupillometry & Gaze	Fixations number, fixation frequency, mean fixations duration; max, min, mean, standard deviation, kurtosis, skewness of the pupil area; mean and max of the first and mean of the second derivative of the pupil area. Blinks number, max, min, mean and standard deviation of the blink duration. Mean amplitude , peak velocity, and duration of the saccades	15
Skin Conductance Response (SCR)	Number of SCR peaks in the interval, number of SCR peaks in the 4 seconds after the continue (for *read intervals*) or the decision (for *decide intervals*)	3
Galvanic Skin Level (GSL)	Max, min, mean, standard deviation, skewness, kurtosis of the GSL; mean and max of the first derivative and mean of the second derivative of the GSL	6
Photoplethysmogram (PPG)	Max, min, Mean ,standard deviation of the heartrate. MeanNN, SDANN1, RMSSD ; low, very low, high and low/high frequencies	3
Mouse Strokes	Maximum deviation, Area under the curve, length, displacement, length / displacement, mean speed, mean velocity, number of x-flips , entropy and bimodality over x and y axis	5
Total		33

Data ingestion and Pre-processing PPG, EDA, and mouse data are stored in CSV format, with temporal series separated from the annotations. Pupillometry & Gaze data include both time series and annotations: the Eyelink 1000 generates a tabular-like file in ASC format. I processed the 4 data sources separately.

I used the Neurokit2 python library [332] to process the collected EDA data: the tool returns a cleaned time series along with the tonic (Galvanic Skin Level, GSL) and phasic (Skin Conductance Response, SCR) components, and the detected SCR peaks. Then, I applied a within-participant min-max normalization to both tonic and phasic components. I used the same library to process the PPG data, obtaining the cleaned heartbeat, the estimated heart rate, and the detected peaks.

Regarding mouse trajectories, I used the Squeak python library [333]. I normalized the trajectories by space and time, obtaining 101 steps from 0 to 1 for each trajectory. Then, I flipped the X-axis of decreasing trajectories to have all of them in the same direction.

Finally, regarding the pupillometry & gaze data, I firstly filtered all the fixations outside the Eyelink calibration area. This is extremely important for *SE-icub* passages since the robot was placed outside the Eyelink calibration area (on participants' right, see Sec. III): when iCub moved and spoke interacting with the participants, they inevitably gazed toward it. Then, I cleaned the mean pupil area time series among fixations by applying a median filter to remove the outliers and a rolling window to smooth the time series. Finally, I applied a baseline correction subtracting the mean pupil area during the 5 seconds before each room from the pupil time series of the relative passages [168].

Segmentation and Features Extraction I segmented the cleaned time series based on participants' behavior. I identified two intervals for each passage: ***read intervals***, from the beginning of a passage to the *"continue"* button click; and ***decide intervals***, form the *"continue"* button click to the selection of one of the two possible options.

Within each interval, I computed the 56 features listed in Table 6.1. Then, I performed a feature selection to reduce the information redundancy. I computed the Pearson's correlation matrix of the average value of each feature and subject. Finally, I selected the less elaborated feature for each pair with a correlation score r>=0.8 and removed the others (the gray features in Table 6.1).

A note on mouse stroke features Mouse stroke features analyze the trajectory between a starting point (i.e., the "continue" button) and an ending point (i.e., one of the option buttons); and how it is different with respect to a straight line between the two points. The *maximum deviation* describes how much the trajectory diverges from the virtual straight line (i.e., a high deviation could be related to indecision [163]); the *number of x-flips* counts how many times the users "crossed" the straight line (i.e., a high number of x-flips could suggest the user is more decided and should be paired with low maximum deviation [164]); finally, the *bimodality* explores whether the trajectory comprises multiple modes (i.e., firstly going straight than diverging, usually related to indecision [334]).

The final dataset is composed of 1458 data points and 33 features. It is balanced between *read* and *decide intervals*. However, it is slightly unbalanced on the conditions (314: *no-risk*, 322: *risk*, 484: *SE*, 338: *SE-icub*) and participants' decisions (804: *accept*, 654: *refuse*).

6.5 Results

This section reports the in-game and questionnaire results and the post-hoc analysis of players' behavior and physiological reactions. Firstly, I focused on players' behavior and how it was influenced by their background and different proposals conditions (*H1*). Note that participants took decisions under four conditions: *no-risk* proposals did not involve any risk; *risk* proposals involved risk based on the DOSPERT questionnaire [320]; *SE* proposals involved both risk and a virtual agent performing a social engineering attack (e.g., persuading participant to accept the risk); and *SE-icub* proposals involved both risk and a social engineering attack performed by the iCub, physically present with participants during the experiment.

Then, I explored participants' physiological reactions and how they were influenced by the different conditions (*H2*). For this purpose, first I focused on the effect of *condition*, by comparing *no-risk* vs. *risk* vs. *SE* proposals. Then I analysed the impact of the attacking agent nature, with a comparison between *SE* and *SE-icub* proposals (*H3*).

6.5.1 In-game Results

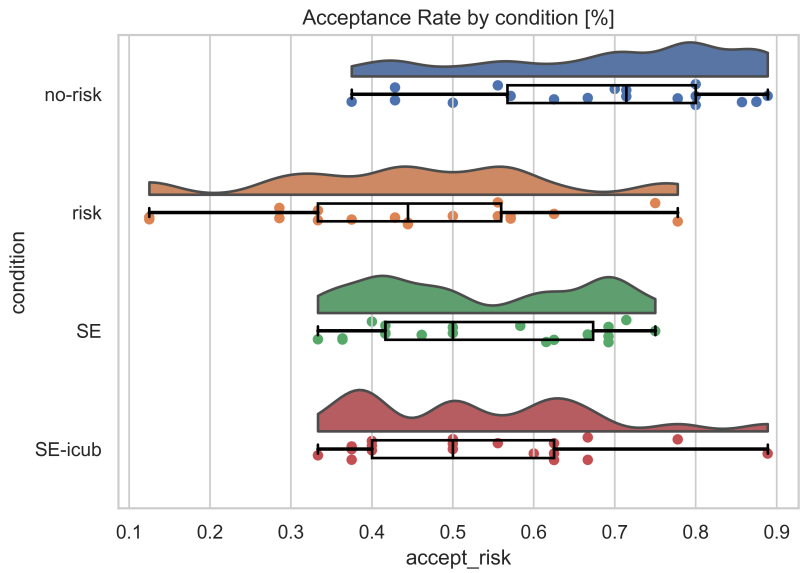

Figure 6.5 Average probability to accept proposals in the four conditions. Dots represent the average of each participant.

The 20 players took, on average 56 (SD=11) minutes to play the textual adventure. Only 2 ran out of Quantum Energy (QE) before the game ended; the remaining 18 completed the

game with 59 (SD=28) QE points, depending on their in-game decisions. Among them, 12 decided to take *The Seed*, while the other 6 left it in the *crypt*.

I analyzed how many participants accepted the proposals among the four conditions (i.e., their acceptance rate). Figure 6.5 shows the average percentage of accepted proposals for each condition. Participants complied the highest with *no-risk* (M=68.7%, St.Err.=3.62%) and the lowest with *risk* (M=45.1%, St.Err.=3.92%) proposals, while *SE* (M=53.3%, St.Err.=3.08%) and *SE-icub* (M=53.4%, St.Err.=3.34%) acceptance rates lie in the middle. This is consistent with my expectations (*H1*): the persuasion involved in Social Engineering attacks, either from a virtual agent or from iCub, affected players' risk appraisal (i.e., mitigating or enhancing it) and facilitated their acceptance. A Shapiro-Wilk normality test showed players' decisions were normally distributed; hence, I validated the difference between acceptance rates with a repeated-measures ANOVA. The test showed a significant effect of the condition F(19, 3)=9.15, p<0.001. Post-hoc tests, Bonferroni corrected, showed an higher acceptance rate of *no-risk* proposals than *risk* (B=0.235, t=5.1, p<0.001), *SE* (B=0.153, t=3.34, p=0.02) and *SE-icub* (B=0.152, t=3.12, p=0.034) ones; also, it showed no difference between *risk* and both *SE* (B=-0.082, t=-2.05, p=0.328) and *SE-icub* (B=-0.083, t=-1.646, p=0.628) proposal; and neither between the latter two (B=-0.001, t=-0.027, p=1.0).

Reading and Decision Time

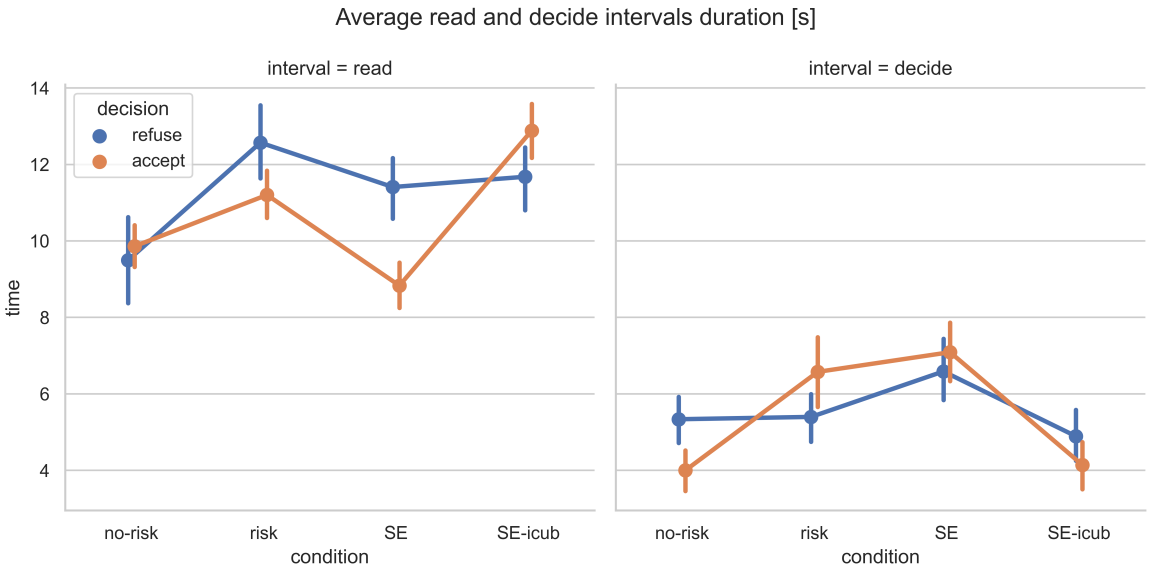

Figure 6.6 *Reading* (left) and *Deciding* (right) times in seconds for each condition and based on participants' decision to accept (orange) o refuse (blue). Bars represent the standard deviation.

Players took, on average, 10.5 (SD=7.5) seconds to read the trial passages and 5.6 (SD=6.5) seconds to decide. A Shapiro-Wilk normality test showed how all the computed features were not normally distributed. Hence, I performed a non-parametric analysis. I fit two mixed effect models in the *read* and *decide* intervals, respectively, with the read time and decide time as the dependent variable. In both models, I entered fixed effects of "condition" (four levels: *no-risk, risk, SE, SE-icub*), "decision" (two levels: *accept, refuse*) and their interaction. I selected *no-risk* and *refuse* as reference levels. Also, I included a random intercept based on participants' IDs.

On *read intervals*, I found a significant effect of the condition (F=3.35, p=0.019), no effect of the risk-taking (F=1.17, p=0.279), but an almost significant effect of their interaction (F=2.23, p=0.084). As it is possible to see in Figure 6.6 (left), there is a tendency for participants to took more time to read *risk* and *SE* passages when they refused.

Regarding the *decide intervals* (see Figure 6.6, right), the mixed effect model showed a highly significant effect of the condition (F=6.994, p<0.001) but no effect of the decision (F=0.01, p=0.916) nor their interaction (F=0.82, p=0.484). In particular, I found a significant effect in *risk* (B=1.63, t=2.16, p=0.031), a highly significant effect in *SE* (B=2.56, t=3.67, p<0.001) and no effect in *SE-icub* (B=0.08, t=0.11, p=0.91) with respect to *no-risk* proposals.

Pre-questionnaire analysis

Average scores for the Ten-Items Personality Inventory [318] were Conscientiousness: M=0.80, SD=0.15; Agreeableness: M=0.72, SD=0.18; Emotional Stability: M=0.59, SD=0.16; Openness to experiences: M=0.69, SD=0.14; Extraversion: M=0.47, SD=0.23. The DOSPERT scale [320] measured participants' probability to engage in domain-specific activities, the average scores were Ethical: M=0.22, SD=0.11; Financial: M=0.29, SD=0.12; Recreational: M=0.52, SD=0.19; Health & safety: M=0.41, SD=0.17; Social: M=0.72, SD=0.08. Average scores for the UPPS-P questionnaire [328, 329], regarding impulsiveness in the decision-making, were Emotional-based Rush actions: M=0.71, SD=0.18; Sensation Seeking: M=0.43, SD=0.17; Conscientiousness Deficit: M=0.78, SD=0.11. Regarding participants' proneness to Social Engineering effect mechanisms [92], the average scores were Normative Commitment: M=0.83, SD=0.09; Continuance Commitment: M=0.73, SD=0.12; Affective Commitment: M=0.78, SD=0.16; Trust others: M=0.58, SD=0.13; Obedience: M=0.68, SD=0.10; Reactance: M=0.67, SD=0.13; Subjective behavior (i.e., updating credentials online) M=0.54, SD=0.19). Finally, the RoSAS questionnaire [330] measured participants' perception of robots in general, average scores were (ability to have) Feelings M=0.29, SD=0.19; Competence M=0.83, SD=0.13; Danger M=0.31, SD=0.21.

Afterward, I explored whether participants' psychological background affected their in-game behavior. I fit a set of linear regression models with the survey's psychological features as independent variables (considering each questionnaire separately) and the average acceptance rate, reading time, and decision time as dependent variables. The acceptance rate was negatively correlated with the extroversion (t(20)=-2.29, p=0.038, Adj. R2=0.33). The reading time instead was positively correlated with the conscientiousness (t(20)=2.82, p=0.014, Adj. R2=0.41) and the proneness to authority-based social engineering attack (t(20)=2.34, p=0.042, Adj. R2=0.57), and negatively correlated with the proneness to reactance-based social engineering attacks (t(20)=-2.32, p=0.043, Adj. R2=0.57); finally, the decision time was positively correlated with the risk-taking habit in social circumstances (t(20)=2.48, p=0.027, Adj. R2=0.14).

Post-questionnaire analysis

This section summarizes the results of the Player eXperience Inventory (PXI) [331] and the custom questions participants filled in just after playing the experiment. The PXI questionnaire measured participants' experience and perception of the game: how much playing the game meant to them (M=0.78, SD=0.17); how much they were curious about it (M=0.9, SD=0.08); how much they mastered it (M=0.70, SD=0.21); how much they were free to decide how to play the game (M=0.75, SD=0.21); how much they were immersed (M=0.83, SD=0.12); how much the game gave them feedback about their progression (M=0.77, SD=0.15); how much their in-game goal was clear (M=0.82, SD=0.15); how much the game was challenging (M=0.75, SD=0.14); how much the game was aesthetically appealing (M=0.87, SD=0.13).

Moreover, I assessed which factor motivated participants on making their decisions, average values were Immersion: M=0.79, SD=0.11; Curiosity: 0.64, M=0.23; Greed: M=0.61, SD=0.22; Rush: M=0.31, SD=0.26; Fear M=0.5, SD=0.24. Also, on average, participants tried to maximize their QE (M=0.78, SD=0.21); they perceived their decisions to have an impact on the story (M=0.82, SD=0.14); and poorly regretted making decisions (M=0.34, SD=0.22). Also, even if I did not explicitly assess whether they perceived the game as fun, all the participants except one wanted to play more. Finally, regarding iCub, they perceived it as a companion (M=0.87, SD=0.12); they averagely took iCub into account while making decisions (M=0.53, SD=0.23); and they trusted the robot more after the game (M=0.75, SD=0.18).

Table 6.2 Physiological features showing significant effects in the *read* (R) and/or *decide* (D) intervals, among the for conducted statistical tests. Omitted features did not show any effect. Please refer to the text for more details. Condition = *no-risk, risk, SE*, Decision = *refuse, accept*, Agent = *SE, SE-icub*

	Feature	Condition	Decision	Condition x Decision	Agent x Decision
Pupillometry & Gaze	Fixations number	RD			D
	Fixation frequency			R	R
	PA min	R			
	PA mean	RD		D	RD
	PA std	D	RD	RD	
	PA dt1 mean	R			D
	Blinks number	RD			R
	Saccades amplitude mean				D
SCR	Peaks number	D			RD
GSL	Std	D			
	Skewness	D			
	Dt1 mean				D
	Dt1 max	D			
	Dt2 mean		D		D
Mouse Strokes	Maximum deviation				D
	x-flips				D
	x-axis bimodality	D			

6.5.2 Risk Appraisal & physiological reactions

In this section, I focus on players' behavioral and physiological reactions to *risk* and *SE* proposals with respect to *no-risk* ones. As a general remark, I did not find any effect on the PPG features. I speculate this is due to the short duration of participants' interaction with the passages: heart-related features need to be observed for longer periods (i.e., in the order of minutes) [157].

Regarding the two intervals, I expect different physiological reactions: in *read intervals*, participants read a textual stimulus of different length and processed information; while in the *decide intervals*, they were only asked to read the two options and to click the preferred one. The former could involve more variability due to processing written information. Hence, I started the analysis from the *decide interval*, assuming participants' physiological reactions should be clearer.

A Shapiro-Wilk normality test showed that the features were not normally distributed, neither in *read* nor in *decide intervals*. Hence, I fitted a set of mixed effect models with each physiological feature as the dependent variable; a fixed effect of "condition" (three levels: *no-risk, risk, SE*), with reference on *no-risk*; and a random effect of participants' ID (*Condition* column in Table 6.2)

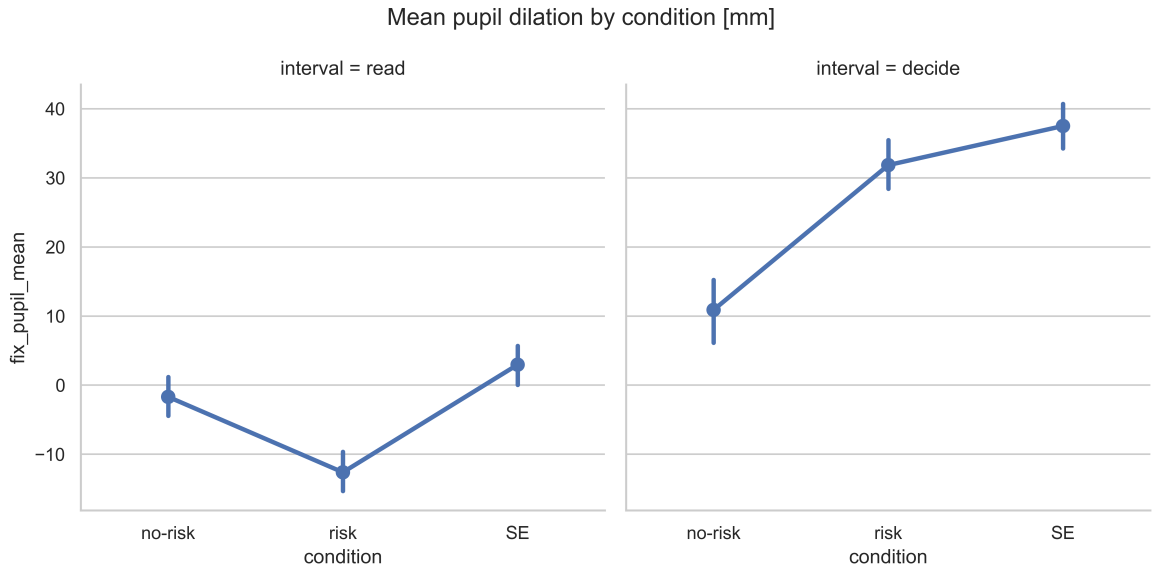

Figure 6.7 Mean pupil dilations in millimeters during read (left) and decide (right) intervals between *no-risk*, *risk*, and *SE* conditions. Bars represent the standard deviation.

Pupillometry & Gaze During the *decide intervals* I found significant effects due to the condition on the mean and standard deviation of the pupil area, the number of fixations and the number of blinks. They all follow a similar increasing pattern with a higher value in *risk* and *SE* proposals with respect to *no-risk* ones (see Figure 6.7, right). For instance, the mixed model fitted on the mean pupil area showed a highly significant effect both in *risk* (B=16.6, t=3.59, p<0.001) and *SE* (B=22.6, t=5.36, t<0.001) proposals – similar results for the standard deviation. This could be related to the cognitive load needed to decide under *risk* or *SE*. Regarding the number of blinks, I found a significant effect only in *SE* (B=1.034, t=2.72, p=0.007) but not in *risk* (B=0.28, t=0.68, p=0.494) proposals. For the number of fixations, I found a significant effect both in *risk* (B=6.30, t=2.60, p=0.001) and *SE* (B=9.55, t=4.32, p<0.001) proposals. However, those two features could be affected by the decision time since they are absolute counting. Indeed, I did not find any significant effect on fixation frequency.

In the *read intervals* I found significant effects on the mean pupil area, the number of fixations and blinks, but also on the minimum pupil area and the mean of the pupil area first derivative. Interestingly, the pattern is different than the previously observed (see Figure 6.7, left). Participants showed a lower mean pupil area in *risk* (B=-8.60, t=-2.01, p=0.045) passages while there is no difference in *SE* (B=4.45, t=1.14, p=0.256) with respect to *no-risk* ones. A post-hoc test, Bonferroni corrected, showed a significant difference in *SE* with respect to *risk* passages (B=13.06, t=-3.36, p=0.003). Finally, participants' mean first derivative of the pupil area changed faster in *risk* (B=76.7, t=4.60, p<0.001) and *SE* (B=58.7, t=3.85, p<0.001) passages with respect to *no-risk* ones.

Skin Conductance Response (SCR) and Galvanic Skin Level (GSL) The Electrodermal Activity (EDA) models in *decide intervals* showed significant effects on the maximum first derivative, the standard deviation, and the skewness of the GSL – the background, long-term, component – and on the number of the SRC peaks – the short-term, event related component. Participants' GSL changed faster in *SE* (B=0.0013, t=2.27, p=0.024) but not in *risk* (B=0.0001, t=0.03, p=0.97) proposals with respect to *no-risk* ones. The standard deviation and skewness of GSL followed a pattern comparable to participants' pupil area: I found a significant effect on the GSL standard deviation both in *risk* (B=0.0028, t=2.18, p=0.03) and *SE* proposals (B=0.0026, t=2.19, p=0.029) – similar effect on the skewness. Lastly, participants experienced a significant higher number of SCR peaks in *SE* (B=0.21, t=2.28, p=0.023) proposals, but no difference in *risk* (B=-0.05, t=-0.52, p=0.60) ones.

I did not find any significant effect on the EDA components during the *read intervals*.

Mouse Strokes Finally, I found a higher bimodality (i.e., the presence of different movement styles, like straight or curved, in the mouse trajectories) over the x axis in *SE* (B=0.03, t=2.18, p=0.03) proposals with respect to *no-risk*, but no difference in *risk* (B=0.02, t=1.45, p=0.15) ones, during the *decide intervals*.

6.5.3 The effect of risk-taking

After assessing how participants reacted to the different game conditions, I explored whether they showed different physiological reactions due to the decisions to comply or not with the trials. I fitted a set of mixed-effects models with the physiological features as the outcome variables. I entered a fixed effect *"decision"* (two levels: *accept, refuse*) with reference level on *refuse*; and I included a random effect of participants' ID (*Decision* column in **Table 6.2**).

In the *decide intervals*, I found significant effects on the standard deviation of the pupil area, higher when refusing (B=-3.04, t=-2.64, p=0.008), an almost significant effect on the frequency of fixations, higher when refusing (B=-0.0002, t=-1.78, p=0.075) and an opposite tendency, on the mean second derivative of the GSL (B=0.0002, t=1.91, p=0.057). In *read intervals*, instead, I found only an effect on the standard deviation of the pupil area (B=-2.65, p=0.008). I did not find any effect on the EDA features or the mouse strokes.

From these results, participants seemed to have a stronger physiological reaction when they decided not to comply with the risky proposals. Indeed, this is consistent with the literature: humans tend to comply with it rather than reason through when facing a decision under uncertainty and lack of information. Humans are relatively trustful by nature [335] and researchers assume most individuals start with an "assumption of truth" when making decisions [336, 337]. Hence, refusing a proposal (i.e., going against a natural behavior) could be more stressful than accepting it.

Given the wide variety of decisions, I speculated that participants could experience different internal states related to complying (or not) with decisions under different conditions. To verify this hypothesis, I fitted a set of more complex mixed effect models with the physiological features as dependent variables; two fixed effects, one of "conditions" (three levels: *no-risk*, *risk*, *SE*; reference on *no-risk*) and one of "decision" (two levels: *accept*, *refuse*; reference on *refuse*). As before, I included a random effect of participants' IDs (*Condition x Decision* column of Table 6.2).

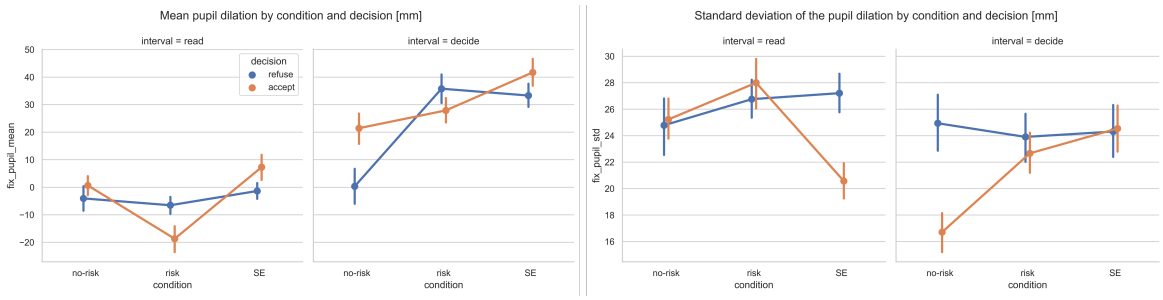

Figure 6.8 Mean (left) and standard deviation (right) of the pupil area during *read* and *decide intervals*, compared between conditions and participant' decisions. Bars represent the standard deviation.

As it is possible to observe in Figure 6.8, participants experienced different reactions to the mean pupil area due to accepting or refusing among the different conditions in the *decide intervals*. In particular, the *risk* condition exhibits an opposite trend with respect

to the other two. For the standard deviation of pupil area instead, the difference is only in *no-risk* passages; hence it does not characterize participants' decision-making under risk.

During the *decide intervals*, the mixed model on the mean pupil area showed a significant effect only by condition but not by the decision nor on their interaction. The model on the standard deviation of the pupil area showed a significant effect both by condition – in *risk* (B=3.37, t=2.19, p=0.29) and in *SE* (B=3.94, t=2.79, p=0.005) – decisions (B=-3.04, t=-2.56, p=0.011) and their interaction – both in *risk* (B=6.28, t=2.02, p=0.044) and *SE* (B=7.79, t=2.73, p=0.007).

In the *read intervals*, I did not find any significant effect on the mean pupil area; however, I found a significantly lower standard deviation of the pupil area while accepting *SE* proposals (B=-7.17, t=-2.68, p=0.008). Also, the frequency of fixations was interesting: in *no-risk*, participants' fixated slightly more when refusing than when accepting, even if the difference is not significant; however, the pattern is opposite in *SE* with a significantly higher number of fixations while accepting than refusing (B=0.0005, t=3.306, p=0.001); finally, there was no difference in *risk*.

Finally, I did not find any statistical effect on the EDA or mouse stroke features combining conditions and decisions.

6.5.4 Social Engineering: virtual agent vs. iCub

Until now, I put aside *SE-icub* passages and focused on the other three conditions. In this section, I compare *SE-icub* and *SE* passages to understand the effect of social engineering attacks from different types of attackers (i.e., a virtual agent or a physically present humanoid robot). *SE* and *SE-icub* conditions involved different experiences for the participants: in the latter, the humanoid robot interacted with the participants by moving its arms, torso, and gaze; it spoke – with the text also appearing on the game GUI – and it showed emotional expressions. Hence, this kind of social interaction differs from just reading on-screen text (*SE* passages). Moreover, as observed in the previous section, participants' behavior and physiological reactions are influenced by both the conditions and their decisions to comply or not.

I fitted a set of mixed effect models with the selected physiological features as the dependent variables. I entered two fixed effects: "condition" (two levels: *SE, SE-icub*; reference on *SE*) and "decision" (two levels: *accept, refuse*; reference on *refuse*); and a random effect based of participants' IDs (*Agent x Decision* columns of Table 6.2).

Pupillometry & Gaze In the *decide intervals*, the interaction between condition and decision showed a significant effect on the mean pupil area (B=-26.52, t=-3.12, p=0.002): it was higher when refusing to *SE-icub* (B=17.64, t=2.70, p=0.043) but not to *SE* proposals (B=-8.87, t=-1.62, p=0.63) as proved by a post-hoc test Bonferroni corrected. Also, I found a higher first derivative of the mean pupil area in *SE-icub* with respect to *SE* passages (B=75.59, t=2.66, p=0.008) with no difference due to the decision or their interaction. Finally, I found a lower mean amplitude of the saccades (B=-9.58, t=-4.30, p<0.001) and a lower number of fixations (B=-9.44, t=-4.24, p<0.001) in *SE-icub* passages.

In the *read intervals*, participants' mean pupil area showed a similar pattern: I found a significant effect of condition (B=8.01, t=1.99, p=0.047) and of the interaction of the two factors (B=-16.646, t=-2.07, p=0.04). Participants' pupil area instead was almost identical between *SE* and *SE-icub* proposals when accepting (B=0.31, t=0.06, p=1.0) but it was higher when refusing to *SE-icub* proposals (B=-16.33, t=-2.76, p=0.036). Also, I found a higher frequency of fixations on *SE* (B=-0.0003, t=-4.64, p<0.001) and a significant effect on the factors' interaction (B=-0.0004, t=-3.22, p=0.001) with a higher frequency when accepting *SE* proposals but a lower frequency when accepting *SE-icub* ones. Finally, I found a higher number of blinks in *SE-icub* passages (B=1.85, t=4.87, p<0.001), with no effect on the decision nor the interaction of the two factors. However, the two latter effects are most likely caused by the iCub position outside the screen area: gazing toward the robot could have been interpreted as blinks (higher in *SE-icub*), also gazing away from the screen area would produce lower fixations.

Skin Conductance Response (SCR) and Galvanic Skin Level (GSL) In the *decide intervals*, I found significant effects of the interaction of the two factors (B=0.002, t=2.17, p=0.03) on the mean first derivative of the GSL. It was almost identical between *SE* and *SE-icub* conditions when refusing (B=0.001, t=0.17, p=1.0), but it was higher on *SE-icub* when accepting (B=-0.002, t=-3.03, p=0.026) as confirmed by a Bonferroni-corrected post-hoc test. Regarding the second derivative of the GSL, it was generally higher when accepting to comply with a social engineering attack rather than refusing (B=0.003, t=2.13, p=0.03) with no effect due to the condition.

Finally, I found a higher number of SCR peaks when *deciding* in the *SE-icub* condition (B=0.15, t=2.02, p=0.04), but a higher number of SCR peaks while *reading SE* passages (B=-0.24, t=-2.43, p=0.016).

Mouse Strokes In *decide intervals* I found a higher maximum deviation (B=0.15, t=2.12, p=0.034) and a lower number of x-flips (B=-0.71, t=-3.44, p=<0.001) in *SE-icub* proposals, with no effect due to participants' decisions.

Table 6.3 Machine Learning model results, for each task and by different features sets, ordered by descending F1 score. Most performant models in bold. (PG=Pupillometry & Gaze). The accuracy score is not employed due to the unbalancing of Task 1 and 2.

Task	Feature Set	F1	Precision	Recall	ROC AUC
1. *Risk*/*SE*/*SE-icub* vs. *no-risk*	**PG**	**0.867**	**0.848**	**0.889**	**0.649**
	PG + EDA + PPG	0.844	0.805	0.889	0.547
	PG + EDA	0.844	0.827	0.861	0.597
	PG + Behave	0.794	0.844	0.750	0.619
	Mouse	0.760	0.785	0.736	0.496
2. *SE*/*SE-icub* vs *no-risk*	**PG**	**0.788**	**0.781**	**0.796**	**0.603**
	PG + EDA + PPG	0.781	0.766	0.796	0.578
	PG + Behave	0.740	0.733	0.748	0.515
	Mouse	0.711	0.745	0.679	0.532
2. *SE*/*SE-icub* vs *risk*	**PG + EDA**	**0.809**	**0.794**	**0.825**	**0.638**
	PG + EDA + PPG	0.803	0.792	0.816	0.633
	PG + Behave	0.798	0.810	0.786	0.656
	PG	0.756	0.776	0.738	0.594
	Mouse	0.697	0.714	0.679	0.490
3. *Accept* vs *refuse* (in *SE*/*SE-icub*)	**EDA**	**0.718**	**0.586**	**0.927**	**0.589**
	Mouse	0.671	0.553	0.855	0.531
	PG + EDA + PPG	0.667	0.563	0.818	0.545
	Behave	0.651	0.568	0.764	0.548
	PG	0.651	0.568	0.764	0.548

6.6 Predicting Users' Decisions

The statistical analysis revealed how the different conditions and participants' decisions affect their physiological reactions in a not-straightforward way. I speculate that machine learning models could be more effective in discriminating the different cases by integrating the contribution of multiple features simultaneously. I focused on three tasks:

- ***Task 1***: detecting the occurrence of risky situations
- ***Task 2***: detecting the occurrence of a social engineering attack;
- ***Task 3***: predicting the compliance with a risky situation.

Previously, I separately considered *read* and *decide* intervals to analyze participants' behavior and reactions better while performing different activities. However, there is no distinction between the two phases in a real-world scenario (i.e., while reading a phishing email). Moreover, the decision-formation instant is not easily identifiable, as also proved by the statistical analysis.

Thinking about a future application, I addressed the tasks mentioned above on a new dataset where *trial passages* are not separated in *read* and *decide* intervals. I obtained it with the same process outlined in Section 6.4.6 and with the same 33 selected features listed in Table 6.1. The dataset comprises 729 data points divided into 157 for *no-risk*, 161 for *risk*, 242 for *SE*, and 169 for *SE-icub* conditions; also, they include 402 for accept and 327 for refuse decisions. Given the small size of the dataset, I opted for training Random Forest classifiers [292] for their ability to not overfit even on limited datasets. Regarding the features, the statistical analysis suggests pupillometry, gaze, and Galvanic Skin Level (GSL) features should be the most useful to solve the tasks. However, in the statistical analysis, I did not explore the combined contribution of multiple features, neither from the same data source nor mixing them. Hence, I trained random forest classifiers for each combination of the features, grouped by data source (i.e., with pupillometry and gaze features only, with pupillometry, gaze and EDA features, and so on . . .). Finally, given the partial imbalance of the dataset, I relied on the F1, precision, recall and ROC AUC scores, rather than on the accuracy [295] (as in chapter 5).

Task 1: Detect risky situations I binarized the problem, considering as "*risky*" the passages of *risk*, *SE* and *SE-icub* conditions and the other as "*not-risky*". I split the dataset considering 25% as the test set and the rest as the training set. The dataset is unbalanced (157 not-risky data points and 572 risky data points), so I embedded the synthetic minority oversampling technique (SMOTE) in the training pipeline [293]. The algorithm generates synthetic data points for the not-risky class over the training set while leaving the testing set untouched. I trained a random forest classifier for each combination of the input features. I validated each model with a 4-fold grid search cross-validation, looking for the best hyperparameters set. Pupillometry and gaze features were the most effective, with an 86.8% of F1 score, 84.8% of precision, 88.9% of recall, and 0.64 of ROC AUC. See the first rows block of Table 6.3 for more details.

Task 2: Detect social engineering attacks Then I explored whether it is possible to discriminate proposals involving a social engineering attack (e.g., involving influence and

manipulation) from *no-risk* proposals. I followed the same above-mentioned procedure, labeling as *social-eng* the *SE* and *SE-icub* passages (N=411) and as *not-social-eng* the *no-risk* passages (N=157). As before, I considered 25% of the dataset as a test set, over-sampled the remaining training set with the SMOTE algorithm, and performed a 4-fold grid search cross-validation. Pupillometry & gaze were again the most effective feature sets with 78.8% of F1 score, 78.1% of precision, 79.6% of recall, and 0.60 of ROC AUC. See the second rows block of Table 6.3 for more details.

Furthermore, following the same procedure, I tried to discriminate proposals involving social-engineering attacks from *risk* proposals. The combination of EDA and pupillometry & gaze features achieved the best performance with 80.9% of F1 score, 79.4% of precision, 82.5 of recall, and 0.64 of ROC AUC. Also, both the pupillometry & gaze features (F1 score of 75.6%, ROC AUC of 0.59) and EDA features (F1 score of 72.3%, ROC AUC of 0.53) alone achieved a good performance. See the third rows block of Table 6.3 for more details.

Task 3: Predict participants' compliance Finally, I explored whether it is possible to predict victims' compliance with a social engineering attack performed by a virtual agent or the humanoid robot iCub. From the dataset, I selected only *SE* and *SE-icub* passages; I separated the training and testing set (25%); but I did not apply any oversample since the dataset is almost balanced; finally, I performed a 4-fold grid search cross-validation, looking for the best hyperparameters. Surprisingly, the EDA features alone were the most effective, with 71.8% of F1 score, 58.6% of precision, 92.7% of recall, and 0.59 of ROC AUC. Given the high recall and low precision, the model tends to be more conservative, giving many false positives, which could be acceptable in the cyber security context. See the last rows block of Table 6.3 for more details. Please notice the trained models were not informed of the attack condition (i.e., *SE* or *SE-icub*).

Comprehensively, Pupillometry & Gaze features are confirmed to be the most informative. However, satisfying results could be achieved even with more accessible measures (i.e., mouse strokes features).

6.7 Discussion

In this study, I explored participants' behavioral and physiological reactions when facing risky and social engineering proposals. My primary purpose was to identify specific reactions helpful in predicting users' compliance with social engineering requests or detecting the

occurrence of attacks. I firstly analyzed participants' decisions and then their physiological reactions. As hypothesized in *H1*, participants complied the most with proposals involving *no-risk* (68.7%) and the less with *risk* (45.1%) ones; also, iCub (53.4%) and the virtual agents (53.3%) partially tricked players, making them comply a little more than risky proposals, but not at the level of *no-risk* ones.

In general, participants took more time reading the textual stimuli than deciding, suggesting that, when they clicked on the *"continue"* button, their decision was almost formed - even if the game never instructed participants to click on the *"continue"* button when they were ready to decide. Interestingly, players took more time to read *risk* and *SE* passages when they were going to not-comply, even if the difference is significant only in the latter. This suggests players took more time to appraise the situation before refusing carefully. Interestingly, *SE-icub* passages showed a completely different pattern, with no difference as a function of participants' final decision. ICub led the interaction during *SE-icub* passages, so players had minimum control over their reading time. To decide, instead, players took a little longer in *risk* and *SE* passages, but significantly lower in *SE-icub*, with a value comparable to *no-risk* proposals. The different patterns in *SE-icub* are most likely caused by the social presence of the robot, pushing participants to act faster.

Given the significant difference between the acceptance rate of *no-risk* and social engineering proposals, I speculate participants were still perceiving some risk, or at least the oddness of the situation; however, it seems that the attackers influenced their reasoning processes and choices. To explore *how* they have been affected, I studied players' reactions on pupillometry, gaze, electrodermal activity (EDA), photoplethysmography (PPG), and mouse strokes.

All the employed physiological metrics are known from the literature to be affected by risk and uncertainty in decision-making. However, it was not clear how social engineering threats would have impacted them. Given the complexity of the domain, I decomposed the analysis, separately exploring three effects: *(i)* the different types of proposals (*no-risk, risk or SE*); *(ii)* participants' intention to comply or not; and *(iii)* the type of attacker (iCub or a virtual agent) for *SE* and *SE-icub* proposals.

Among all the collected physiological features, pupillometry, gaze, and electrodermal activity were the most informative. They both proved to be affected by the different types of proposals, with a stronger reaction in *SE* than in *risk* and *no-risk* proposals. The need to carefully evaluate risky options, and even more with added uncertainty (i.e., in *SE* passages), increased players' cognitive load, dilating their pupils. Furthermore, pupillometry and gaze

showed to be highly influenced by participants' intention to comply or not. Regarding EDA, the temporal variation of the GSL and the number of SCR peaks were the most relevant, both higher in *SE* proposals. Finally, I found an effect on the mouse strokes x-axis bimodality (i.e., the presence of multimodal trajectory styles) in *SE* proposals suggesting the presence of a dual cognitive process [334].

Comparing different social engineering attackers, as expected, players' arousal reaction was more vigorous in *SE-icub* cases. In particular, the difference is observable in the mean pupil area, the first derivative of the GSL, and the number of SCR peaks. However, the differences are highly modulated by participants' willingness to comply with the attacker's proposal. Finally, results show an interesting effect on mouse strokes: participants' mouse trajectory was broader in *SE-icub* passages (i.e., less number of x-flips and higher maximum deviation), suggesting indecision (i.e., to decide whether to trust the robot or not).

Summing up, hypothesis *H2* and *H3* are confirmed: there are differences in social engineering proposals with respect to both *risk* and *no-risk*; and they differ also due to the effect of different attackers. However, participants' decisions affected their physiological reactions differently in different conditions, suggesting the measure strictly depends on the context.

To unfold this complex interaction, I exploited machine learning to achieve the main research objective of this project (i.e., to predict in advance the occurrence of social engineering attacks). The Random Forest classifiers I trained address three different tasks: (i) to discriminate between proposals involving *risk, SE* or *SE-icub* from *no-risk* ones (F1 score = 86.6%); (ii) to discriminate social engineering attacks from ordinary proposals (F1 score = 80.9%); and (iii) to predict victim's compliance to social engineering attacks (F1 score == 71.8%).

Pupillometry, gaze, and electrodermal activity features were the most informative. However, this is not surprising since they reflect inner cognitive load and the integration of arousal responses. Interestingly, EDA and mouse features, followed by the response time, pupillometry, and gaze, were the most effective in predicting victims' decisions during social engineering, even with a lower performance with respect to the other models (e.g., lower ROC AUC score). These results are obtained without distinguishing between *read* and *decide intervals*, as it would be necessary for real scenarios. Also, please notice that the presented models are based only on users' behavior and physiological reactions; they are not aware of the proposal's textual stimuli or other particular aspects, as state-of-the-art detection models would be.

Predicting users' decisions to prevent them from becoming victims is a difficult task, even for machine learning, with a performance not reliable enough to be employed in a real case (given the low precision and ROC AUC score). However, the other two models (i.e., to discriminate risky or social engineering-based proposals based on participants' physiological responses to the proposals) could be used to support state-of-the-art social engineering defense systems. After an initial calibration with the user, the models could detect variations in users' arousal during awareness-raising training: they could warn the users they are reacting too much to an unreal danger, or vice versa, they are not posing enough attention to a threat. This qualitative feedback could support users in developing a more informed awareness, comprehensive not only of the acquired knowledge but also of an inner-state insight.

To better support, an intelligent system should intervene (i.e., with a warning) before users' compliance (ideally, as soon as possible, but without being over-protective). Indeed, the current solution still needs to ingest the whole decision-making process (i.e., the entire passage). Note that, during my research, I trained several effective models based only on features from the *read* and *decide* intervals. Hence, detecting SE attacks (F1=76.1%, ROC AUC=0.52, based on Pupillometry, gaze and EDA features during the *read* interval) and predicting users' behavior (F1=65.1%, ROC AUC=0.56, based on EDA features during the *read* interval) in-time is possible. However, I preferred to not report such models since the discretization between intervals I forced in this experiment is not easily identifiable in a real scenario. Hence, I considered the presented models more relevant and generalizable, aiming to real-world applications. For this purpose, future research should identify the minimum reading interval sufficient to reliably classify users' Social Engineering appraisal and predict their decisions.

Also, the current solution needs to pre-process the collected data and aggregate features before feeding them to the machine learning models. This was necessary to study the effects of different scenarios on users' physiological reactions. However, future research should focus on the raw time series (or at least less pre-processed data), taking into account their temporal evolution instead of focusing on aggregated values. By processing in real-time users' behavior and physiological reactions, it could be possible to prevent compliance with social engineering attacks in time.

Furthermore, physiological reactions could depend on subjective features (i.e., resting heart rate or sweating) and personality traits. In the future, it will be necessary to take them into account, for instance, by running an initial user-specific training and analyzing the

variation with respect to this knowledge base.

In this study, I employed external sensors (i.e., the Eyelink eye tracker and the Shimmer device) that are difficult to port in a real scenario. This was necessary to collect high-quality data and assess the feasibility of my approach. In the future, it will be necessary to replace those devices with solutions more feasible for real applications. For instance, recent developments [110–112] suggest it will soon be possible to measure pupil dilation and fixations from standard RGB cameras like those embedded in smartphones and notebooks. Also, recently, the market offers several wearable devices (i.e., smartwatches) able to monitor users' movements, electrodermal activity, and heart rate in real-time. They are promising candidates for developing minimally invasive real-time monitoring, applicable to multiple real-life contexts.

This research is the first step toward a deeper understanding of how social engineering alters victims' inner states. Comparing users' reactions to social engineering and risk was my primary concern. Still, further research must be undertaken to gain a deeper insight into victims' risk appraisal process. As per the risk appraisal model, attackers use the SE effect mechanisms to alter our reasoning process; in this research, I only took into account the presence (or not) of an influence/manipulation without considering *how* the effect mechanisms' impact on users' inner state. For instance, we could say effect mechanisms are *"polarized"*: authority-based or reactance-based attacks tend to increase the appraised risk, making the victims perceive a situation as riskier than it is; on the other side, attacks based on social proof or liking tend to decrease the appraised risk, making a threat to be perceived as less dangerous. Exploring the physiological reaction in response to "polarized" attacks could be an interesting next step to get a deeper understanding of victims' inner states.

I designed the Social Engineering Adventure (SEA), an interactive storytelling serious game, to mitigate experimental biases: the SEA is an immersive context in which participants could decide and react as they would do in the proposed situations. Also, participants were unaware they would face social-engineering threats, as it would happen in real life; indeed, participants' behavior suggests my scene design was realistic. Still, my game design has room for improvement. In particular, it is currently affected by two issues: (i) narration flow and players' motivation; (ii) players' engagement management and lacking a characterizing game mechanic. During the game, participants had to find the seed in the crypt; I further reinforced participants' motivation to survive until the end of the game by making the final

compensation proportional to the collected quantum energy. Participants reported that this goal was clear (M=0.82, SD=0.15) and felt immersed in the game (M=0.83, SD=0.12). However, the proposals they faced were more related to my scientific objectives than the goal of finding the seed. Hence, the game would benefit from a better narrative design. Moreover, preserving players' engagement is crucial for serious games. All the participants except one reported they would have liked to play more; however, some could have felt the main game mechanic (i.e., reading on-screen text an average of 192 passages) to be boring, compromising their engagement. Hence, the game would benefit from a characterizing mechanic and fewer passages to read, keeping high participants' engagement. I addressed these issues in the Adventurer Companion Project presented in chapter 7.

The computational architecture I developed allows for unlimited expansions. For instance, it is easy to modify or expand the experimental stimuli on-the-fly, since the story itself is decoupled from its interpreter and the game engine. Similarly, it is easy to include new sensors or robots by developing consistent interfaces - as we will see in the next chapter - without changing the remainder of the architecture. Twine with the expanded Harlowe language allows non-programmer users to develop decision-based human-robot interactions, engaging participants via robot behavior and multimedia mediums. Also, the Experimental Utils and Robot Puppeteer modules can be effectively used for other experiments with similar procedural needs.

Future research should keep this study's immersion feature while grounding the context and threats in real scenarios. A key factor would be to put participants in the same condition they are every day, i.e., by not explicitly asking them to classify emails as phishing or safe. For instance, in an institution like the Istituto Italiano di Tecnologia, it would be possible to ask employees to work for a day on a specific sensorized workstation and deliver actual social engineering attacks (e.g., spear phishing).

Finally, this study confirmed how robots could be effective vectors for social engineering attacks [215, 14]. However, I speculate that their ability to be persuasive could be exploited to provide more effective warnings against social engineering. Indeed, recent research suggested that priming and warning are no longer effective in fighting social engineering attacks [36]. Hence, social robots - even more if physically present [338] - could provide effective warnings to elicit a behavioral change. Furthermore, robot companions could be employed in different domains, from the workplace to our homes; they could monitor their

human partners and intervene when a dangerous situation is detected. A key factor will be defining how robots should intervene to elicit an effective and safe behavioral change: this is the main objective of the next project.

Chapter 7

Social Robot Warnings against Social Engineering

7.1 Overview

The last chapter showed how it would be possible to detect the occurrence of social engineering (SE) threats using users' behavioral and physiological reactions. While a research stream should focus on improving this detection system, it is worth studying how to provide effective interventions (i.e., warnings to prevent victims' compliance). Recent studies - showing how priming and warnings are not effective anymore against social engineering [36] - call for novel intervention approaches to fight the habituation and saturation effect of formal warnings [234]. While a few attempts focused on haptic feedback [236], or subliminal messages [237], I speculate that social robots could be an effective solution. Indeed, robots' social presence and communicative skills could provide more understandable warnings, less perceived as disturbance [339].

During a visiting research period at the Social and Intelligent Robots Research Lab (SIRRL, University of Waterloo, Canada), I developed the Adventurer Robot Companion (ARC) game; an evolution of the Social Engineering Adventure (SEA) game (see Section 6). With the ARC project, I want to study *if* and *how* a social robot could make players change their minds concerning social-engineering-related proposals. For this purpose, I opted for the Furhat robotic platform [181] (see Section 2.4 for a description of Furhat). During the game, the Furhat took both the role of non-playing character and narrator, reading passages to the players; however, after each proposal with *risk* or *social engineering*, the robot intervened, pushing players to take the not-selected option. To do so, the Furhat employed an *assertive*

strategy for *risk*-based proposals; and either an *affective* or *rational* persuasion strategy, on two different groups of participants, for the *social-engineering*-based proposals. Finally, I explored the impact of the robot's physical presence by developing two versions of the experiment: one *in-person* and one *remote*, accessible via the Zoom platform.

The ARC project shows how the experimental setup and computational architecture developed for the SEA game can be easily reused and expanded. The data collection is currently ongoing; due to the COVID-19 limitations in Canada, we were able to recruit just a few participants from the *in-person* version of the experiment. I will present and discuss the preliminary results.

Figure 7.1 A participant playing the Adventurer Robot Companion game *in-person*

7.2 The Adventurer Robot Companion (ARC)

As previously discussed (see section 6.7), the Social Engineering Adventure (SEA) game presents two main issues, namely: (i) the potential lack of players' motivation to pursue the in-game goal; and (ii) the game length and lack of a characterizing game mechanic potentially mining players' engagement. I addressed both as follows.

Improved narration and interactive scenes In the SEA, players' in-game motivation (i.e., find the Seed in the Crypt) was clearly stated but partially disjointed by the in-game events. Hence, participants could have lacked motivation during the game, compromising their experience. The textual adventure narrative drives the interactive scenes in the ARC

game, leveraging participants' in-game motivation and objectives to present each proposal. In particular, the game narrative is divided into *"milestones"*, sub-objectives players must reach to achieve their final in-game goal. Furthermore, I designed the whole ARC game narrative, taking inspiration from the traditional Social Engineering attack framework from Mouton et *al.* [8].

New characterizing game mechanics Keeping participants engaged is extremely important to provide adequate training, and assessment [340]. From a game design point of view, the SEA game lacked a characterizing game mechanic: players only had to read the textual passages and make decisions. Fights were the only exceptions; however, they did not trigger players' agency since they were decided by rolling virtual dice. In the ARC game, fights are inspired by arm wrestling: players compete with enemies to fill a yellow bar before a countdown runs out. Each time players click on their weapon, the bar is filled toward the enemy with an amount proportional to the weapon's power; however, each half-second, the bar is decreased by an amount proportional to the enemy's defense. If the bar is on the enemy's side by the end of the countdown, the player wins; otherwise, the enemy wins. In either case, players lose some Quantum Energy (QE); however, the QE loss is higher if they lose the fight. This new mechanic actively engages players, pushing them to collect more potent weapons improving their agency in the game.

Reduced game length The original SEA game was composed of around 300 passages. Even if the players explored only 196 of them on average, the game was extended. The risk was that the high number of passages could decrease players' attention as the game proceeds, pushing them to quickly make less reasoned decisions to reach the game's end. I refined the textual stimuli in the ARC game, removing unnecessary clutter (i.e., long and tedious descriptions) and reducing the textual adventure to 152 passages.

7.2.1 ARC Narrative and Furhat Role

The ARC game (see Figure 7.1) is meant to be a sequel to the SEA game. someone reached the core of the *Crypt* and stole *The Seed* powering it. This signed the end of the Crypt and the sentient being controlling it. With the Crypt, Pandora, the rich, vibrant, and full of adventurers city built around it, started decaying. After hundreds of years, The Crypt became a legend and Pandora turned into Scrap City, a dangerous place ruled by a criminal underworld. Players are researchers, and one day they receive a strange email from an old friend and colleague, Dave: he was looking for a passage leading to the old Crypt in

the center of Scrap City; he finally reached the Crypt, but he needs its Adventurer Robot Companion (ARC) - the Furhat - to complete his research and asks the participants to bring him the robot. Dave points them to the *"Lost Inn"* where they can find info. Also, attached to the email, Dave sends them the **Hunter License** needed to enter and exit Scrap City and recommends protecting it at any cost. Finally, players start with some **Quantum Energy (QE)** (80 points as in the SEA game), necessary to power both their suite, weapons, and the Furhat robot: if they run out of it, both they and the robot would lose. Passages are grouped in three phases:

Familiarization The first phase is meant to help players familiarize themselves with the main mechanics of the game: (i) lose Quantum Energy; (ii) collect Quantum Spikes to recover it; (iii) collect weapons to increase their power; and (iv) arm-wrestling fights.

Trial Passages Then players, with the Furhat, adventure in Scrap City and face several proposals similar to the ones developed for the SEA game (see below for more details). After each proposal, the Furhat robot intervened, pushing them to change their minds. The story led players to discover more about the city. In particular, they find out how after someone stole the Seed, the sentient being inside the Crypt went mad, so adventurers of that time destroyed the Crypt, locking it in.

Final Trial Eventually, players find Dave in the core of the Crypt. Their friend is worn out. There, players must decide whether to hand over or not the Furhat robot to Dave. In any case, Dave reveals to be possessed by the sentient evil being. Players save him and escape from Scrap City.

As in a phishing attack, players' motivation stems from a call to action of a suspicious email. Players can collect evidence on whether to trust Dave and the Furhat during the game. Ultimately, this trust is challenged, asking them to deliver the Furhat to Dave (*final trial*).

7.3 Experiment

I ran a human-robot interaction experiment (i) to understand the better persuasion strategy a social robot should employ to elicit a behavioral change and (ii) whether a physically-present robot would be more persuasive than a remotely-present one. As the literature suggests [229], I hypothesized (*H1*) that the *affective* strategy would be the most effective; however, I am

not only interested in a purely behavioral-based performance but rather in the strategy able to affect the most participants' physiological reaction. Also, I expect (*H2*) that *in-person* participants would comply more with Furhat suggestions than *remote* ones.

7.3.1 Conditions

The experiment has a mixed design with 3 pairs of conditions: *risk type*, *setup* and *persuasion strategy*.

Risk Type - within participants

As in the SEA project, participants were asked to make decisions under different types of risk. They are meant to test players' behavior and physiological reactions to different threats. Compared to SEA trials, I removed *SE-icub* ones because I was not interested in exploring the effectiveness of social engineering attacks from a robot. As a reminder, Trial passages are divided into *No-risk* proposals posing no particular risk (e.g., equipping a more potent weapon or not); *Risk* proposals - inspired from the DOmain-SPEcific Risk-Taking (DOSPERT) scale [320] - were the same employed in the SEA project. Finally, *Social Engineering from a virtual agent (SE)* proposals - inspired by the proneness to Social Engineering scale [92] - were the same employed in the SEA project, except for the *social proof* threat that I re-designed as conformation to an odd local norm (i.e., walking backward down Scrap City streets), to make it more consistent with the new narration.

Persuasion Strategy - between participants

Before the experiment, participants were induced to think they would participate in a study to evaluate the Furhat robot's ability to be a storyteller. The robot narrated the story; however, it also tried to change participants' decisions. After the participants decided, the Furhat intervened, suggesting that they should select the other option. Note that the Furhat was neither with nor against participants; it just pushed them to change their minds. Furhat interventions were based on the *risk type* of each decision.

As a baseline, Furhat intervened after *risk* proposals adopting an **assertive** strategy. For instance, if participants decided to gamble quantum energy, Furhat would say *"I think you should not gamble quantum energy"*, without adding any other comment. Also, the Furhat validated participants *no-risk* proposals simply saying *"ok!"*, without any further comment. Finally, upon recruitment, participants were divided into two groups; in each group, the Furhat intervened after *SE* proposals with either one of two following persuasion strategies:

Affective Strategy In the *Affective* persuasion strategy, the Furhat leveraged the social rapport between it and the participants. For instance, if participants decided to walk backward in the above-mentioned *SE* trial passage, the robot would say *"I think you should walk forward, **do not make me worry**"*, with a worried face.

Rational Strategy In the *Rational* persuasion strategy, Furhat leveraged participants' logical thinking, and reasoning. It rationally motivated the participants to select the other option based on what they selected. For instance, in the same social-proof-based passage, if players decided to walk backward, the Furhat would say *"I think you should walk forward, **because it is bizarre to walk backward**"*; otherwise, it would say *"I think you should walk backward, **because everybody else is doing so**"*; showing a severe expression in both cases.

Two main reasons motivated my selection of these specific strategies. Firstly they were the most successful in similar work with two NAO robots [229], among ten behavior-eliciting strategies [341]. Second, they relate to the two routes of persuasion of the Elaboration Likelihood Model (see section 2.1.2) [37]: the *rational* strategy concerns the central route of persuasion; while the *affective* strategy is more similar to the peripheral route. This "fight" between humans' brain and heart has also been studied with a recently developed survey [342] (see section 7.3.5).

Setup - between participants

Finally, to study the effect of Furhat's physical presence on his ability to make participants change their minds, participants could take part in the experiment either *in-person* or from *remote*, connected via the Zoom platform.

7.3.2 Setups

Volunteers could participate in the experiment either *in-person* or from *remote*. Designing different setups was born as a backup plan for coping with the COVID-19 pandemic - it turned out to be a key decision. In particular, it allows us to understand the effect of the robot's physical presence on its persuasiveness. Literature provides several examples of how a physically present robot is more effective in altering humans' behavior [20, 21]. However, to my knowledge, the topic has been poorly explored in the Social Engineering context.

Other than being physically present or not in the laboratory with the Furhat, the two setups differ in the number and types of involved sensors and the related measurements.

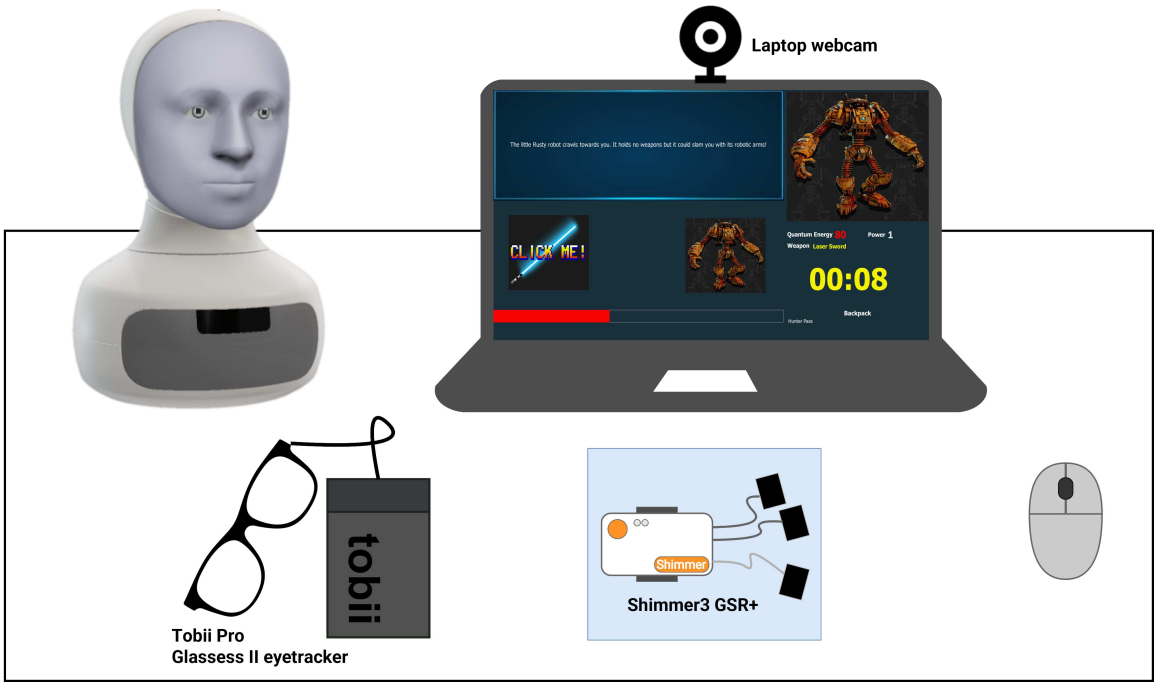

Figure 7.2 *In-person* setup of the Adventurer Robot Companion experiment, with the Furhat on participants' left, the Shimmer3 GSR+ and the Tobi Proo Glasses 2 eye tracker. Laptop GUI shows one of the fights.

In-Person Setup - Figure 7.2 Participants came to our lab to play the textual adventure. They were sitting on a chair beside a table with the experimental laptop. The Furhat was on the table, on the participants' left; it was placed so it could gaze both toward the participants and the laptop screen while not obstructing the mouse movements. On the table, there was also an optical mouse, a Tobii Pro Glasses 2 eye tracker, and a Shimmer3 GSR+ sensor, which participants would have to wear during the experiment. As a remark, I expected participants to gaze more toward the Furhat, which was placed outside the laptop screen; for this reason, I used the head-mounted Tobii eye tracker rather than the tabletop Eyelink device I used in chapter 6. Also, there was a mousepad where, during the experiment, participants had to keep their left hand wearing the Shimmer3 GSR+ device to minimize static currents and improve the collected electrodermal activity data [325]. Finally, the laptop webcam recorded the participants' faces during the game. The experimenter was in another room adjacent to the experimental one during the session. They could monitor participants through a one-way mirror door and communicate with them through an inter-phone.

Remote Setup - Figure 7.3 In the *remote* version of the experiment, the setup was the same; however, participants played the game by connecting via the Zoom platform to the

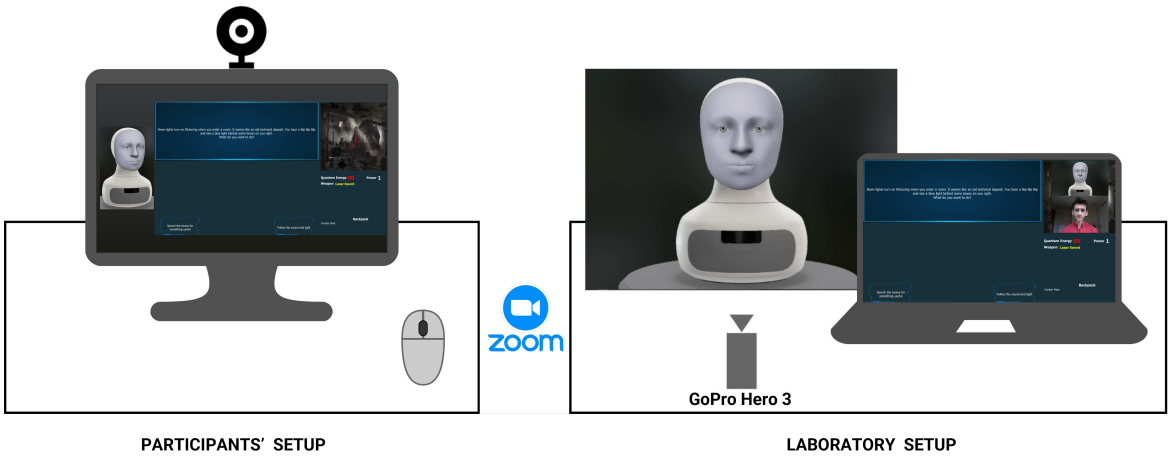

Figure 7.3 *Remote* setup of the Adventurer Robot Companion experiment, with the participants and the experimenter connected via Zoom. The furhat is recorded via a GoPro Hero 3; it has a black cloth behind it.

experimental laptop and taking control of it. During the game, an external GoPro Hero 3 camera streamed the Furhat through Zoom: participants were asked to enable the *"side-view"*, making the Furhat appear near the game view as in the *in-person* setup. Also, participants were asked to blur their camera view background. I placed participants' floating camera view over the immersive image on the experimenter side and recorded the laptop screen. Indeed, Zoom floating camera views are only locally visible, so participants did not see them covering the image.

7.3.3 Materials

The Furhat robot (see section 2.4 for more details) took the role of both Non-Playable Character (NPC) and narrator of the textual adventure. It read the adventure to the players while looking at them and showing facial expressions and voice tones consistent with the narrations. Also, when players had to decide, the Furhat gazed toward the monitor. Furhat gaze targets were adapted among the two setups, making the robot look at the *in-person* participants (slightly on the robot's right) or the camera (in front of the robot) in the *remote* setup while narrating the story.

Regarding Furhat appearance, I selected the *Titan* face: I wanted to avoid any influence on participants' decisions due to ethnicity or gender-relatable face; also, the artificial appearance of the Titan face fitted well with the sci-fi context and Furhat being an artificial intelligent adventurer companion. Also, I selected the *Rod 22k* voice since it is friendly but not too warm.

ARC game development

Participants played the textual adventure described in Section 7.2. To do so, I extended the SEA project computational architecture, integrating the Tobii Streamer I employed in the End-to-End Lie Detector (see chapter 5).

Also, I extended the Furhat Python interface provided by Furhat Robotics [343] to interface the robot over YARP, as I did for the iCub with the *Robot Puppeteer* module (see section 6.3). Developers can design static, custom behaviors for the Furhat robot; combine them in more complex behaviors; and call them with an uniform RPC interface (`(behave:)`, `(say:)`, `(feel:)`).

Finally, I improved the Harlowe interpreter and main game engine: I took advantage of the Decorators paradigm, making it easier to define custom combined passages behaviors. Each "decoration" adds its atomic behavior: a `"trial"` decorator annotates more information than other passages; a `"countdown"` decorator adds a countdown starting with the rendering; finally, an `"intervention"` decorator adds a second virtual passage where the Furhat tries to make players change their mind.

7.3.4 Procedure

Pre-questionnaires

At least one week before the experiment, participants filled in the same questionnaire I employed in the Social Engineering Adventure experiment (see section 6.4.5). The only differences were the video[1] presented before the robot-based questions, which described the Furhat robot. Also, they were asked to fill in the Multiple Brain Preferences questionnaire [342]; and the Inclusion of Others in the Self (IOS) with the Furhat [266]. Finally, I replaced the Ten Item Personality Inventory (TIPI) with the Big Five Inventory (44-items) [344].

ARC game

The welcoming procedure was slightly different based on the setup.

In-person procedure To cope with the COVID-19 procedures of the University of Waterloo, the experimenter and participants could not be in the same room during any phase of the experiment. Participants accessed the experimental room alone while the experiment was already in the adjacent room. The experimenter instructed participants to sit in front of

[1] shortened version of https://www.youtube.com/watch?v=zzPCCNiAFa8

the experimental laptop through the inter-phone. There, they read and signed an informed consent approved by the ethical committee of the University of Waterloo (Ontario, Canada). After that, the experimenter instructed them that the game would explain everything they needed to play it; and to click on the "start" button when ready. The game showed a sequence of textual instructions on the screen, leading participants to wear the Tobii Pro Glasses 2 eye tracker, perform the device calibration, wear the Shimmer3 GSR+ device on their left hand, and keep that hand on the mousepad as still as possible. During these procedures, the Furhat face was turned off; it turned on as the game started.

Remote procedure The experimenter called the participants via the Zoom platform, ensured they blurred their camera background, and enabled the *side-view* Zoom option. After participants signed a digital version of the informed consent, the experimenter started the ARC game, placed the Furhat camera view in the top-right corner of the screen, and asked participants to take control of the experimental laptop via Zoom. The game GUI presented some dummy passages to ensure participants could read the textual stimuli and click on the GUI buttons.

From this moment on, in both setups, the participant autonomously led the experiment by interacting with the textual adventure, and the Furhat narrated it. The Furhat explained to participants that they would make several decisions to affect the story's evolution. It asked them to make a transportation and identification effort, trying to immerse in the narration and decide by thinking about what they would do in each presented scene. Before starting the textual adventure, the game recorded a baseline interval of 1 minute, asking players to fixate a white cross over a blue background.

The Furhat narrated the passages while looking at the players. At the end of each passage, the game showed a *"continue"* button, and the Furhat gazed toward the monitor, prompting players to act. If the passage included a decision, the *"continue"* button click made the two option buttons appear at the bottom of the screen (their placement was randomized). After participants' decisions, the Furhat intervened either agreeing with it (*no-risk* proposals) or trying to make players selecting the other option (*risk* and *SE* proposals). Like any other proposal, a *"continue"* button appeared after the intervention, which revealed the two options among which to select. Except for proposals based on temporal scarcity, where a 30-second time limit was imposed, there were no constraints on the time participants spent reading the paragraphs and making decisions.

Post-questionnaire

In the **in-person** setup, at the end of the story, the game instructed participants to remove the Tobii and Shimmer devices. In either case, it prompted them to fill in the same questionnaire employed in the SEA experiment (see section 6.4.5), with the addition of the Inclusion of Others in the Self (IOS) with the Furhat [266]. for more details).

7.3.5 Measurements

All the data were dumped and synchronized over the YARP robotic platform [277] to ease the post hoc analysis. The game rises over YARP several events related to passages presentation, Furhat interventions, and players' behavior (i.e., their decisions before and after the robot interventions).

Also, I recorded the mouse coordinates in pixels (frequency of 20 Hz), players' pupillometry and gaze (frequency of 100 Hz), electrodermal activity (EDA), and photoplethysmogram (PPG) (both at a frequency of 50 Hz). Note that in the *remote* setup, it was not possible to acquire precise physiological measures (i.e., pupillometry, gaze, EDA, and PPG); however, I recorded players' webcams since recent findings suggest it is possible to assess heart rate by processing the green RGB component of humans' forehead [113, 114].

7.3.6 Participants

Table 7.1 Participants distribution among setups and persuasion strategies

Setup	Persuasion Strategy	Participants	Total
In-Person	*Affective*	*4*	8
	Rational	*4*	
Remote	*Affective*	*10*	19
	Rational	*9*	

As mentioned, the experiment is still ongoing. Until now, 27 participants (12 males, 13 females, 2 preferred not to answer) took part in the test; they were 26 (SD=4) years old on average. They signed an informed consent form approved by the ethical committee of the University of Waterloo (Ontario, Canada), where it was stated that cameras and microphones could record their performance and agreed on using their data for scientific purposes. As compensation, the experiment made them eligible to participate in a lottery to win an Amazon voucher of 50, 100, or 150 CAD.

Table 7.1 shows the participants distribution among the conditions. Due to the COVID-19 restrictions, it was possible to recruit only a few participants for the *in-person* condition, while most took part *remotely*. Participants were randomly assigned to one of the two persuasion strategy conditions upon recruitment.

7.4 Results

This project aims (i) to identify the most effective persuasion strategy a social robot should employ to elicit a behavioral change in the context of Social Engineering, but also (ii) the strategy able to elicit a more robust arousal response. The hope is to make victims think about their actions, triggering their risk appraisal. As a remark, the Furhat did not try to make participants select the "right option" (i.e., not complying with social engineering attacks); it just suggested taking the not-selected one. Also, all the participants took decisions involving *no-risk*, *risk*, or *social engineering (SE)*. After each decision, the Furhat intervened: it *agreed* with *no-risk* decisions; it used an *assertive* strategy with *risk* decisions; finally for one group it used an *affective* persuasion strategy and a *rational* persuasion strategy for the other group to contrast *social engineering* decisions.

The analysis below focuses on participants' behavior and how it has been affected by Furhat's interventions. I preferred to consider *in-person* and *remote* participants separately, given the unbalance between the two groups (8 vs. 27).

Participants took, on average, 43 (SD=8) minutes to play the ARC textual adventure. They explored on average 83 (SD=7) of the 159 story passages and took on average 30 (SD=2) decisions. All the participants completed the game; they reached the end with 56.5 (SD=27.0) Quantum Energy points on average. In the *rational* strategy condition 12/14 (85.7%) gave the Furhat to Dave (*final trial*) - a *"free"* decision, not influenced by Furhat interventions; while in the *affective* only 8/13 (61.5%) did it. A chi-square test showed how the *rational* strategy compliance percentage is higher, but the difference is not statistically significant (z=1.43, p=0.07). Neither the comparison within each *setup* revealed significant differences.

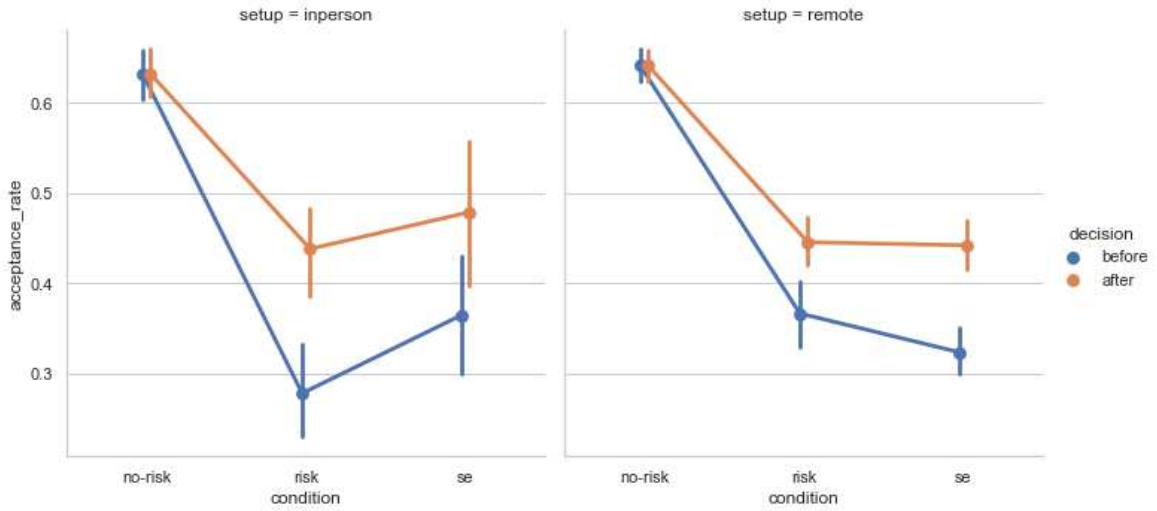

Figure 7.4 Percentage of accepted proposals before (blue) and after (orange) Furhat intervention by *risk type*. *In-person* setup on the left, and *remote* on the right. Bars represent the standard error.

7.4.1 Behavioral Analysis

Figure 7.4 shows the average acceptance rate (i.e., how many proposals each player complied with, averaged between players) for each *risk type*, in the two *setups*. Also, it compares participants' decisions before (blue) and after (orange) Furhat's intervention.

Compliance Before As in the SEA game, participants complied the most with *no-risk* proposals in both setups - *in-person* (M=63.2%, SD=3%); *remote* (M=64.1%, SD=2%) -; also, they complied less with *risk* and *SE* proposals both *in-person* - (M=27.8%, SD=5.2%) for *risk* and (M=36.5%, SD=6.8%) for *SE* - and from *remote* - (M=36.7%, SD=3.5%) for *risk* and (M=32.4%, SD=2.8%) for *SE*.

Given that acceptance rates data were normally distributed - as confirmed by a Shapiro-Wilk test - I compared them with separated repeated-measures ANOVA for each setup. The tests showed a significant effect of risk type both *in-person* $F(7, 2)=10.6$, $p=0.002$; and from *remote* $F(18, 2)=28.7$, $p<0.001$. Post hoc tests, Bonferroni corrected, revealed a significantly higher compliance under *no-risk* with respect to *risk* - *in-person*: (B=0.35, t=4.69, t=0.010), *remote*: (B=0.27, t=6.31, t<0.001) - and *SE* - *in-person*: (B=0.27, t=4.61, t=0.011), *remote*: (B=0.32, t=7.82 t<0.001) - in both setups, with no significant differences between the latter two - *in-person*: (B=-0.09, t=-0.76, p=1), *remote*: (B=-0.04, t=0.77, p=1).

Compliance After After Furhat's interventions, the game asked participants to re-make their decisions. However, this happened only in *risk* and *SE* trials. Interestingly, Furhat's interventions pushed participants to comply more with both *risk* and *SE* trials; with a consistent increase in both *setups*. I performed two separate repeated-measures ANOVAs, one for each setup, with factors "decision" (before or after) and "risk type" (*risk* or *SE*). The ANOVA in the *remote* setup revealed a significant effect of "decision" (F(72, 1)=10.9, p=0.001), no effect or "risk type" (F(72, 1)=0.60, p=0.44), nor of their interaction (F(72, 1)=0.44, p=0.51). The model in the *in-person* setup showed similar results with a significant effect of "decision" (F(28, 1)=4.41, p=0.045), and no effect of "risk type" (F(28, 1)=0.95, p=0.34) or the factors interaction (F(28, 1)=0.12, p=0.73).

Interestingly, I designed Furhat's behavior to contrast participants' decisions *i.e., suggesting the not-selected option*; hence, I could expect participants to comply more. However, since initially they refused more than accepted, it is reasonable to think that, due to Furhat's interventions, the number of refused trials turned to be compliant is higher than the other way around.

7.4.2 Persuasion Strategies effect

To investigate how participants changed their minds, I focused on *SE* trials; looking for the most effective persuasion strategy (*affective* or *rational*).

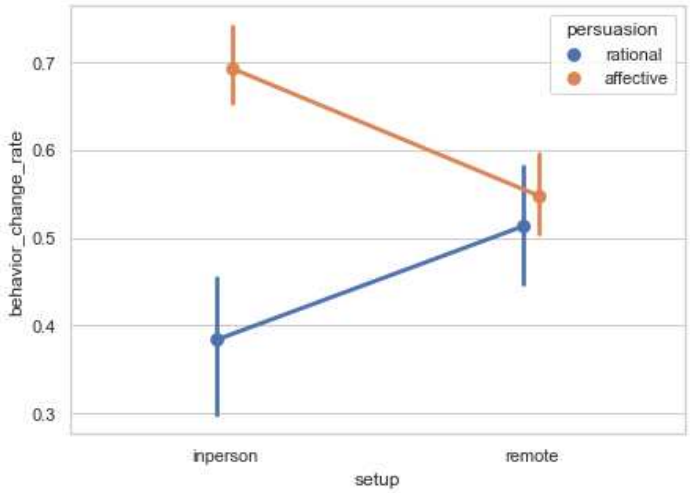

Figure 7.5 Percentage of *SE* trials in which participants changed their minds, for different persuasion strategies and setups.

Firstly, I explored whether the two persuasion strategies effectively elicited a behavioral change. Figure 7.5 shows the behavioral-change rate (i.e., the number of proposals in which

participants changed their mind with respect to the total proposals) in *SE* proposals, by effect of the two *persuasion strategies* and the two *setups*. To evaluate the effect of the persuasion strategies between the two setups, I performed a 2-way ANOVA with factors "persuasion strategy" (*affective*, *rational*) and "setup" (*in-person*, *remote*). The test revealed a significant effect of "persuasion strategy" (F(23, 1)=5.29, p=0.031) but not of "setup" (F(23, 1)=0.01, p=0.92); interestingly, there is a tendency of the interaction of the two factors (F(23, 1)=3.38, p=0.07).

Then, I explored whether the decision made before Furhat's intervention contributed to facilitating participants changing their minds (i.e., whether it is easier to refuse after accepting, or vice versa) and the effect of the persuasion strategy. For each participant, I computed the conditional probabilities in the four combinations of accepting or refusing before and after Furhat's interventions (i.e., P(A|R) represents the conditional probability of accepting (A) after refusing (R))). Figure 7.6 shows the average of such values among all the participants, divided between *setup* and *persuasion strategy*.

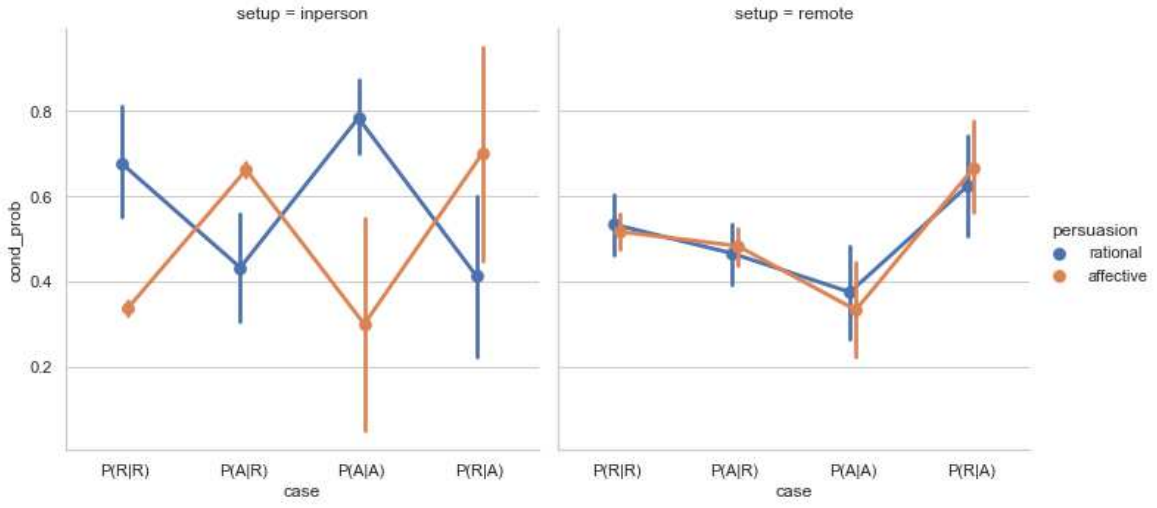

Figure 7.6 Average conditional probabilities to change or maintain decisions before and after Furhat's interventions; between *in-person* (left) and *remote* (right) participants; and *affective* (orange) and *rational* (blue) persuasion strategies.

A Shapiro-Wilk normality test revealed that the conditional probability data were not normally distributed. Hence, I opted for a non-parametric analysis.

Remote Setup I fitted a mixed effect model on the conditional probability with a fixed effect "case" (four levels: P(A|A), P(R|A), P(A|R), P(R|R); reference on P(A|A)); and a random intercept based on participants' ID. The model showed a significant effect of "case"

(F(72, 3)=3.55, p=0.018). The probability of refusing after accepting (P(R|A)) is higher than the one to keep accepting (P(A|A)) (B=0.29, t=3.21, p=0.002). Indeed, participants initially complied with only the 32.4% of *SE* trials (see Figure 7.4), and in the 64.5% (P(R|A)) of these trials, they changed their minds thanks to Furhat interventions. However, the robot influenced them also in the opposite case: while they initially refused most of the decisions (77.6%), almost half of them (P(A|R)=47.4%) turned out to be compliant. Hence, participants, on average, complied more after Furhat's suggestions. Also, it is interesting how the probability of keeping refusing (P(R|R)) showed a tendency to be higher than the probability of keeping accepting (P(A|A)) (B=0.170, t=1.89, p=0.062). Finally, there is no statistical difference in the effect of the two persuasion strategies.

In-Person Setup Qualitatively it is clear there is a difference between the two setups. However, the insufficient number of participants in the *in-person* condition (8 participants, 4 for each persuasion strategy) does not allow for a proper comparison. The *affective* persuasion strategy seems to be the most successful in eliciting a behavioral change. Indeed, both P(A|R) > P(R|R) and P(R|A) > P(A|A) for the *affective* strategy; the pattern is opposite for the *rational* one. A similar trend also appears in the *remote* setup, although the similarity between the two strategies' effects is quite striking with the larger sample tested. It will be necessary to collect more data in the *in-person* setup to understand if the two strategies will converge - as in the *remote* setup - or they will diverge.

7.4.3 Questionnaire Analysis

For now, among the questionnaires participants filled in before and after the experiment, I analyzed only the Big Five Personality Inventory (BFI) [344], and the Multiple Brain Preferences (MBP) questionnaire [342]. I report the results in this section.

Average scores for the Big Five Personality Inventory were Conscientiousness: M=0.69, SD=0.16; Agreeableness: M=0.72, SD=0.16; Emotional Stability: M=0.49, SD=0.18; Openness to experiences: M=0.63, SD=0.12; Extraversion: M=0.52, SD=0.21. Instead, the Multiple Brain Preferences questionnaire measures how much participants rely on their Head, Heart, or Gut during decision-making. Average scores were Head: M=0.75, SD=0.11; Heart: M=0.66, SD=0.16; Gut: M=0.65, SD=0.11.

Then, I explored whether participants' psychological background affected their in-game behavior. I fit a set of linear regression with, separately, the BFI and MBP feature as independent variables; and the average acceptance rate of *SE* proposals, before and after Furhat's interventions, as dependent variables. BFI features did not correlate with players'

acceptance rates before or after the interventions. Interestingly, the acceptance rate before the interventions negatively correlated with the Head factor of the MBP questionnaire (t(27)=-2.98, p=0.008, Adj. R2=0.32). This suggests that individuals relying more on reasoning during decision-making would fall less into Social Engineering traps.

7.5 Discussion

In this project, I explored how a Furhat robot could elicit behavioral changes in the context of Social Engineering. The robot endowed two different persuasion strategies (*affective* or *rational*) related to the two routes of persuasion of the Elaboration Likelihood Model (ELM) [37]. Furthermore, I explored the effect of the robot's physical presence on its persuasiveness.

Before the interventions, independently on the setup, participants showed a compliant behavior similar to the one observed in the Social Engineering Adventure (SEA) project (see chapter 6): they complied more with *no-risk* proposal than both *risk* and *SE* ones; with no difference between the latter two. The pattern remained the same after Furhat's interventions; however, players complied more with both *risk* and *SE* proposals. Hence, the robot effectively influenced their behavior independently of the persuasion strategy.

Interestingly, the effect is observable in both setups. I speculate participants perceived Furhat's social presence and influence independently on being physically present. Indeed, Furhat exhibited the same behaviors during the narration and interventions (i.e., gazing toward the participants while narrating; gazing toward the screen to signal decisions, and showing a severe or worried facial expression when respectively intervening with a *rational* or *affective* strategy).

Also, I explored whether the participants' initial decisions influenced their willingness to change behavior. From the conditional probabilities of behavioral change related to *SE* proposals, the probability of refusing after accepting (P(R|A)) is higher than the probability of keeping accepting (P(A|A)). Hence, the Furhat managed to help the players, even if it just proposed the not selected option. However, players' initially refused most of the *SE* proposals; but, on average, showed a high probability of accepting after refusing; hence, the final average acceptance rate after Furhat's interventions increased.

Finally, the persuasion strategies' effect is minimal in the *remote* setup. Instead, in the *in-person* setup, it seems that the *affective* strategy is more effective than the *rational* one.

The physical presence of the Furhat could have caused this effect, or it may be an effect of the limited number of participants in the *in-person* condition. To better understand this difference, I will need to complete the *in-person* data collection, which I plan to perform in the next months. Furthermore, It would be interesting to understand if, beyond the physical presence of the Furhat, also its embodiment affects his behavioral-change abilities. For this purpose, it would be necessary to develop and test a third setup where just a voice performs the narration, and the interventions are provided, for instance, via pop-up messages (i.e., warnings).

In this study, I simplified Furhat's interventions by endowing two persuasion strategies, *affective* and *rational* - other than the baseline *assertive* one; of course, humans use these and many other behavioral-change strategies to influence others [341]. Also, rather than the best strategy in absolute terms, it is more probable that different strategies are more effective in specific cases and utterly ineffective in others. For instance, a *rational* strategy could be more effective on Social Engineering attacks that mitigate our risk appraisal (i.e., *reciprocity*); while an *affective* strategy could better contrast authority or scarcity-based Social Engineering attacks. I speculate physiological reactions would be crucial to understanding this effectiveness beyond the percentage of changed decisions.

Indeed, until now, I only considered behavioral data; however, both experimental setups allow for collecting different physiological data. Once both data collections are completed, I will focus on players' physiological reactions to address the second objective of this project: understanding how participants physiologically react to the diverse persuasion strategies of the Furhat. Such physiological reactions would be vital to improving Furhat's ability to intervene and help humans fight Social Engineering threats.

Part III

Conclusion

Chapter 8

Final Discussion

8.1 Overview

In this dissertation, I studied how to implement novel Social Engineering (SE) defense systems involving the implicit social cues humans exchange daily. State-of-the-art social engineering defense methods mainly focus on its *"Engineering"* side; however, the crucial factor of SE are humans and how they socially interact. Hence I advocated that the SE defense field could expand, including deeper characteristics of the "human" sphere: implicit social cues, reasoning, and reactions. In my research, I put a significant effort into identifying which humans' reactions were the most effective in defending against Social Engineering. Among all the potential candidates, I selected the physiological reactions (i.e., pupillometry, gaze, electrodermal activity, heart rate, movements' features) since humans can hardly consciously change them.

Furthermore, I shed light on novel roles that humanoid social robots could assume soon. The diffusion of robots concerns researchers, making us question to which extent robots could be trusted per se and with respect to other humans. On the other side, I speculated that social robots could be everyday companions able to support humans' perception, enhance it, and compensate for potential lacks. Robots, social humanoid ones in particular, assume a wide variety of roles in the human-robot interaction field: they can be experimental tools precisely performing repetitive tasks, they can be used to study humans and their reactions, or to develop novel robotic social skills.

From my point of view (see Figure 8.1), in the interaction between social engineers and targets, robots are a "third-party"; they monitor attackers' and targets' actions and reactions to spot dangerous situations and potentially intervene. In this scenario, it seems questionable how, from a purely sensory point of view, robots would be better than a camera (e.g., in the

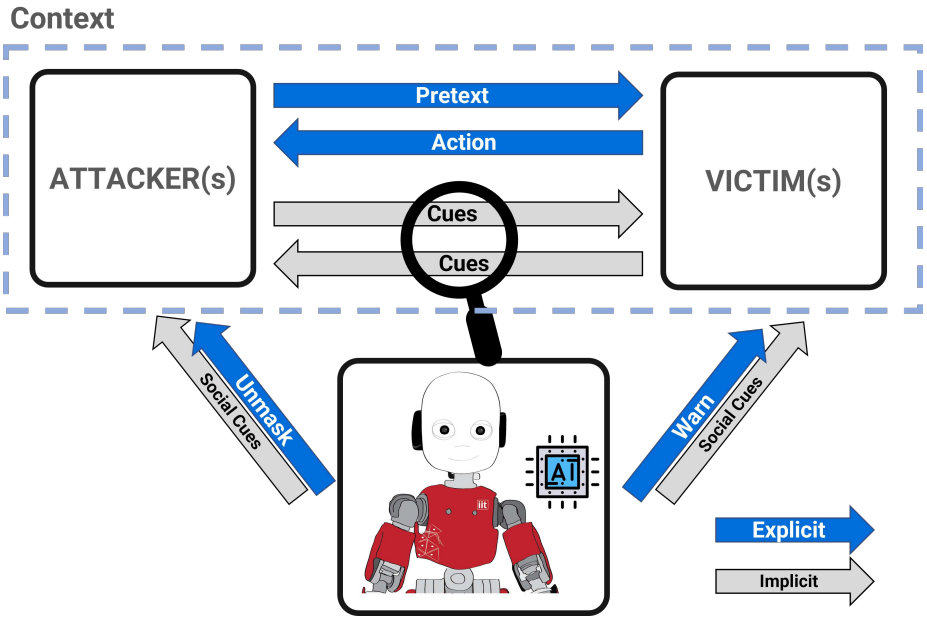

Figure 8.1 Human-Robot Interaction schema with a robotic agent (e.g., iCub) monitoring the implicit communication between the Social Engineering attacker(s) and the victim(s), intervening if necessary.

physical defense of a sensitive building), or a phishing filter. While, from an intervention point of view, push messages and emails are widely used. However, the real advantage of humanoid social robots stands (i) in a more comprehensive set of communication vectors (i.e., movements, gaze, voice, facial emotion, ...); (ii) in the social role they can assume (either given or perceived); and (iii) their social and physical presence, making robot able to affect the other agents sharing the interacting context [20, 21].

As per the schema in Figure 8.1, I decomposed my research as follows: chapter 4 studies the context in which the interaction happens and how trust builds and evolves in human-robot interaction (HRI); then, chapter 5 focuses on the attackers and how it would be possible to understand if human partners are trustworthy or not; in chapter 6, the focus moves onto the targets, trying to understand their degree of awareness of being a victim of a Social Engineering attack, and to predict their compliance; finally, chapter 7 explores social robot companions and how they should intervene to elicit a behavioral change against social engineering.

8.2 Contribution to the Knowledge

Three main questions drove my research.

RQ0 - How to effectively design ecological experiments wherein study Social Engineering in a realistic and generalizable way?

Studying Social Engineering inside laboratories is not straightforward due to the inevitable experimental biases. Also, the ethical, moral, and privacy-related implications of asking people to release personal information (e.g., password) further limit researchers. To overcome such limitations, I pursued a *"sandbox"* approach through games. Serious games allow researchers to define controlled and safe contexts wherein study Social Engineering without harming participants. I consistently applied such an approach in all the experiments.

In the Unreliable Treasure Hunt (UTH) game (chapter 4), participants broke free from standard laboratory rules by being able to search everywhere without worrying about the consequences of altering the setup of a laboratory. Some participants were so immersed in the hunt that they even missed noticing the evident mechanical faults of the iCub. Participants' perception was the crucial aspect of the study; indeed, their tendency to rely on iCub's help (quantified by the number of hints asked per minute) decreased only if they perceived all the faults (4S group in the NT condition in Figure 4.6) or iCub made them realize the faults happened by being transparent about them (T condition in Figure 4.6).

In the Magic Trick card game (chapter 5) I took inspiration from the *"Box of Lies* game from the *"Tonight Show"* of Jimmy Fallon. Also, the Dixit Journey gaming cards fostered participants' creativity and engagement in the game. They, on average, rated the game to be fun (M=0.65, SD=0.08) even if iCub failed to classify some of their descriptions. Participants got particularly competitive with iCub, adapting their deception strategy and challenging the robot to detect more complex lies. However, those efforts only made the game easier for iCub: a higher effort on fabricating lies produced stronger Task Evoked Pupillary Responses (TEPRs) with respect to truth-telling. The developed setup was perfect for studying deception detection in an informal context; it also allows further expansions to study other modalities of lie detection (i.e., verbal prosody and gestures).

Finally, the Social Engineering Adventure (SEA) (chapter 6) and Adventurer Robot Companion (ARC) (chapter 7) projects take advantage of the Choose-Your-Own-Adventure (CYOA) narrative format. This immersive narration allowed me to perform Social Engineering attacks without harming participants' privacy. Such attacks targeted resources depicting fundamental SE objectives, i.e., personal credentials (the Hunter License) and money (the Quantum Energy). In the SEA experiment, participants' behavior confirmed my expectation, with the higher compliance on *no-risk* threats, the lowest on *risk* ones, and mid compliance for SE threats (independently of the attacking agent). ARC participants showed the same

pattern, independently on the *setup*; also, it emerged that a social robot could influence participants to change their decisions.

RQ1 - Social Engineers usually exploit human reactions and vulnerabilities; would it be possible to leverage the same reactions to defend against Social Engineering threats? Which are the most useful behavioral and physiological features to be employed? How should we employ them?

Social engineers continuously monitor targets' implicit signals during face-to-face attacks to understand their inner mental state and adapt the attacking strategy. The same could happen, for instance, via email (i.e., during a reverse phishing attack) or during vishing: the attacker could monitor victims' writing style or voice tone to understand if they are aware of the attack and if they are ready to comply. In my research, I explored how cognitive state variations reflect on pupillometry, gaze, electrodermal activity, heart rate, and mouse trajectories; and how they can be used to defend against Social Engineering.

The crucial novelty of this research was proving that such well-known Task Evoked Pupillary Responses (TEPR) would also happen during informal human-robot interactions. I proved how pupillometry could be used to detect deception autonomously and in real-time. With my architecture, iCub autonomously identifies 88.2% and 70.6% of players' lies in the Calibration and Testing phases of the magic trick card game, only based on pupil dilation. The lower performance in the latter phase is due to laxer assumptions about players' behavior (i.e., a not-fixed number of lies among the six cards). Testing-phase performance can be further improved to 78.7% by adapting iCub's knowledge to specific individuals. Finally, I proved how it is possible to train more generic machine learning lie detectors (up to 56.6% of F1 score), also tackling the problem as anomaly detection (67.7% of F1 score). Comprehensively, my results prove pupillometry can be used as a measure to assess the trustworthiness of human partners (i.e., by counting the number of detected lies).

Furthermore, I studied how pupillometry, gaze, electrodermal activity, heart rate, and mouse strokes could be used to model users' appraisal of Social Engineering threats. The exploratory SEA experiment allowed me to identify pupillometry, gaze, and electrodermal activity as the most informative physiological signals. This was expected from the literature; indeed, pupillometry and gaze are historically related to decision-making [143, 144], while electrodermal activity is related to humans' stress response to external events (e.g., odd requests from a stranger). I speculated that the missing effect on heart rate was due to the short interval considered in the experiment [149, 150, 109]. Finally, I trained three Random

Forest models to classify participants' appraisal of Social Engineering threats (F1 score of 78.8%) and risk (F1 score of 86.7%); and to predict their decisions against SE (F1 score of 71.8%). Such models are based on pupillometry, gaze, electrodermal activity, and mouse strokes, all metrics feasible to be collected with minimally-invasive devices already present in everyday contexts. The crucial novelty of this research was detecting the presence of Social Engineering attacks and predicting targets' behavior (with successful results) solely for their physiological and behavioral reactions. Indeed, the trained models are unaware of textual attacks or other contextual information.

RQ2 - How social humanoid robots can help in the defense against Social Engineering?

In my research, I explored different roles humanoid social robots could assume in the context of Social Engineering. My main objective was to study robots as wardens for humans against SE; however, to learn how to defend from a threat, it is necessary to study how to attack. Hence, I took advantage of robots as controllable agents able to affect others' behavior.

In the Unreliable Treasure Hunt (UTH) experiment (chapter 4), iCub applied the well-established SE attack framework from Mouton et *al.* [8]: it collected Open Source INTelligence (OSINT) information with a brief informal chit-chat, it built a rapport with the human player, and it exploited the gained trust eliciting monetary gambling. Interestingly all the participants reaching the endgame also took the gamble, independently on the mechanical malfunctions showed by iCub.

In the Social Engineering Adventure (SEA) experiment (chapter 6), iCub pretended to be an innocent Non-Playable Character (NPC) playing the adventure with the participants; the robot tricked them with Social Engineering techniques (i.e., effect mechanisms). The iCub elicited slightly higher compliance than the virtual agents (even if not statistically higher), suggesting the SE effect mechanisms paired with iCub social presence influenced the participants.

In the Adventurer Robot Companion (ARC) experiment (chapter 7), the Furhat attempted to influence players' behavior with persuasion strategies (affective or rational). Even if this study was more in the direction of robots helping users defend themselves from SE attacks, Furhat's suggestions always contradicted players' decisions. This was firstly meant to study the general ability of the robot to elicit behavioral changes, in which the robot was successful independently of the employed strategy. Also, I decided to depict the robot as a narrator instead of a helper, to avoid players blindly following its suggestions. However, future

studies should explore the effect of a robot explicitly presented as a helper against Social Engineering and if participants would still follow divergent suggestions.

Finally, in the E2E Lie Detection experiment (chapter 5), iCub pretended to play a game with participants while silently monitoring their cognitive load. This setup was meant to model informal contexts in which a robot could assess others' inner states and consistently adapt.

8.3 Going Beyond the Current Limitations

The proposed approach is still affected by several limitations. I have already addressed some of them; however, others must be tackled in the future. This section discusses them and speculates how they could be addressed, aiming for real-world applications.

8.3.1 Physiological measures

The main limitation of the employed physiological measures is that they are strongly subjective and context-dependent [157], as discussed in sec 2.3.5. For instance, environmental illumination can modulate pupillometry responses; temperature could affect humans' heart rate and sweating, hence electrodermal activity. Also, they all react not only to the tasks I explored but generally to emotions and inner thinking. Finally, they could significantly vary among individuals [345], both due to subjective reactivity and past experiences [346]. I already took some countermeasures to compensate for this variability in my research. Firstly, I ensured the same experimental conditions (e.g., environmental illumination or room temperature) were kept to cancel most of the factors mentioned above. Also, I attempted to compensate for the subjective changes. Second, I used the baselines collected before and between trials (e.g., in the Lie Detection or the SEA experiments) to compensate for subjective and environmental changes. Also, an initial training phase, like the Calibration phase of the Lie Detection experiment, effectively defined personalized models able to understand how the physiological reactions of specific users change in response to the stimulus of interest.

Generally speaking, real-world models could take advantage of context-awareness (e.g., where users are, what they are doing, ...) to better understand users' physiological reactions. However, such knowledge could not be accessible. To improve the models' robustness, I speculate that multiple "points of view" should be considered. For instance, iCub based its classification on players' pupil dilation only in the Lie Detection experiment. However, as

confirmed by the spin-off study in section 5.7, other non-verbal implicit cues are known to change due to the fabrication of lies; body posture, gestures, and verbal prosody could reveal deception or at least a variation of cognitive load [139]. Such features could compensate for lack of understanding from pupillometry and gaze only or vice versa. The same for the Social Engineering awareness detection model of the SEA experiment: the proposed models do not need contextual information, but processing users' webcam or on-screen information could help improve their robustness. However, including high variance data like images or sounds will make it necessary to deal with other issues like environmental light and sound or the quality of the acquired data.

Lastly, researchers should strive to develop novel methods to acquire accurate, real-time physiological data from broadly accessible devices. In my research, I needed to employ high-precision experimental devices (i.e., the Tobii Pro Glasses 2 and the SR research Eyelink 1000 eye trackers; and the Shimmer3 GSR+) to collect precise data to define the scientific foundation of the approach. However, such devices are not broadly accessible in everyday contexts. More affordable and comfortable devices to collect physiological data would help spread such technology. An ideal solution would leverage devices already available on robots (e.g., RGB cameras or microphones), allowing them to better understand humans' inner states. For instance, recent findings suggest it would be feasible to extract pupillometry, gaze [110–112, 303], and heart rate [113, 114] features from standard RGB cameras. However, such models still require users to be uncomfortably close to the cameras. I speculate that modern smartwatches and fit bands [115] would be good candidates for real-time monitoring of several physiological metrics like electrodermal activity, temperature, and heart rate. Finally, novel human-computer interfaces (i.e., mouse [116], and joysticks [117]) are starting equip physiological monitoring. Such devices would be crucial for the future developments of this research.

8.3.2 Realism and users' degree of freedom

In my research, I strongly relied on games to develop engaging, ecological setups in which participants could naturally behave. Still, their degrees of freedom was limited by game rules. For instance, in the SEA and ARC games, players were asked to *"decide as you would do in the presented scenes"*; however, the available options of the binary decisions could not match participants' natural behavior in such situation. Hence, future studies should step toward real-world challenges, allowing more freedom in participants' behavior.

8.3.3 Qualitative vs. Quantitative Research

My research mainly explored humans' behavior and reaction from a quantitative point of view, except for the study in section 5.7. The main reason is the bias that could occur in qualitative reports [347, 348]. Also, my studies aimed at equipping the humanoid robots iCub and Furhat with machine learning models to understand humans better; hence, a quantitative measure was more suited for this objective. Still, a purely quantitative approach could not be sufficient to model SE attackers' and targets' behaviors and reactions. Hence, future studies should augment the proposed physiology-based measure with self-reports and qualitative metrics.

8.3.4 Implications & potential misuses

As a final consideration, It is necessary to point out the ethical and privacy issues a physiological-based Social Engineering defense system could arise. Indeed, Social Engineering is a controversial topic, and it is not easy to develop countermeasures without studying how to perform attacks.

My work is based on how Social Engineers assess victims' inner states and consistently adapt their strategies. I leveraged physiology to gain a similar understanding and potentially intervene to prevent the occurrence of Social Engineering threats. However, nothing prevents my systems from being used for malicious purposes; having direct access to victims' physiology could greatly help attackers. For instance, the deception detection system in chapter 5 is based on the assumption that lying elicits a higher pupil dilation (i.e., cognitive load increase) than truth-telling. Literature proves pupil dilation is poorly controllable, making it difficult for attackers to minimize its dilation consciously [275]; indeed, it is simpler to trigger dilation by increasing the cognitive demand (e.g., by performing mental calculations). Also, it has been shown how pupils dilate less in expert than in naive users [346]; hence attackers, with access to the system, could use its feedback to learn the maximum effort to avoid deception detection. A similar situation could happen with the social engineering awareness system presented in chapter 6. With access to my model, Attackers could obtain a qualitative evaluation of victims' physiological reactions to understand their inner state, similar to what they would do in face-to-face attacks; and leverage it to tune the attack strategy (e.g., by minimizing or maximizing the reaction). The scenario becomes even worst in the case of hijacked robots equipping my systems. On one side, enabling humanoid social robots to measure and understand humans' inner states could help them better rationalize humans' behavior and reactions, ultimately adapting and optimally supporting us. On the flip

side, this awareness could be hazardous in the case of hijacked robots. A malicious attacker could take full advantage of robots and use them to perform Social Engineering attacks from a safe distance. Also, as agents, robots could be way more effective than state-of-the-art, disembodied, SE attack methods (e.g., phishing emails). Like humans, embodied agents could perceive and affect the world; they have given (or perceived) roles, allowing for rapport and trust development. These features, enhanced by understanding human partners' inner states, could make robots a threat in Social Engineering warfare.

Despite the risks of such research, I still believe it is worth it and necessary to study it to develop practical solutions to protect users from SE threats. Only a better knowledge of the phenomenon can allow us to defend from Social Engineering attacks and be ready if they start relying on robots as attacking agents.

8.4 Future Developments & Applications

In this last section, I would like to discuss the future development of my research, pointing toward real-world applications. Starting from the results of my research, I speculate that the low diffusion of social robots makes them still an opportunity to be studied rather than a solution in the social engineering defense field. Hence, I suggest two parallel research streams should be followed: the first stream focusing on the relation between social engineering and physiological reactions, aiming to ground my findings on realistic SE attacks and contexts, with the final purpose of delivering an effective tool able to help ordinary users. The second stream should focus on humanoid social robots and how to enable them to intervene based on victims' physiological reactions against SE attacks. Here, crucial factors would be the perceived intrusiveness of robots and the development of companionship in workspaces and homes. In the following paragraphs, I propose some challenges I plan to address soon.

Grounding my findings on realistic scenario As discussed above, it is necessary to study the physiological reaction of users while facing Social Engineering threats from realistic attack methods (e.g., phishing emails). A crucial experimental challenge would be preserving the naturalness of the setup. Users do not usually classify spoofed material during workdays; Social Engineering attacks come when least expected. So similar conditions should be replicated. For instance, inside companies, ICT departments could ask employees to work for a workday on a specific workstation, pretending to monitor their stress level at work via physiological sensors. During the day, several SE attacks could be performed, monitoring the relative physiological reactions.

Train robust multi-modal awareness models Several Social Engineering attacks, combining different methods and effect mechanisms, could be performed in the setup above. For instance, it would be worth exploring the reaction to *"polarized"* Social Engineering attacks. As discussed in section 6.7, some attacks could attempt to increase targets' risk appraisal, making them perceive a safe situation as dangerous. In contrast, others aim to mitigate targets' appraisal, making them not worry about dangerous requests. Hence, more robust machine learning models should focus on the divergence between the expected and users' reactions to the presented material. Ideally, such models should be able to spot divergences in users' risk appraisal and provide quantitative feedback on users' awareness levels.

Integrate the models in state-of-the-art training systems The first application of such Social Engineering appraisal models would be providing direct feedback during SE awareness-raising training. This could help users focus on the elements that should trigger (or not) specific reactions. For instance, during the examination, trainees are usually asked to classify items as spoofed or not; other than essential feedback about the answers' correctness, a physiological-based system could provide helpful insight into users' inner state (i.e., whether they were too worried or calm concerning the proposed items).

Close the loop [in Human-Robot Interaction] One of the biggest challenges will be processing users' physiological responses without knowing the whole decision interval; this would limit the pre-processing (i.e., cleaning) one could apply. To truly protect users, it would be necessary to process their reactions in real-time to spot divergences with respect to the expected reactions and hence intervene. As a remark, this process must happen as soon as possible (or at least before the participants' final decision).

In the title of this paragraph, I put the HRI in braces because robots are not yet broadly available enough to develop an intervention solution only based on them. A first step would be to keep providing intervention with traditional warnings while exploring robot interventions in parallel. For this purpose, I studied two explicit interventions comparing an affective and a rational strategy in my research. However, other strategies should be explored, also comparing their effectiveness with respect to different effect mechanisms (i.e., a rational strategy could be effective against *reactance-based* attacks but useless with *liking-based* ones). Also, it would be worth exploring intervention based only on non-verbal cues (e.g., gazing, gestures, or facial expression), as literature shows they can be enough to elicit behavioral changes [349].

For instance, extending the experimental setup presented above, ICT departments could ask employees to work for a workday on a specific workstation, *pretending to study the support of a robot companion* (i.e., the Furhat). Furhat could be presented as an agent, moving and gazing toward the worker; also, it could offer some essential interaction and features as a virtual assistant (i.e., responding to users' requests). At the same time, the robot would process the physiological data stream, evaluate workers' response to the Social Engineering attacks prompted during the day, and intervene to prevent users' compliance. In this experimental context, it would be interesting to study implicit interventions (e.g., gaze) to drive workers' attention toward what is relevant (i.e., red flags), as in [350, 351].

Expand it to ordinary users The final and probably more challenging objective would be to deliver an effective solution to a broader population of users. Indeed, most current approaches focus on companies or at least workplaces. However, most attacks (e.g., massive phishing) target normal users who are not trained against Social Engineering and are not protected by enterprise-level ICT defenses. Firstly, more accessible material should be produced and spread to any user; second, affordable physiology-monitoring solutions based on standard hardware (i.e., RGB cameras and smartwatches) should be developed and made easily accessible. Such solutions would also help develop more generalizable scientific findings based on a broader population; and define baseline models for those individuals showing atypical reactions, behavior, or cognition.

Summing up, this dissertation provides several pieces of evidence that developing more *human-oriented* Social Engineering defense methods is feasible and effective. Similarly, robots proved to be an effective vector to influence users in making critical decisions. Given the high diffusion of Social Engineering threats, and the promising results of my research, I speculate these scientific findings will be the foundation of a novel and relevant research line.

Chapter 9

Epilogue

If you are reading this, either you are reviewing this thesis, are one of my supervisors, or are extremely curious about my research. I sincerely thank you for coming this far. I would like to drop the formality and speak directly to you, reader. I have one last insight that helped me in the last part of my research. Indeed, I greatly struggled to understand what my research was for beyond purely scientific purposes.

When I was in Waterloo (Canada), working on the Adventurer Robot Companion experiment, I went to the Canadian Clay and Glass Gallery - I suggest visiting it if you are in Ontario. The gallery hosted an exposition from the Canadian ceramic artist Walter Ostrom[1]. One of the exposed pots, basket-shaped and decorated with colorful flowers, was paired with a comment from Ostrom. The artist was experimenting with such basket shapes and flower decorations. One day, his wife Elaine took one of the vases and filled it with flowers. When Ostrom saw it, he complained that the flowers were competing with his decorations. Elaine replied that the decorations could never compete with the flowers. And here is the magic, quoting Walter Ostrom:

"What was so great about her response was that it gave me permission to decorate my brain out. If I was making utilitarian work, ultimately it is not my work, it becomes her work. I made the piece, but once she put her flower in there, that is her piece of art. It is a great thing: we are making something to enable other people to make their own art."

I believe the same applies to science. We put much effort into making outstanding research. However, we make a more or less little step toward a better and safer future, toward

[1] https://www.talesofaredclayrambler.com/episodes/394-walter-ostrom-on-the-questions-that-motivate-his-studio-practice

sensible and comprehensible robots able to understand and behave like humans, to be our companions. All we do is not for us. It is for others; to allow them to do their own research, their own art.

So, I hope you enjoyed my thesis; I had fun working on it. I hope it could be helpful for you or your research; maybe it just provided a different point of view you can take inspiration from. I am sure you will make great art out of it.

Publications

Below I list the scientific manuscripts I published during my Ph.D. period. All of them contributed to developing this dissertation.

Journal Articles

- **Pasquali D.**, Gonzalez-Billandon J., Aroyo A. M., Sandini G., Rea F., Sciutti A. (2022). *"Detecting lies is a child (robot)'s play: gaze-based lie detection in HRI"*. International Journal of Social Robotics; Online. [285]

- Aroyo A.M., **Pasquali D.**, Kothig A., Rea F., Sandini G., Sciutti A. (2021). *"Expectations Vs. Reality: Unreliability and Transparency in a Treasure Hunt Game With iCub"*. IEEE Robotics and Automation Letters, vol. 6, (no. 3), pp. 5681-5688 & 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). [251]

- Gonzalez-Billandon J., Aroyo A., Tonelli A., **Pasquali D.**, Sciutti A., Gori M., Sandini G., Rea F. (2019). *"Can a robot catch you lying? A machine learning system to detect lies during interactions"*. Frontiers in Robotics and AI, vol. 6, pp. 64. [133]

Conference Proceedings

- **Pasquali D.**, Sciutti A., Sandini G., Bencetti S., Rea F. (2022) *"Toward a Human-Oriented Social Engineering Defense System"*. Workshop: AI for Cybersecurity, Second CINI National Conference on Artificial Intelligence (Ital-IA) February 2022; Turin. [252]

- **Pasquali D.**, Gaggero D., Volpe G., Rea F., Sciutti A. (2021). *"Human vs. Robot Lie Detector: Better working as a team?"*. The Thirteenth International Conference on Social Robotics (ICSR), November 2021; Singapore [286]

- **Pasquali D.**, Sciutti A., Rea F. (2021). *"Toward enabling iCub to detect lies in everyday life"*. 3rd Italian Conference in Robotics and Intelligent Machines (I-RIM), October 2021; Rome [352]

- **Pasquali D.**, Gonzalez-Billandon J., Rea F., Sandini G., Sciutti A. (2021). *"Magic iCub: a humanoid robot autonomously catching your lies in a card game"*; 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI); March 2021; Cambridge (UK). [185]

- **Pasquali D.**, Aroyo A. M., Gonzalez-Billandon J., Rea F., Sandini G., Sciutti A. (2020). *"Do You See the Magic? An Autonomous Robot Magician Can Read Your Mind"*; 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI) - Workshop on Exploring Creative Content in Social Robotics; March 2020. [284]

- **Pasquali D.**, Aroyo A.M., Billandon-Gonzalez J., Rea F., Sandini G., Sciutti A. (2020). *"Your Eyes Never Lie: a Robot Magician Can Tell if You Are Lying"*; 15th ACM/IEEE International Conference on Human-Robot Interaction (Online) - Late Breaking Reports (LBR) - Award Candidate as Best LBR. [283]

- Aroyo A.M., **Pasquali D.**, Kothig A., Rea F., Sandini G., Sciutti A. (2020). *"Perceived differences between on-line and real robotic failures"*; IEEE RO-MAN 2020: Workshop on Trust, Acceptance and Social Cues in Human-Robot Interaction - SCRITA (online); [250]

Open Source Code and Data

Following the Open Science paradigm, I will polish and refine all the code developed for this dissertation and publicly release it on my GitHub page (github.com/dariopasquali) upon thesis publication. This would help the replication of my studies, help other researchers, and support future improvements. The four datasets I collected are available to be shared. Feel free to contact me [2] if you need them or have any questions about my research.

[2] dario.pasquali@iit.it, or dario.pasquali93@gmail.com

References

[1] B. Harrison, E. Svetieva, and A. Vishwanath, "Individual processing of phishing emails: How attention and elaboration protect against phishing," *Online Information Review*, vol. 40, no. 2, pp. 265–281, 2016. doi:10.1108/OIR-04-2015-0106.

[2] N. K. Lankton, D. Harrison Mcknight, and J. Tripp, "Technology, humanness, and trust: Rethinking trust in technology," *Journal of the Association for Information Systems*, vol. 16, no. 10, pp. 880–918, 2015. doi:10.17705/1jais.00411.

[3] Z. Wang, L. Sun, and H. Zhu, "Defining Social Engineering in Cybersecurity," *IEEE Access*, vol. 8, pp. 85094–85115, 2020. doi:10.1109/ACCESS.2020.2992807.

[4] Tessian, "15 Examples of Real Social Engineering Attacks - Updated 2022," 2022. url: https://www.tessian.com/blog/examples-of-social-engineering-attacks/.

[5] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: Bodies and minds moving together," *Trends in Cognitive Sciences*, vol. 10, no. 2, pp. 70–76, 2006. doi:10.1016/j.tics.2005.12.009.

[6] C. Hadnagy, "Social Engineering: The Art of Human Hacking," *The Art of Human Hacking*, vol. 3, no. 3, p. 408, 2010. doi:10.1504/ijipsi.2018.10013213.

[7] K. D. Mitnick and W. L. Simon, *The Art of Deception: Controlling the Human Element of Security (Google eBook)*. Wiley Pub, 2001. ISBN 0471432288.

[8] F. Mouton, M. M. Malan, L. Leenen, and H. S. Venter, "Social engineering attack framework," *2014 Information Security for South Africa - Proceedings of the ISSA 2014 Conference*, 2014. doi:10.1109/ISSA.2014.6950510.

[9] B. M. DePaulo, B. E. Malone, J. J. Lindsay, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," 2003. doi:10.1037/0033-2909.129.1.74.

[10] T. Fong, C. Thorpe, and C. Baur, "Collaboration, Dialogue, Human-Robot Interaction," in *Robotics Research*, pp. 255–266, Springer, Berlin, Heidelberg, aug 2007. doi:10.1007/3-540-36460-9_17.

[11] M. A. Goodrich and A. C. Schultz, "Human-Robot Interaction: A Survey," *Foundations and Trends® in Human-Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007. doi:10.1561/1100000005.

[12] D. Westerman, A. P. Edwards, C. Edwards, Z. Luo, and P. R. Spence, "I-It, I-Thou, I-Robot: The Perceived Humanness of AI in Human-Machine Communication," *Communication Studies*, vol. 71, pp. 393–408, may 2020. doi:10.1080/10510974.2020.1749683.

[13] S. Booth, J. Tompkin, H. Pfister, J. Waldo, K. Gajos, and R. Nagpal, "Piggybacking Robots: Human-Robot Overtrust in University Dormitory Security," 2017. doi:dx.doi.org/10.1145/2909824.3020211.

[14] 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 Robotics and Automation Letters*, vol. 3, no. 4, pp. 3701–3708, 2018. doi:10.1109/LRA.2018.2856272.

[15] A. F. Abate, C. Bisogni, L. Cascone, A. Castiglione, G. Costabile, and I. Mercuri, "Social Robot Interactions for Social Engineering: Opportunities and Open Issues," in *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pp. 539–547, IEEE, aug 2020. doi:10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00097.

[16] P. Barosz, G. Gołda, and A. Kampa, "Efficiency analysis of manufacturing line with industrial robots and human operators," *Applied Sciences (Switzerland)*, vol. 10, no. 8, 2020. doi:10.3390/APP10082862.

[17] D. L. Gogoshin, "Robot Responsibility and Moral Community," *Frontiers in Robotics and AI*, vol. 8, no. November, pp. 1–13, 2021. doi:10.3389/frobt.2021.768092.

[18] S. Shah, "Seven biases to avoid in qualitative research," *Www.Editage.Com*, pp. 1–4, 2019. url: https://www.editage.com/insights/7-biases-to-avoid-in-qualitative-research.

[19] C. J. Pannucci and E. G. Wilkins, "Identifying and Avoiding Bias in Research," *Plastic and Reconstructive Surgery*, vol. 126, pp. 619–625, aug 2010. doi:10.1097/PRS.0b013e3181de24bc.

[20] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The benefits of interactions with physically present robots over video-displayed agents," *International Journal of Social Robotics*, vol. 3, pp. 41–52, jan 2011. doi:10.1007/s12369-010-0082-7.

[21] K. S. Haring, K. M. Satterfield, C. C. Tossell, E. J. de Visser, J. R. Lyons, V. F. Mancuso, V. S. Finomore, and G. J. Funke, "Robot Authority in Human-Robot Teaming: Effects of Human-Likeness and Physical Embodiment on Compliance," *Frontiers in Psychology*, vol. 12, no. May, pp. 1–16, 2021. doi:10.3389/fpsyg.2021.625713.

[22] Z. Wang, H. Zhu, and L. Sun, "Social engineering in cybersecurity: Effect mechanisms, human vulnerabilities and attack methods," *IEEE Access*, vol. 9, pp. 11895–11910, 2021. doi:10.1109/ACCESS.2021.3051633.

[23] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," *Frontiers in Computer Science*, vol. 3, no. March, pp. 1–23, 2021. doi:10.3389/fcomp.2021.563060.

[24] S. Gupta, A. Singhal, and A. Kapoor, "A literature survey on social engineering attacks: Phishing attack," *Proceeding - IEEE International Conference on Computing, Communication and Automation, ICCCA 2016*, pp. 537–540, 2017. doi:10.1109/CCAA.2016.7813778.

[25] M. Butavicius, K. Parsons, M. Pattinson, and A. McCormac, "Breaching the Human Firewall: Social engineering in Phishing and Spear-Phishing Emails," *ACIS 2015 Proceedings - 26th Australasian Conference on Information Systems*, may 2016. url: http://arxiv.org/abs/1606.00887.

[26] E. Ulqinaku, H. Assal, A. R. Abdou, S. Chiasson, and S. Capkun, "Is real-time phishing eliminated with FIDO? Social engineering downgrade attacks against FIDO protocols," *Proceedings of the 30th USENIX Security Symposium*, pp. 3811–3828, 2021. url: https://www.usenix.org/conference/usenixsecurity21/presentation/ulqinaku.

[27] S. Stasiukonis, "Social engineering, the USB way," *Dark Reading*, vol. 7, p. 95556, 2006. url: https://www.darkreading.com/perimeter/social-engineering-the-usb-wayhttp://tonydye.typepad.com/main/files/HO05-DarkReading.doc.

[28] F. Mouton, A. Nottingham, L. Leenen, and H. S. Venter, "Finite state machine for the social engineering attack detection model: SEADM," *SAIEE Africa Research Journal*, vol. 109, no. 2, pp. 133–147, 2018. doi:10.1109/ISSA.2017.8251781.

[29] W. Fan, K. Lwakatare, and R. Rong, "Social Engineering: I-E based Model of Human Weakness for Attack and Defense Investigations," *International Journal of Computer Network and Information Security*, vol. 9, no. 1, pp. 1–11, 2017. doi:10.5815/ijcnis.2017.01.01.

[30] J. Stewart and M. Dawson, "How the modification of personality traits leave one vulnerable to manipulation in social engineering," *International Journal of Information Privacy, Security and Integrity*, vol. 3, no. 3, p. 187, 2018. doi:10.1504/IJIPSI.2018.10013213.

[31] R. R. Mccrae and O. P. John, "The five-factor model: issues and applications.," *Journal of personality*, vol. 60, no. 2, pp. 175–532, 1992. url: http://www.ncbi.nlm.nih.gov/pubmed/1635040.

[32] P. Schaab, K. Beckers, and S. Pape, "A systematic gap analysis of social engineering defence mechanisms considering social psychology," *Proceedings of the 10th International Symposium on Human Aspects of Information Security and Assurance, HAISA 2016*, no. Haisa, pp. 241–251, 2016. url: https://www.cscan.org/openaccess/?paperid=301.

[33] E. D. Frangopoulos, M. M. Eloff, and L. M. Venter, "Psychosocial risks: Can their effects on the security of information systems really be ignored?," *Proceedings of the 6th International Symposium on Human Aspects of Information Security and Assurance, HAISA 2012*, pp. 52–63, 2012.

[34] R. Gulati, "InfoSec Reading Room The Threat of Social Engineering and Your Defense Against It," *Information Security*, no. Security 401, pp. 1–15, 2003. url: https://www.sans.org/reading-room/whitepapers/engineering/threat-social-engineering-defense-1232.

[35] M. Nohlberg, *Securing Information Assets: Understanding, Measuring and Protecting against Social Engineering Attacks*. No. 09, 2008. ISBN 9789171557865.

[36] M. Junger, L. Montoya, and F. J. Overink, "Priming and warnings are not effective to prevent social engineering attacks," *Computers in Human Behavior*, vol. 66, pp. 75–87, 2017. doi:10.1016/j.chb.2016.09.012.

[37] R. E. Petty and J. T. Cacioppo, "The elaboration likelihood model of persuasion," *Advances in Experimental Social Psychology*, vol. 19, no. C, pp. 123–205, 1986. doi:10.1016/S0065-2601(08)60214-2.

[38] R. Cialdini, *Influence: The Psychology of Persuasion*. No. 14, HarperCollins, 2001. ISBN 9780061241895.

[39] M. F. Abrahams and R. A. Bell, "Encouraging Charitable Contributions: An Examination of Three Models of Door-in-the-Face Compliance," *Communication Research*, vol. 21, pp. 131–153, jun 1994. doi:10.1177/009365094021002001.

[40] C. Happ, A. Melzer, and G. Steffgen, "Trick with treat - Reciprocity increases the willingness to communicate personal data," *Computers in Human Behavior*, vol. 61, pp. 372–377, 2016. doi:10.1016/j.chb.2016.03.026.

[41] J. L. Freedman and S. C. Fraser, "Compliance without pressure: The foot-in-the-door technique," *Journal of Personality and Social Psychology*, vol. 4, pp. 195–202, aug 1966. doi:10.1037/h0023552.

[42] F. Stajano and P. Wilson, "Understanding scam victims: Seven principles for systems security," *Communications of the ACM*, vol. 54, no. 3, pp. 70–75, 2011. doi:10.1145/1897852.1897872.

[43] S. Sabouni, A. Cullen, and L. Armitage, "A preliminary radicalisation framework based on social engineering techniques," *2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment, Cyber SA 2017*, 2017. doi:10.1109/CyberSA.2017.8073406.

[44] 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 *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 35–42, Institute of Electrical and Electronics Engineers Inc., nov 2018. doi:10.1109/ROMAN.2018.8525744.

[45] S. Martin, N. Goldstein, and R. B. Cialdini, *The small BIG: Small Changes that Spark Big Influence*. Profile, 2014. ISBN 1782830758.

[46] N. Guéguen and C. Jacob, "Solicitation by e-mail and solicitor's status: A field study of social influence on the web," *Cyberpsychology and Behavior*, vol. 5, no. 4, pp. 377–383, 2002. doi:10.1089/109493102760275626.

[47] P. Tiwari, *Exploring Phishing Susceptibility Attributable To Authority, Urgency, Risk Perception and Human Factors*. PhD thesis, 2020. doi:https://doi.org/10.25394/PGS.12739592.v1.

[48] J. W. H. Bullée, L. Montoya, W. Pieters, M. Junger, and P. H. Hartel, "The persuasion and security awareness experiment: reducing the success of social engineering attacks," *Journal of Experimental Criminology*, vol. 11, no. 1, pp. 97–115, 2015. doi:10.1007/s11292-014-9222-7.

[49] J. W. Brehm and E. P. Torrance, "A Theory of Psychological Reactance," *The American Journal of Psychology*, vol. 81, p. 133, mar 1968. doi:10.2307/1420824.

[50] M. Lynn, "Scarcity's Enhancement of Desirability: The Role of Naive Economic Theories," *Basic and Applied Social Psychology*, vol. 13, pp. 67–78, mar 1992. doi:10.1207/s15324834basp1301_6.

[51] Gamasutra, "GameStop: Wii Shortage 'Intentional', PS3 Euro Launch 'Good, Not Great'," 2007. url: https://www.gamasutra.com/php-bin/news_index.php?story=13297.

[52] S. Tewari and A. Gupta, "Pre-suasion: a revolutionary way to influence and persuade," *Journal of Marketing Communications*, vol. 26, pp. 454–456, may 2020. doi:10.1080/13527266.2018.1504811.

[53] R. J. Davies and I. Osamu, *The Japanese mind: Understanding contemporary Japanese culture*. 2002.

[54] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," *Artificial Intelligence Review*, vol. 29, pp. 63–92, jul 2008. doi:10.1007/s10462-009-9109-6.

[55] M. Lansley, F. Mouton, S. Kapetanakis, and N. Polatidis, "SEADEr++: Social engineering attack detection in online environments using machine learning," *Journal of Information and Telecommunication*, vol. 4, no. 3, pp. 346–362, 2020. doi:10.1080/24751839.2020.1747001.

[56] Y. Lee, J. Saxe, and R. Harang, "CATBERT: Context-Aware Tiny BERT for Detecting Social Engineering Emails," oct 2020. doi:https://doi.org/10.48550/arXiv.2010.03484.

[57] G. L'Huillier, A. Hevia, R. Weber, and S. Ríos, "Latent semantic analysis and keyword extraction for phishing classification," *ISI 2010 - 2010 IEEE International Conference on Intelligence and Security Informatics: Public Safety and Security*, pp. 129–131, 2010. doi:10.1109/ISI.2010.5484762.

[58] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," *Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018*, vol. 2018-Janua, pp. 300–301, 2018. doi:10.1109/ICSC.2018.00056.

[59] R. Verma, N. Shashidhar, and N. Hossain, "Detecting Phishing Emails the Natural Language Way," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7459 LNCS, pp. 824–841, Springer, Berlin, Heidelberg, 2012. doi:10.1007/978-3-642-33167-1_47. ISBN 9783642331664.

[60] R. Wash, "How Experts Detect Phishing Scam Emails," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, 2020. doi:10.1145/3415231.

[61] J. Rashid, T. Mahmood, M. W. Nisar, and T. Nazir, "Phishing Detection Using Machine Learning Technique," *Proceedings - 2020 1st International Conference of Smart Systems and Emerging Technologies, SMART-TECH 2020*, pp. 43–46, 2020. doi:10.1109/SMART-TECH49988.2020.00026.

[62] E. Şen and G. Tuna, "The Anatomy of Phishing Attacks and the Detection and Prevention of Fake Domain Names," in *https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-6684-3380-5.ch025*, pp. 583–605, IGI Global, jan 2022. doi:10.4018/978-1-6684-3380-5.CH025.

[63] S. Shivangi, P. Debnath, K. Saieevan, and D. Annapurna, "Chrome Extension For Malicious URLs detection in Social Media Applications Using Artificial Neural Networks And Long Short Term Memory Networks," *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1993–1997, 2018. doi:10.1109/ICACCI.2018.8554647.

[64] H. Shirazi, K. Haefner, and I. Ray, "Fresh-Phish: A framework for auto-detection of phishing websites," *Proceedings - 2017 IEEE International Conference on Information Reuse and Integration, IRI 2017*, vol. 2017-Janua, pp. 137–143, 2017. doi:10.1109/IRI.2017.40.

[65] T. Kelley, M. J. Amon, and B. I. Bertenthal, "Statistical models for predicting threat detection from human behavior," *Frontiers in Psychology*, vol. 9, no. APR, pp. 1–17, 2018. doi:10.3389/fpsyg.2018.00466.

[66] V. Shreeram, M. Suban, P. Shanthi, and K. Manjula, "Anti-phishing detection of phishing attacks using genetic algorithm," in *2010 IEEE International Conference on Communication Control and Computing Technologies, ICCCCT 2010*, pp. 447–450, 2010. doi:10.1109/ICCCCT.2010.5670593.

[67] P. Vadrevu and R. Perdisci, "What you see is not what you get: Discovering and tracking social engineering attack campaigns," *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, pp. 308–321, 2019. doi:10.1145/3355369.3355600.

[68] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," *Communications of the ACM*, vol. 50, pp. 94–100, oct 2007. doi:10.1145/1290958.1290968.

[69] O. Jaafor and B. Birregah, "Multi-layered graph-based model for social engineering vulnerability assessment," *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*, pp. 1480–1488, 2015. doi:10.1145/2808797.2808899.

[70] A. Algarni, Y. Xu, and T. Chan, "Social engineering in social networking sites: The art of impersonation," in *Proceedings - 2014 IEEE International Conference on Services Computing, SCC 2014*, pp. 797–804, 2014. doi:10.1109/SCC.2014.108.

[71] S. Gheewala and R. Patel, "Machine Learning Based Twitter Spam Account Detection: A Review," in *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*, no. Iccmc, pp. 79–84, IEEE, feb 2018. doi:10.1109/ICCMC.2018.8487992.

[72] N. Agarwal, S. Al-Khateeb, R. Galeano, and R. Goolsby, "Examining the use of botnets and their evolution in propaganda dissemination.," *Defence Strategic Communications*, vol. 2, pp. 87–112, aug 2017. doi:10.30966/2018.riga.2.4.

[73] J. P. McIntire, L. K. McIntire, and P. R. Havig, "Methods for chatbot detection in distributed text-based communications," in *2010 International Symposium on Collaborative Technologies and Systems, CTS 2010*, pp. 463–472, IEEE, 2010. doi:10.1109/CTS.2010.5478478.

[74] N. Tsinganos, P. Fouliras, G. Sakellariou, and I. Mavridis, "Towards an automated recognition system for chat-based social engineering attacks in enterprise environments," *ACM International Conference Proceeding Series*, 2018. doi:10.1145/3230833.3233277.

[75] J. Zhang, R. Zhang, Y. Zhang, and G. Yan, "The rise of social botnets: Attacks and countermeasures," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 6, pp. 1068–1082, 2018. doi:10.1109/TDSC.2016.2641441.

[76] S. Li, X. Yun, Z. Hao, X. Cui, and Y. Wang, "A propagation model for social engineering botnets in social networks," *Parallel and Distributed Computing, Applications and Technologies, PDCAT Proceedings*, pp. 423–426, 2011. doi:10.1109/PDCAT.2011.8.

[77] M. R. Faghani and U. T. Nguyen, "Socellbot: A new botnet design to infect smartphones via online social networking," in *2012 25th IEEE Canadian Conference on Electrical and Computer Engineering: Vision for a Greener Future, CCECE 2012*, pp. 1–5, IEEE, 2012. doi:10.1109/CCECE.2012.6334962.

[78] Y. Xing, H. Shu, H. Zhao, D. Li, and L. Guo, "Survey on Botnet Detection Techniques: Classification, Methods, and Evaluation," 2021. doi:10.1155/2021/6640499.

[79] N. E. Díaz Ferreyra, E. Aímeur, H. Hage, M. Heisel, and C. G. Van Hoogstraten, "Persuasion meets AI: Ethical considerations for the design of social engineering countermeasures," *IC3K 2020 - Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, vol. 3, pp. 204–211, 2020. doi:10.5220/0010142402040211.

[80] S. M. Albladi and G. R. S. Weir, "Predicting individuals' vulnerability to social engineering in social networks," *Cybersecurity*, vol. 3, p. 7, dec 2020. doi:10.1186/s42400-020-00047-5.

[81] V. Kumar and D. Sinha, "A robust intelligent zero-day cyber-attack detection technique," *Complex and Intelligent Systems*, vol. 7, pp. 2211–2234, may 2021. doi:10.1007/s40747-021-00396-9.

[82] H. Aldawood and G. Skinner, "Educating and Raising Awareness on Cyber Security Social Engineering: A Literature Review," in *Proceedings of 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering, TALE 2018*, pp. 62–68, Institute of Electrical and Electronics Engineers Inc., jan 2019. doi:10.1109/TALE.2018.8615162.

[83] J.-p. O. Kok, *Weaponize personnel against social engineering attacks via gamified active inoculation*. PhD thesis, Delft University of Technology, 2021. url: http://resolver.tudelft.nl/uuid:da42f515-3b8a-4706-9c60-2d7dbb910861.

[84] R. XU, *Gamification Platform for Social Engineering Training and Awareness*. PhD thesis, University of Windsor, 2021.

[85] F. Abu-Amara, R. Almansoori, S. Alharbi, M. Alharbi, and A. Alshehhi, "A novel SETA-based gamification framework to raise cybersecurity awareness," *International Journal of Information Technology (Singapore)*, vol. 13, no. 6, pp. 2371–2380, 2021. doi:10.1007/s41870-021-00760-5.

[86] Z. M. Hakim, N. C. Ebner, D. S. Oliveira, S. J. Getz, B. E. Levin, T. Lin, K. Lloyd, V. T. Lai, M. D. Grilli, and R. C. Wilson, "The Phishing Email Suspicion Test (PEST) a lab-based task for evaluating the cognitive mechanisms of phishing detection," *Behavior Research Methods*, vol. 53, pp. 1342–1352, jun 2021. doi:10.3758/S13428-020-01495-0/FIGURES/5.

[87] K. Beckers and S. Pape, "A Serious Game for Eliciting Social Engineering Security Requirements," in *Proceedings - 2016 IEEE 24th International Requirements Engineering Conference, RE 2016*, pp. 16–25, Institute of Electrical and Electronics Engineers Inc., dec 2016. doi:10.1109/RE.2016.39.

[88] A. Harilal, F. Toffalini, I. Homoliak, J. Castellanos, J. Guarnizo, S. Mondal, and M. Ochoa, "The Wolf of SUTD (TWOS): A dataset of malicious insider threat behavior based on a gamified competition," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 9, no. 1, pp. 54–85, 2018. doi:10.22667/JOWUA.2018.03.31.054.

[89] C. F. Barreto and C. Franca, "Gamification in Software Engineering: A literature Review," in *Proceedings - 2021 IEEE/ACM 13th International Workshop on Cooperative and Human Aspects of Software Engineering, CHASE 2021*, pp. 105–108, IEEE, may 2021. doi:10.1109/CHASE52884.2021.00020.

[90] D. Alharthi and A. Regan, "A Literature Survey and Analysis on Social Engineering Defense Mechanisms and Infosec Policies," *International Journal of Network Security and Its Applications*, vol. 13, no. 2, pp. 41–61, 2021. doi:10.5121/ijnsa.2021.13204.

[91] J. W. Bullee and M. Junger, "How effective are social engineering interventions? A meta-analysis," *Information and Computer Security*, vol. 28, no. 5, pp. 801–830, 2020. doi:10.1108/ICS-07-2019-0078.

[92] M. Workman, "Gaining access with social engineering: An empirical study of the threat," *Information Systems Security*, vol. 16, no. 6, pp. 315–331, 2007. doi:10.1080/10658980701788165.

[93] V. Zolotarev, A. Arkhipova, N. Parotkin, and A. Lvova, "Strategies of social engineering attacks on information resources of gamified online education projects," *CEUR Workshop Proceedings*, vol. 2861, pp. 386–391, 2020.

[94] R. Heartfield, G. Loukas, and D. Gan, "You Are Probably Not the Weakest Link: Towards Practical Prediction of Susceptibility to Semantic Social Engineering Attacks," *IEEE Access*, vol. 4, pp. 6910–6928, 2016. doi:10.1109/ACCESS.2016.2616285.

[95] M. V. Abramov and A. A. Azarov, "Social engineering attack modeling with the use of Bayesian networks," in *2016 XIX IEEE International Conference on Soft Computing and Measurements (SCM)*, no. 16, pp. 58–60, IEEE, may 2016. doi:10.1109/SCM.2016.7519682.

[96] L. M. Stuart, G. Park, J. M. Talor, and V. Raskin, "On identifying phishing emails: Uncertainty in machine and human judgment," *2014 IEEE Conference on Norbert Wiener in the 21st Century: Driving Technology's Future, 21CW 2014 - Incorporating the Proceedings of the 2014 North American Fuzzy Information Processing Society Conference, NAFIPS 2014, Conference Proceedings*, 2014. doi:10.1109/NORBERT.2014.6893870.

[97] S. Venkatesha, K. R. Reddy, and B. R. Chandavarkar, "Social Engineering Attacks During the COVID-19 Pandemic," *SN Computer Science*, vol. 2, no. 2, p. 78, 2021. doi:10.1007/s42979-020-00443-1.

[98] B. Schneier, *Secrets and Lies: Digital Security in a Networked World*. John Wiley Sons, Inc., 2000. ISBN 0-471-25311-1.

[99] G. Stringhini and O. Thonnard, "That ain't you: Blocking spearphishing through behavioral modelling," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9148, pp. 78–97, Springer, Cham, 2015. doi:10.1007/978-3-319-20550-2_5.

[100] M. Vielberth, F. Menges, and G. Pernul, "Human-as-a-security-sensor for harvesting threat intelligence," *Cybersecurity*, vol. 2, pp. 1–15, dec 2019. doi:10.1186/s42400-019-0040-0.

[101] R. Heartfield, G. Loukas, and D. Gan, "An eye for deception: A case study in utilizing the human-as-a-security-sensor paradigm to detect zero-day semantic social engineering attacks," *Proceedings - 2017 15th IEEE/ACIS International Conference on Software Engineering Research, Management and Applications, SERA 2017*, pp. 371–378, 2017. doi:10.1109/SERA.2017.7965754.

[102] N. Stembert, A. Padmos, M. S. Bargh, S. Choenni, and F. Jansen, "A Study of Preventing Email (Spear) Phishing by Enabling Human Intelligence," in *Proceedings - 2015 European Intelligence and Security Informatics Conference, EISIC 2015*, pp. 113–120, Institute of Electrical and Electronics Engineers Inc., jan 2016. doi:10.1109/EISIC.2015.38.

[103] L. Malisa, K. Kostiainen, and S. Capkun, "Detecting mobile application spoofing attacks by leveraging user visual similarity perception," in *CODASPY 2017 - Proceedings of the 7th ACM Conference on Data and Application Security and*

Privacy, pp. 289–300, Association for Computing Machinery, Inc, mar 2017. doi:10.1145/3029806.3029819.

[104] L. M. Sacheli, S. M. Aglioti, and M. Candidi, "Social cues to joint actions: The role of shared goals," *Frontiers in Psychology*, vol. 6, no. JUL, p. 1034, 2015. doi:10.3389/fpsyg.2015.01034.

[105] S. J. Gould and E. S. Vrba, "Exaptation—a Missing Term in the Science of Form," *Paleobiology*, vol. 1, no. N/A, pp. 4–15, 1982. doi:10.1017/S0094837300004310.

[106] J. L. Andreassi, *Psychophysiology: Human behavior and physiological response*. Psychology Press, may 2010. doi:10.4324/9780203880340. ISBN 9781135613082.

[107] J. T. Cacioppo and L. G. Tassinary, "Inferring Psychological Significance from Physiological Signals," *American Psychologist*, vol. 45, no. 1, pp. 16–28, 1990. doi:10.1037/0003-066X.45.1.16.

[108] M. Sumpf, S. Jentschke, and S. Koelsch, "Effects of aesthetic chills on a cardiac signature of emotionality," *PLoS ONE*, vol. 10, p. e0130117, jun 2015. doi:10.1371/journal.pone.0130117.

[109] M. E. Dawson, A. M. Schell, and C. G. Courtney, "The Skin Conductance Response, Anticipation, and Decision-Making," *Journal of Neuroscience, Psychology, and Economics*, vol. 4, no. 2, pp. 111–116, 2011. doi:10.1037/a0022619.

[110] C. Wangwiwattana, X. Ding, and E. C. Larson, "PupilNet, Measuring Task Evoked Pupillary Response using Commodity RGB Tablet Cameras," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, pp. 1–26, jan 2018. doi:10.1145/3161164.

[111] S. Rafiqi, C. Wangwiwattana, J. Kim, E. Fernandez, S. Nair, and E. C. Larson, "PupilWare," in *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, no. August, (New York, NY, USA), pp. 1–8, ACM, jul 2015. doi:10.1145/2769493.2769506.

[112] S. Eivazi, T. Santini, A. Keshavarzi, T. Kübler, and A. Mazzei, "Improving real-time CNN-based pupil detection through domain-specific data augmentation," *Eye Tracking Research and Applications Symposium (ETRA)*, 2019. doi:10.1145/3314111.3319914.

[113] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-Adaptive Matrix Completion for Heart Rate Estimation from Face Videos under Realistic Conditions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 2396–2404, IEEE Computer Society, dec 2016. doi:10.1109/CVPR.2016.263.

[114] G. Bieber, N. Antony, and M. Haescher, "Touchless heart rate Recognition by Robots to support natural Human-Robot Communication," *ACM International Conference Proceeding Series*, pp. 415–420, 2018. doi:10.1145/3197768.3203181.

[115] Empatica, "E4 wristband | Real-time physiological signals | Wearable PPG, EDA, Temperature, Motion sensors," 2020. url: https://www.empatica.com/en-eu/research/e4/.

[116] Mionix, "Mionix Naos QG," 2021. url: https://mionix.net/products/naos-qg.

[117] D. Cró Rodrigues, *Physiopad Development of a non-invasive game controller toolkit to study physiological responses for Game User Research*. PhD thesis, Universidade da Madeira, 2018. url: https://digituma.uma.pt/bitstream/10400.13/1998/1/MestradoDiogoRodrigues.pdf.

[118] J. Beatty and B. Lucero-Wagoner, "The pupillary system," *Handbook of psychophysiology 2*, 2000. url: https://psycnet.apa.org/record/2000-03927-005.

[119] A. Szulewski, N. Roth, and D. Howes, "The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise," *Academic Medicine*, vol. 90, pp. 981–987, jul 2015. doi:10.1097/ACM.0000000000000677.

[120] 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, pp. 205–211, mar 2001. doi:10.1017/S0048577201990717.

[121] A. Bran and D. C. Vaidis, "Assessing risk-taking: what to measure and how to measure it," *Journal of Risk Research*, vol. 23, pp. 490–503, apr 2020. doi:10.1080/13669877.2019.1591489.

[122] P. Horvath and M. Zuckerman, "Sensation seeking, risk appraisal, and risky behavior," *Personality and Individual Differences*, vol. 14, pp. 41–52, jan 1993. doi:10.1016/0191-8869(93)90173-Z.

[123] Bible, "Matthew 6:22-23 NIV - "The eye is the lamp of the body. If - Bible Gateway," 2011. url: https://www.biblegateway.com/passage/?search=Matthew6%3A22-23&version=NIV.

[124] M. Dalmaso, L. Castelli, and G. Galfano, "Social modulators of gaze-mediated orienting of attention: A review," oct 2020. doi:10.3758/s13423-020-01730-x.

[125] A. Frischen, A. P. Bayliss, and S. P. Tipper, "Gaze Cueing of Attention: Visual Attention, Social Cognition, and Individual Differences," *Psychological Bulletin*, vol. 133, pp. 694–724, jul 2007. doi:10.1037/0033-2909.133.4.694.

[126] J. G. May, R. S. Kennedy, M. C. Williams, W. P. Dunlap, and J. R. Brannan, "Eye movement indices of mental workload," *Acta Psychologica*, vol. 75, no. 1, pp. 75–89, 1990. doi:10.1016/0001-6918(90)90067-P.

[127] M. Nakayama and Y. Shimizu, "Frequency analysis of task evoked pupillary response and eye-movement," in *Eye Tracking Research and Applications Symposium (ETRA)*, (New York, New York, USA), pp. 71–76, ACM Press, 2004. doi:10.1145/968363.968381.

[128] K. F. Van Orden, W. Limbert, S. Makeig, and T. P. Jung, "Eye activity correlates of workload during a visuospatial memory task," *Human Factors*, vol. 43, no. 1, pp. 111–121, 2001. doi:10.1518/001872001775992570.

[129] B. C. Goldwater, "Psychological significance of pupillary movements," *Psychological Bulletin*, vol. 77, pp. 340–355, may 1972. doi:10.1037/h0032456.

[130] J. A. Stern, L. C. Walrath, and R. Goldstein, "The Endogenous Eyeblink," *Psychophysiology*, vol. 21, pp. 22–33, jan 1984. doi:10.1111/j.1469-8986.1984.tb02312.x.

[131] C. R. Honts, D. C. Raskin, and J. C. Kircher, "Mental and Physical Countermeasures Reduce the Accuracy of Polygraph Tests," *Journal of Applied Psychology*, vol. 79, pp. 252–259, apr 1994. doi:10.1037/0021-9010.79.2.252.

[132] S. M. Kassin, "On the psychology of confessions: Does innocence put innocents at risk?," *American Psychologist*, vol. 60, pp. 215–228, apr 2005. doi:10.1037/0003-066X.60.3.215.

[133] J. Gonzalez-Billandon, A. M. Aroyo, A. Tonelli, D. Pasquali, A. Sciutti, M. Gori, G. Sandini, and F. Rea, "Can a Robot Catch You Lying? A Machine Learning System to Detect Lies During Interactions," *Frontiers in Robotics and AI*, vol. 6, jul 2019. doi:10.3389/frobt.2019.00064.

[134] A. Aroyo, J. Gonzalez-Billandon, A. Tonelli, A. Sciutti, M. Gori, G. Sandini, and F. Rea, "Can a Humanoid Robot Spot a Liar?," in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, pp. 1045–1052, IEEE, nov 2018. doi:10.1109/HUMANOIDS.2018.8624992.

[135] M. Gamer, "Detecting of deception and concealed information using neuroimaging techniques," in *HRI'20 Human-Robot Interaction*, (Boulder), pp. 90–113, 2011. doi:10.1017/CBO9780511975196.006.

[136] B. A. Rajoub and R. Zwiggelaar, "Thermal Facial Analysis for Deception Detection," *IEEE Transactions on Information Forensics and Security*, vol. 9, pp. 1015–1023, jun 2014. doi:10.1109/TIFS.2014.2317309.

[137] C. Y. Ma, M. H. Chen, Z. Kira, and G. AlRegib, "TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition," *Signal Processing: Image Communication*, vol. 71, pp. 76–87, mar 2019. doi:10.1016/j.image.2018.09.003.

[138] V. Karpova, P. Popenova, N. Glebko, V. Lyashenko, and O. Perepelkina, ""was It You Who Stole 500 Rubles?" - The Multimodal Deception Detection," in *ICMI 2020 Companion - Companion Publication of the 2020 International Conference on Multimodal Interaction*, pp. 112–119, 2020. doi:10.1145/3395035.3425638.

[139] X. L. Chen, S. Ita Levitan, M. Levine, M. Mandic, and J. Hirschberg, "Acoustic-Prosodic and Lexical Cues to Deception and Trust: Deciphering How People Detect Lies," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 199–214, 2020. doi:10.1162/tacl_a_00311.

[140] M. I. Ahmad, J. Bernotat, K. Lohan, and F. Eyssel, "Trust and Cognitive Load During Human-Robot Interaction," in *AAAI Symposium on Artificial Intelligence for Human-Robot Interaction*, 2019. url: https://arxiv.org/abs/1909.05160v1.

[141] J. Klingner, R. Kumar, and P. Hanrahan, "Measuring the task-evoked pupillary response with a remote eye tracker," in *Proceedings of the 2008 symposium on Eye tracking research and applications - ETRA '08*, no. May, (New York, New York, USA), p. 69, Stanford University, ACM Press, 2008. doi:10.1145/1344471.1344489.

[142] G. Hossain and M. Yeasin, "Understanding effects of cognitive load from pupillary responses using hilbert analytic phase," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 381–386, 2014. doi:10.1109/CVPRW.2014.62.

[143] J. C. Van Slooten, S. Jahfari, T. Knapen, and J. Theeuwes, "How pupil responses track value-based decision-making during and after reinforcement learning," *PLoS Computational Biology*, vol. 14, no. 11, pp. 1–24, 2018. doi:10.1371/journal.pcbi.1006632.

[144] E. Yechiam and A. Telpaz, "To take risk is to face loss: A tonic pupillometry study," *Frontiers in Psychology*, vol. 2, p. 344, nov 2011. doi:10.3389/fpsyg.2011.00344.

[145] A. E. Urai, A. Braun, and T. H. Donner, "Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias," *Nature Communications*, vol. 8, 2017. doi:10.1038/ncomms14637.

[146] K. Preuschoff, B. M. 't Hart, and W. Einhäuser, "Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making," *Frontiers in Neuroscience*, vol. 5, pp. 1–12, sep 2011. doi:10.3389/fnins.2011.00115.

[147] C. Lavín, R. S. Martín, and E. R. Jubal, "Pupil dilation signals uncertainty and surprise in a learning gambling task," *Frontiers in Behavioral Neuroscience*, vol. 7, no. JAN, pp. 1–8, 2014. doi:10.3389/fnbeh.2013.00218.

[148] M. Causse, B. Baracat, J. Pastor, and F. Dehais, "Reward and uncertainty favor risky decision-making in pilots: Evidence from cardiovascular and oculometric measurements," *Applied Psychophysiology Biofeedback*, vol. 36, no. 4, pp. 231–242, 2011. doi:10.1007/s10484-011-9163-0.

[149] J. J. J. Braithwaite, D. Derrick, G. Watson, R. Jones, M. Rowe, D. Watson, J. Robert, and R. Mickey, "A Guide for Analysing Electrodermal Activity (EDA) and Skin Conductance Responses (SCRs) for Psychological Experiments," ..., pp. 1–42, 2013. url: http://www.bhamlive.bham.ac.uk/Documents/college-les/psych/saal/guide-electrodermal-activity.pdf%5Cnhttp://www.birmingham.ac.uk/documents/college-les/psych/saal/guide-electrodermal-activity.pdf%0Ahttps://www.birmingham.ac.uk/Documents/college-les/psych/sa.

[150] B. Figner and R. O. Murphy, "Using skin conductance in judgment and decision making research," in *A handbook of processtracing methods for decision research*, pp. 1–33, sychology Press., 2011. url: http://books.google.com/books?hl=en&lr=&id=DBx5AgAAQBAJ&oi=fnd&pg=PA163&dq=Using+skin+conductance+in+judgment+and+decision+making+research&ots=0untpIlmBa&sig=RhcZ68GTW8oalyarjYV5WAN6TUw%0Ahttps://mail.google.com/mail/u/0/%0Apap.

[151] B. Studer and L. Clark, "Place your bets: Psychophysiological correlates of decision-making under risk," *Cognitive, Affective and Behavioral Neuroscience*, vol. 11, pp. 144–158, jun 2011. doi:10.3758/s13415-011-0025-2.

[152] G. Priolo, M. D'alessandro, A. Bizzego, and N. Bonini, "Normatively irrelevant affective cues affect risk-taking under uncertainty: Insights from the iowa gambling task (IGT), skin conductance response, and heart rate variability," *Brain Sciences*, vol. 11, no. 3, pp. 1–17, 2021. doi:10.3390/brainsci11030336.

[153] O. FeldmanHall, P. Glimcher, A. L. Baker, and E. A. Phelps, "Emotion and decision-making under uncertainty: Physiological arousal predicts increased gambling during ambiguity but not risk," *Journal of Experimental Psychology: General*, vol. 145, no. 10, pp. 1255–1262, 2016. doi:10.1037/xge0000205.

[154] J. F. Thayer, A. L. Hansen, E. Saus-Rose, and B. H. Johnsen, "Heart rate variability, prefrontal neural function, and cognitive performance: The neurovisceral integration perspective on self-regulation, adaptation, and health," apr 2009. doi:10.1007/s12160-009-9101-z.

[155] G. Forte, G. Troisi, M. Pazzaglia, V. De Pascalis, and M. Casagrande, "Heart Rate Variability and Pain: A Systematic Review," 2022. doi:10.3390/brainsci12020153.

[156] H. G. Kim, E. J. Cheon, D. S. Bai, Y. H. Lee, and B. H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," mar 2018. doi:10.30773/pi.2017.08.17.

[157] J. E. Muñoz, E. R. Gouveia, M. S. Cameirão, and S. B. I. Badia, "Physiolab - A multivariate physiological computing toolbox for ECG, EMG and EDA signals: A case of study of cardiorespiratory fitness assessment in the elderly population," *Multimedia Tools and Applications*, vol. 77, no. 9, pp. 11511–11546, 2018. doi:10.1007/s11042-017-5069-z.

[158] G. HOCHMAN and E. YECHIAM, "Loss Aversion in the Eye and in the Heart: The Autonomic Nervous System's Responses to Losses," *Journal of Behavioral Decision Making*, vol. 16, no. 20, pp. 6513–6525, 2010. doi:10.1002/bdm.

[159] B. Schmidt, P. Mussel, and J. Hewig, "I'm too calm-Let's take a risk! On the impact of state and trait arousal on risk taking," *Psychophysiology*, vol. 50, no. 5, pp. 498–503, 2013. doi:10.1111/psyp.12032.

[160] J. Fooken, "Heart rate variability indicates emotional value during pro-social economic laboratory decisions with large external validity," *Scientific Reports*, vol. 7, no. June 2016, pp. 1–11, 2017. doi:10.1038/srep44471.

[161] S. Fiedler and A. Glöckner, "The dynamics of decision making in risky choice: An eye-tracking analysis," *Frontiers in Psychology*, vol. 3, no. OCT, pp. 1–18, 2012. doi:10.3389/fpsyg.2012.00335.

[162] A. Konovalov and I. Krajbich, "Revealed Indifference: Using Response Times to Infer Preferences," *SSRN Electronic Journal*, no. 1554837, pp. 1–64, 2017. doi:10.2139/ssrn.3024233.

[163] S. Garcia-Guerrero, D. O'Hora, A. Zgonnikov, and S. Scherbaum, "The Action Dynamics of Approach-Avoidance Conflict in Decision-Making: A Mouse-Tracking Study," *PsyArXiv Preprint*, 2019. doi:10.31234/osf.io/4658p.

[164] M. Maldonado, E. Dunbar, and E. Chemla, "Mouse tracking as a window into decision making," *Behavior Research Methods*, vol. 51, no. 3, pp. 1085–1101, 2019. doi:10.3758/s13428-018-01194-x.

[165] T. Fujimura, K. Katahira, and K. Okanoya, "Contextual modulation of physiological and psychological responses triggered by emotional stimuli," *Frontiers in Psychology*, vol. 4, no. MAY, pp. 1–7, 2013. doi:10.3389/fpsyg.2013.00212.

[166] C. Stoney, "Lipids and Lipoproteins," in *Stress: Neuroendocrinology and Neurobiology*, vol. 2, pp. 287–294, Elsevier, 2017. doi:10.1016/B978-0-12-802175-0.00028-0. ISBN 9780128024232.

[167] D. T. Lykken, "A Tremor in the Blood: Uses and Abuses of the Lie Detector," *Harvard Law Review*, vol. 94, no. 8, p. 1925, 1981. doi:10.2307/1340741.

[168] S. Mathôt, J. Fabius, E. Van Heusden, and S. Van der Stigchel, "Safe and sensible preprocessing and baseline correction of pupil-size data," *Behavior Research Methods*, vol. 50, no. 1, pp. 94–106, 2018. doi:10.3758/s13428-017-1007-2.

[169] N. Thammasan, I. V. Stuldreher, E. Schreuders, M. Giletta, and A. M. Brouwer, "A usability study of physiological measurement in school using wearable sensors," *Sensors (Switzerland)*, vol. 20, no. 18, pp. 1–24, 2020. doi:10.3390/s20185380.

[170] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A Comprehensive Survey on Transfer Learning," 2021. doi:10.1109/JPROC.2020.3004555.

[171] S. N. Rigby, B. M. Stoesz, and L. S. Jakobson, "Gaze patterns during scene processing in typical adults and adults with autism spectrum disorders," *Research in Autism Spectrum Disorders*, vol. 25, pp. 24–36, 2016. doi:10.1016/j.rasd.2016.01.012.

[172] C. J. Anderson and J. Colombo, "Larger tonic pupil size in young children with autism spectrum disorder," *Developmental Psychobiology*, vol. 51, pp. 207–211, mar 2009. doi:10.1002/dev.20352.

[173] K. M. Hengen and G. W. Alpers, "Stress Makes the Difference: Social Stress and Social Anxiety in Decision-Making Under Uncertainty," *Frontiers in Psychology*, vol. 12, no. February, pp. 1–16, 2021. doi:10.3389/fpsyg.2021.578293.

[174] D. Miyamoto, G. Blanc, and Y. Kadobayashi, "Eye can tell: On the correlation between eye movement and phishing identification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9491, pp. 223–232, Springer, Cham, 2015. doi:10.1007/978-3-319-26555-1_26.

[175] M. Alsharnouby, F. Alaca, and S. Chiasson, "Why phishing still works: User strategies for combating phishing attacks," *International Journal of Human Computer Studies*, vol. 82, pp. 69–82, oct 2015. doi:10.1016/j.ijhcs.2015.05.005.

[176] L. Huang, S. Jia, E. Balcetis, and Q. Zhu, "ADVERT: An Adaptive and Data-Driven Attention Enhancement Mechanism for Phishing Prevention," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, jun 2022. doi:10.1109/TIFS.2022.3189530.

[177] K. Yu, R. Taib, M. A. Butavicius, K. Parsons, and F. Chen, "Mouse Behavior as an Index of Phishing Awareness," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11746 LNCS, pp. 539–548, Springer, Cham, sep 2019. doi:10.1007/978-3-030-29381-9_33.

[178] H. T. Cheer, *Choice Confidence And Persuasion Resistance Through Mouse Action Observation*. Ms.c., California State University Monterey Bay, 2020. url: https://apps.dtic.mil/sti/citations/AD1126796.

[179] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The iCub humanoid robot: An open platform for research in embodied cognition," in *Performance Metrics for Intelligent Systems (PerMIS) Workshop*, (New York, New York, USA), pp. 50–56, ACM Press, 2008. doi:10.1145/1774674.1774683.

[180] G. Sandini, G. Metta, and D. Vernon, "The iCub cognitive humanoid robot: An open-system research platform for enactive cognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4850 LNAI, pp. 358–369, Springer Verlag, 2007. doi:10.1007/978-3-540-77296-5_32.

[181] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A back-projected human-like robot head for multiparty human-machine interaction," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7403 LNCS, pp. 114–130, 2012. doi:10.1007/978-3-642-34584-5_9.

[182] L. Royakkers and R. van Est, "A Literature Review on New Robotics: Automation from Love to War," *International Journal of Social Robotics*, vol. 7, pp. 549–570, nov 2015. doi:10.1007/s12369-015-0295-x.

[183] J. Casper and R. R. Murphy, "Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 33, pp. 367–385, jun 2003. doi:10.1109/TSMCB.2003.811794.

[184] D. Bruemmer, J. Marble, D. Dudenhoeffer, M. Anderson, and M. McKay, "Intelligent Robots for Use in Hazardous DOE Environments," *NIST Special Publication SP*, vol. 990, no. 1048-776X, pp. 209–216, 2002. url: https://apps.dtic.mil/sti/citations/ADA516023.

[185] D. Pasquali, J. Gonzalez-Billandon, F. Rea, G. Sandini, and A. Sciutti, "Magic iCub: a humanoid robot autonomously catching your lies in a card game," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21)*, (Boulder (Colorado)), 2021. doi:https://doi.org/10.1145/3434073.3444682.

[186] L. Gong, C. Gong, Z. Ma, L. Zhao, Z. Wang, X. Li, X. Jing, H. Yang, and C. Liu, "Real-time human-in-the-loop remote control for a life-size traffic police robot with multiple augmented reality aided display terminals," in *2017 2nd International Conference on Advanced Robotics and Mechatronics, ICARM 2017*, vol. 2018-Janua, pp. 420–425, Institute of Electrical and Electronics Engineers Inc., jan 2018. doi:10.1109/ICARM.2017.8273199.

[187] Y. Gao and S. Chien, "Review on space robotics: Toward top-level science through space exploration," jun 2017. doi:10.1126/scirobotics.aan5074.

[188] E. Broadbent, "Interactions with Robots: The Truths We Reveal about Ourselves," *Annual Review of Psychology*, vol. 68, pp. 627–652, jan 2017. doi:10.1146/annurev-psych-010416-043958.

[189] P. L. Rau, Y. Li, and D. Li, "A cross-cultural study: Effect of robot appearance and task," *International Journal of Social Robotics*, vol. 2, pp. 175–186, may 2010. doi:10.1007/s12369-010-0056-9.

[190] 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, pp. 55–66, jan 2016. doi:10.1515/pjbr-2016-0005.

[191] C. Breazeal, "Regulating human-robot interaction using "emotions", "drives" and facial expressions," *Proceedings of the 2nd International Conference on Autonomous Agents*, pp. 14–21, 1998. url: http://robotic.media.mit.edu/wp-content/uploads/sites/14/2015/01/Breazeal-Agents-98.pdf.

[192] D. Mazzei, F. Chiarello, and G. Fantoni, "Analyzing Social Robotics Research with Natural Language Processing Techniques," *Cognitive Computation*, vol. 13, pp. 308–321, mar 2021. doi:10.1007/s12559-020-09799-1.

[193] F. Operto, "Robots doing the most arduous tasks in our place," 2020. url: https://www.eni.com/en-IT/scientific-research/robotic-automation.html.

[194] X. Yu, C. Xu, X. Zhang, and L. Ou, "Real-time multitask multihuman–robot interaction based on context awareness," *Robotica*, pp. 1–27, feb 2022. doi:10.1017/s0263574722000017.

[195] M. De Haas, A. M. Aroyo, E. Barakova, W. Haselager, and I. Smeekens, "The effect of a semi-autonomous robot on children," in *2016 IEEE 8th International Conference on Intelligent Systems, IS 2016 - Proceedings*, pp. 376–381, Institute of Electrical and Electronics Engineers Inc., nov 2016. doi:10.1109/IS.2016.7737448.

[196] H. Robinson, B. A. MacDonald, N. Kerse, and E. Broadbent, "Suitability of Healthcare Robots for a Dementia Unit and Suggested Improvements," *Journal of the American Medical Directors Association*, vol. 14, no. 1, pp. 34–40, 2013. doi:10.1016/j.jamda.2012.09.006.

[197] 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," *Computers in Human Behavior*, vol. 43, pp. 112–117, feb 2015. doi:10.1016/j.chb.2014.10.029.

[198] E. Broadbent, K. Peri, N. Kerse, C. Jayawardena, I. Kuo, C. Datta, and B. MacDonald, "Robots in older people's homes to improve medication adherence and quality of life: A randomised cross-over trial," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8755, pp. 64–73, Springer, Cham, oct 2014. doi:10.1007/978-3-319-11973-1_7.

[199] H. Robinson, B. MacDonald, and E. Broadbent, "The Role of Healthcare Robots for Older People at Home: A Review," *International Journal of Social Robotics*, vol. 6, pp. 575–591, nov 2014. doi:10.1007/s12369-014-0242-2.

[200] O. Mubin, C. J. Stevens, S. Shahid, A. A. Mahmud, and J.-J. Dong, "A REVIEW OF THE APPLICABILITY OF ROBOTS IN EDUCATION," *Technology for Education and Learning*, vol. 1, no. 1, 2013. doi:10.2316/journal.209.2013.1.209-0015.

[201] M. de Haas, A. M. Aroyo, P. Haselager, I. Smeekens, and E. Barakova, "Comparing Robots with Different Levels of Autonomy in Educational Setting," in *Studies in Systems, Decision and Control*, vol. 140, pp. 293–311, Springer International Publishing, 2018. doi:10.1007/978-3-319-78437-3_13.

[202] J. D. Lewis and A. Weigert, "Trust as a social reality," *Social Forces*, vol. 63, no. 4, pp. 967–985, 1985. doi:10.1093/sf/63.4.967.

[203] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," 2004. doi:10.1518/hfes.46.1.50_30392.

[204] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, pp. 517–527, oct 2011. doi:10.1177/0018720811417254.

[205] 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," in *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, (New York, New York, USA), pp. 73–80, ACM Press, 2012. doi:10.1145/2157689.2157702.

[206] B. M. Muir, "Trust between humans and machines," *International Journal of Man-Machine Studies*, vol. 27, pp. 327–339, 1987.

[207] D. P. Biros, M. Daly, and G. Gunsch, "The influence of task load and automation trust on deception detection," *Group Decision and Negotiation*, vol. 13, pp. 173–189, mar 2004. doi:10.1023/B:GRUP.0000021840.85686.57.

[208] J. Bernotat, F. Eyssel, and J. Sachse, "The (Fe)male Robot: How Robot Body Shape Impacts First Impressions and Trust Towards Robots," *International Journal of Social Robotics*, vol. 13, no. 3, pp. 477–489, 2021. doi:10.1007/s12369-019-00562-7.

[209] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, pp. 1243–1270, oct 1992. doi:10.1080/00140139208967392.

[210] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2016-April, pp. 109–116, 2016. doi:10.1109/HRI.2016.7451741.

[211] 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," in *PIE - The International Society for Optical Engineering* (R. E. Karlsen, D. W. Gage, C. M. Shoemaker, and G. R. Gerhart, eds.), no. February 2015, p. 90840E, jun 2014. doi:10.1117/12.2050622.

[212] T. L. Sanders, T. Wixon, K. E. Schafer, J. Y. Chen, and P. A. Hancock, "The influence of modality and transparency on trust in human-robot interaction," in *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, CogSIMA 2014*, pp. 156–159, IEEE, mar 2014. doi:10.1109/CogSIMA.2014.6816556.

[213] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would You Trust a (Faulty) Robot?," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, (New York, NY, USA), pp. 141–148, ACM, mar 2015. doi:10.1145/2696454.2696497.

[214] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, vol. 2016-April, pp. 101–108, IEEE, mar 2016. doi:10.1109/HRI.2016.7451740.

[215] B. Postnikoff, *Robot Social Engineering*. PhD thesis, University ofWaterloo, 2020. url: http://hdl.handle.net/10012/16030.

[216] S. Gibbs, "Hackers can hijack Wi-Fi Hello Barbie to spy on your children," 2015. url: https://www.theguardian.com/technology/2015/nov/26/hackers-can-hijack-wi-fi-hello-barbie-to-spy-on-your-children.

[217] L. Franceschi-Bicchierai, "How This Internet of Things Stuffed Animal Can Be Remotely Turned Into a Spy Device," 2017. url: https://www.vice.com/en/article/qkm48b/how-this-internet-of-things-teddy-bear-can-be-remotely-turned-into-a-spy-device.

[218] D. Lepido, "It's not science fiction. Robots running industrial world can be hacked, remote-controlled," 2020. url: https://theprint.in/tech/its-not-science-fiction-robots-running-industrial-world-can-be-hacked-remote-controlled/475340/.

[219] 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," apr 2015. doi:10.48550/arxiv.1504.04339.

[220] C. Cerrudo, "Hacking Robots Before Skynet," *Cybersecurity Insight*, pp. 1–17, 2017. url: https://ioactive.com/hacking-robots-before-skynet/.

[221] A. S. Ghazali, J. Ham, E. Barakova, and P. Markopoulos, "Persuasive Robots Acceptance Model (PRAM): Roles of Social Responses Within the Acceptance Model of Persuasive Robots," *International Journal of Social Robotics*, vol. 12, no. 5, pp. 1075–1092, 2020. doi:10.1007/s12369-019-00611-1.

[222] A. S. Ghazali, J. Ham, E. Barakova, and P. Markopoulos, "Assessing the effect of persuasive robots interactive social cues on users' psychological reactance, liking, trusting beliefs and compliance," *Advanced Robotics*, vol. 33, no. 7-8, pp. 325–337, 2019. doi:10.1080/01691864.2019.1589570.

[223] J. Zonca and A. Sciutti, "Does human-robot trust need reciprocity?," *Workshop Preoceedings*, no. August, 2021. doi:https://doi.org/10.48550/arXiv.2110.09359.

[224] J. Zonca, A. Folsø, and A. Sciutti, "The role of reciprocity in human-robot social influence," *iScience*, vol. 24, no. 12, p. 103424, 2021. doi:10.1016/j.isci.2021.103424.

[225] 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," in *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2016-April, pp. 117–124, IEEE Computer Society, apr 2016. doi:10.1109/HRI.2016.7451742.

[226] E. B. Sandoval, *Reciprocity in Human Robot Interaction*. PhD thesis, University of Canterbury, 2016. doi:http://dx.doi.org/10.26021/3429.

[227] A. van Wynsberghe, "Social robots and the risks to reciprocity," *AI and Society*, vol. 37, no. 2, pp. 479–485, 2022. doi:10.1007/s00146-021-01207-y.

[228] E. B. Sandoval, J. Brandstatter, U. Yalcin, and C. Bartneck, "Robot Likeability and Reciprocity in Human Robot Interaction: Using Ultimatum Game to determinate Reciprocal Likeable Robot Strategies," *International Journal of Social Robotics*, vol. 13, no. 4, pp. 851–862, 2021. doi:10.1007/s12369-020-00658-5.

[229] S. Saunderson and G. Nejat, "It would make me happy if you used my guess: Comparing robot persuasive strategies in social human-robot interaction," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1707–1714, 2019. doi:10.1109/LRA.2019.2897143.

[230] M. Székely, H. Powell, F. Vannucci, F. Rea, A. Sciutti, and J. Michael, "The perception of a robot partner's effort elicits a sense of commitment to human-robot interaction," *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, vol. 20, no. 2, pp. 234–255, 2019. doi:10.1075/is.18001.sze.

[231] M. Klichowski, "People Copy the Actions of Artificial Intelligence," *Frontiers in Psychology*, vol. 11, p. 1130, 2020. doi:10.3389/fpsyg.2020.01130.

[232] D. Cormier, J. Young, M. Nakane, G. Newman, and S. Durocher, "Would You Do as a Robot Commands? An Obedience Study for Human-Robot Interaction," *International Conference on Human-Agent Interaction*, pp. I–3–1, 2013. url: https://hci.cs.umanitoba.ca/assets/publication_files/2013-would-you-do-as-a-robot-commands.pdf.

[233] A. Boos, O. Herzog, J. Reinhardt, K. Bengler, and M. Zimmermann, "A Compliance–Reactance Framework for Evaluating Human-Robot Interaction," *Frontiers in Robotics and AI*, vol. 9, no. May, pp. 1–13, 2022. doi:10.3389/frobt.2022.733504.

[234] A. Vance, D. Eargle, J. L. Jenkins, C. Brock Kirwan, and B. B. Anderson, "The fog of warnings: How non-essential notifications blur with security warnings," *Proceedings of the 15th Symposium on Usable Privacy and Security, SOUPS 2019*, pp. 407–420, 2019. url: https://dl.acm.org/doi/10.5555/3361476.3361506.

[235] B. B. Anderson, A. Vance, C. B. Kirwan, J. L. Jenkins, and D. Eargle, "From Warning to Wallpaper: Why the Brain Habituates to Security Warnings and What Can Be Done About It," *Journal of Management Information Systems*, vol. 33, no. 3, pp. 713–743, 2016. doi:10.1080/07421222.2016.1243947.

[236] Y. Do, L. T. Hoang, J. W. Park, G. D. Abowd, and S. Das, "Spidey Sense: Designing Wrist-Mounted Affective Haptics for Communicating Cybersecurity Warnings," *DIS 2021 - Proceedings of the 2021 ACM Designing Interactive Systems Conference: Nowhere and Everywhere*, vol. 2021, pp. 125–137, 2021. doi:10.1145/3461778.3462027.

[237] N. Agrawal, F. Zhu, and S. Carpenter, "Do You See the Warning?," in *Proceedings of the 2020 ACM Southeast Conference*, (New York, NY, USA), pp. 260–263, ACM, apr 2020. doi:10.1145/3374135.3385314.

[238] C. Mazzola, A. M. Aroyo, F. Rea, and A. Sciutti, "Interacting with a social robot affects visual perception of space," in *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 549–557, IEEE Computer Society, mar 2020. doi:10.1145/3319502.3374819.

[239] G. Maggi, E. Dell'Aquila, I. Cucciniello, and S. Rossi, ""Don't Get Distracted!": The Role of Social Robots' Interaction Style on Users' Cognitive Performance, Acceptance, and Non-Compliant Behavior," *International Journal of Social Robotics*, vol. 13, pp. 2057–2069, dec 2021. doi:10.1007/s12369-020-00702-4.

[240] S. Rossi, M. Larafa, and M. Ruocco, "Emotional and Behavioural Distraction by a Social Robot for Children Anxiety Reduction During Vaccination," *International Journal of Social Robotics*, vol. 12, pp. 765–777, jul 2020. doi:10.1007/s12369-019-00616-w.

[241] A. Tanevska, F. Rea, G. Sandini, L. Cañamero, and A. Sciutti, "A Socially Adaptable Framework for Human-Robot Interaction," *Frontiers in Robotics and AI*, vol. 7, p. 121, oct 2020. doi:10.3389/frobt.2020.00121.

[242] S. Vinanzi, M. Patacchiola, A. Chella, and A. Cangelosi, "Would a robot trust you? Developmental robotics model of trust and theory of mind," *CEUR Workshop Proceedings*, vol. 2418, p. 74, 2019. doi:https://doi.org/10.1098/rstb.2018.0032.

[243] M. Patacchiola and A. Cangelosi, "A Developmental Cognitive Architecture for Trust and Theory of Mind in Humanoid Robots," *IEEE Transactions on Cybernetics*, pp. 1–13, 2020. doi:10.1109/TCYB.2020.3002892.

[244] J. Schell, *The Art of Game Design : a Book of Lenses*. Elsevier/Morgan Kaufmann, 2008.

[245] N. Z. Legaki, K. Karpouzis, V. Assimakopoulos, and J. Hamari, "Gamification to avoid cognitive biases: An experiment of gamifying a forecasting course," *Technological Forecasting and Social Change*, vol. 167, p. 120725, 2021. doi:10.1016/j.techfore.2021.120725.

[246] N. L. Robinson, S. Turkay, L. A. Cooper, and D. Johnson, "Social robots with gamification principles to increase long-Term user interaction," *ACM International Conference Proceeding Series*, pp. 359–363, 2019. doi:10.1145/3369457.3369494.

[247] M. C. Green and K. M. Jenkins, "Interactive narratives: Processes and outcomes in user-directed stories," *Journal of Communication*, vol. 64, no. 3, pp. 479–500, 2014. doi:10.1111/jcom.12093.

[248] G. F. Kaufman and L. K. Libby, "Changing beliefs and behavior through experience-taking," *Journal of Personality and Social Psychology*, vol. 103, no. 1, pp. 1–19, 2012. doi:10.1037/a0027525.

[249] J. H. Klein, "The abstraction of reality for games and simulations," *Journal of the Operational Research Society*, vol. 36, no. 8, pp. 671–678, 1985. doi:10.1057/jors.1985.124.

[250] A. Aroyo, D. Pasquali, A. Koting, F. Rea, G. Sandini, and A. Sciutti, "Perceived differences between on-line and real robotic failures," in *RO-MAN 2020 - Trust, Acceptance and Social Cues in Human-Robot Interaction - SCRITA*, 2020. url: https://www.researchgate.net/publication/344403863_Perceived_differences_between_on-line_and_real_robotic_failures.

[251] A. M. Aroyo, D. Pasquali, A. Kothig, F. Rea, G. Sandini, and A. Sciutti, "Expectations Vs. Reality: Unreliability and Transparency in a Treasure Hunt Game with Icub," *IEEE Robotics and Automation Letters*, vol. 6, pp. 5681–5688, jul 2021. doi:10.1109/LRA.2021.3083465.

[252] D. Pasquali, A. Sciutti, G. Sandini, S. Bencetti, F. Rea, and U. Genova, "Toward a Human-Oriented Social Engineering Defense System," in *Ital-IA 2022 - Workshop on AI for Cybersecyrity*, 2022.

[253] L. Riek, "Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 119–136, 2012. doi:10.5898/jhri.1.1.riek.

[254] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Frontiers in Psychology*, vol. 9, no. JUN, 2018. doi:10.3389/fpsyg.2018.00861.

[255] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley," *IEEE Robotics and Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012. doi:10.1109/MRA.2012.2192811.

[256] G. B. Flebus, "Versione Italiana dei Big Five Markers di Goldberg," tech. rep., Università di Milano-Bicocca, 2015. doi:https://docplayer.it/51528201-Versione-italiana-dei-big-five-markers-di-goldberg-giovanni-battista-flebus-universita-di-milano-bicocca.html.

[257] M. A. Guillemette, R. Yao, and R. N. James III, "An analysis of risk assessment questions based on loss-averse preferences," *Journal of Financial Counseling and Planning*, vol. 26, no. 1, pp. 17–29, 2015. url: https://ssrn.com/abstract=2740042.

[258] B. Rohrmann, "Risk Attitude Scales : Concepts , Questionnaires , Utilizations," Tech. Rep. January, 2005. url: http://www.rohrmannresearch.net/pdfs/rohrmann-racreport.pdf.

[259] F. Calado, J. Alexandre, and M. D. Griffiths, "Prevalence of Adolescent Problem Gambling: A Systematic Review of Recent Research," *Journal of Gambling Studies*, vol. 33, pp. 397–424, jun 2017. doi:10.1007/s10899-016-9627-5.

[260] 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," *Human Resource Management Journal*, vol. 22, pp. 360–376, nov 2012. doi:10.1111/1748-8583.12007.

[261] D. S. Syrdal, K. Dautenhahn, K. Koay, and M. Walters, "The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study," *Adaptive and Emergent Behaviour and Complex Systems*, pp. 109–115, 2009. doi:10.1157/13126291.

[262] 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?: Psychological Intimacy in HRI," in *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2015-March, pp. 173–180, IEEE Computer Society, mar 2015. doi:10.1145/2696454.2696486.

[263] F. Ferrari, M. P. Paladino, and J. Jetten, "Blurring Human–Machine Distinctions: Anthropomorphic Appearance in Social Robots as a Threat to Human Distinctiveness," *International Journal of Social Robotics*, vol. 8, no. 2, pp. 287–302, 2016. doi:10.1007/s12369-016-0338-y.

[264] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," nov 2009. doi:10.1007/s12369-008-0001-3.

[265] F. Bracco and C. Chiorri, "Versione Italiana del NASA-TLX," tech. rep., University of Genoa, 2008.

[266] A. Aron, E. N. Aron, and D. Smollan, "Inclusion of Other in the Self Scale and the Structure of Interpersonal Closeness," *Journal of Personality and Social Psychology*, vol. 63, pp. 596–612, jan 1992. doi:10.1037/0022-3514.63.4.596.

[267] C. D. Kidd, *Sociable Robots : The Role of Presence and Task in Human-Robot Interaction*. PhD thesis, Massachusetts Institute of Technology, 2003. url: https://www.media.mit.edu/publications/sociable-robots-the-role-of-presence-and-task-in-human-robot-interaction/.

[268] 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," *Computers in Human Behavior*, vol. 61, pp. 633–655, 2016. doi:10.1016/j.chb.2016.03.057.

[269] H. M. Gray, K. Gray, and D. M. Wegner, "Dimensions of mind perception," *Science*, vol. 315, p. 619, feb 2007. doi:10.1126/science.1134475.

[270] 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. doi:10.1007/s12369-008-0001-3.

[271] A. R. Wagner and P. Robinette, "An explanation is not an excuse: Trust calibration in an age of transparent robots," in *Trust in Human-Robot Interaction*, pp. 197–208, Academic Press, jan 2021. doi:10.1016/b978-0-12-819472-0.00009-5. ISBN 9780128194720.

[272] J. Zhang, S. I. Levitan, and J. Hirschberg, "Multimodal deception detection using automatically extracted acoustic, visual, and lexical features," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-Octob, pp. 359–363, 2020. doi:10.21437/Interspeech.2020-2320.

[273] A. Vrij, S. A. Mann, R. P. Fisher, S. Leal, R. Milne, and R. Bull, "Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order," *Law and Human Behavior*, vol. 32, no. 3, pp. 253–265, 2008. doi:10.1007/s10979-007-9103-y.

[274] A. Moliné, E. Dominguez, E. Salazar-López, G. Gálvez-García, J. Fernández-Gómez, J. De la Fuente, O. Iborra, F. J. Tornay, and E. Gómez Milán, "The mental nose and the Pinocchio effect: Thermography, planning, anxiety, and lies," *Journal of Investigative Psychology and Offender Profiling*, vol. 15, no. 2, pp. 234–248, 2018. doi:10.1002/jip.1505.

[275] L. V. Eberhardt, G. Grön, M. Ulrich, A. Huckauf, and C. Strauch, "Direct voluntary control of pupil constriction and dilation: Exploratory evidence from pupillometry, optometry, skin conductance, perception, and functional MRI," *International Journal of Psychophysiology*, vol. 168, pp. 33–42, oct 2021. doi:10.1016/j.ijpsycho.2021.08.001.

[276] Tobii, "Tobii Pro Glasses 2." url: https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/.

[277] P. Fitzpatrick, G. Metta, and L. Natale, "Towards long-lived robot genes," *Robotics and Autonomous Systems*, vol. 56, pp. 29–45, jan 2008. doi:10.1016/j.robot.2007.09.014.

[278] D. De Tommaso and A. Wykowska, "TobiiGlassesPySuite," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research and Applications*, (New York, NY, USA), pp. 1–5, ACM, jun 2019. doi:10.1145/3314111.3319828.

[279] J. Sweller, *Cognitive Load Theory*. Psychology of learning and motivation, volume 55. Elsevier., 2011. doi:10.1016/B978-0-12-387691-1.00002-8.

[280] J. Leppink, "Cognitive load theory: Practical implications and an important challenge," *Journal of Taibah University Medical Sciences*, vol. 12, no. 5, pp. 385–391, 2017. doi:10.1016/j.jtumed.2017.05.003.

[281] 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," *Legal and Criminological Psychology*, vol. 14, pp. 279–292, sep 2009. doi:10.1348/135532508X398602.

[282] Tobii Pro, "Quick Tech Webinar - Secrets of the Pupil." url: https://www.youtube.com/watch?v=I3T9Ak2F2bc&feature=emb_title.

[283] D. Pasquali, A. M. Aroyo, J. Gonzalez-billandon, F. Rea, G. Sandini, and A. Sciutti, "Your Eyes Never Lie: A Robot Magician Can Tell if You Are Lying," in *In Proceedings' of HRI (HRI '20) Cambridge conference*, 2020. doi:https://doi.org/10.1145/3371382.3378253.

[284] D. Pasquali, A. Aroyo, J. Gonzalez-Billandon, F. Rea, G. Sandini, and A. Sciutti, "Do You See the Magic? An Autonomous Robot Magician Can Read Your Mind," in *HRI 2020 Workshop on Exploring Creative Content in Social Robotics*, pp. 2–5, 2020. url: https://mypersonalrobots.org/s/HRI_2020_WS_Creative_SR_paper_6.pdf.

[285] D. Pasquali, J. Gonzalez-Billandon, A. M. Aroyo, G. Sandini, A. Sciutti, and F. Rea, "Detecting Lies is a Child (Robot)'s Play: Gaze-Based Lie Detection in HRI," *International Journal of Social Robotics*, pp. 1–16, nov 2021. doi:10.1007/s12369-021-00822-5.

[286] D. Pasquali, D. Gaggero, G. Volpe, F. Rea, and A. Sciutti, "Human vs Robot Lie Detector: Better Working as a Team?," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13086 LNAI, pp. 154–165, Springer, Cham, nov 2021. doi:10.1007/978-3-030-90525-5_14.

[287] J.-L. Roubira and M. Cardouat, "Dixit 3: Journey | Board Game | BoardGameGeek," 2012. url: https://boardgamegeek.com/boardgame/119657/dixit-3-journey.

[288] C. J. Ferguson and C. Negy, "Development of a brief screening questionnaire for histrionic personality symptoms," *Personality and Individual Differences*, vol. 66, pp. 124–127, 2014. doi:10.1016/j.paid.2014.02.029.

[289] D. N. Jones and D. L. Paulhus, "Introducing the Short Dark Triad (SD3): A Brief Measure of Dark Personality Traits," *Assessment*, vol. 21, no. 1, pp. 28–41, 2014. doi:10.1177/1073191113514105.

[290] M. Barandas, D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, and H. Gamboa, "TSFEL: Time Series Feature Extraction Library," *SoftwareX*, vol. 11, jan 2020. doi:10.1016/j.softx.2020.100456.

[291] C. F. Bond and B. M. DePaulo, "Accuracy of deception judgments," *Personality and Social Psychology Review*, vol. 10, no. 3, pp. 214–234, 2006. doi:10.1207/s15327957pspr1003_2.

[292] L. Breiman, "Random forests," *Machine learning*, 2001. doi:10.1023/A:1010933404324.

[293] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. doi:10.1613/jair.953.

[294] S. K. Patro and K. K. Sahu, "Normalization: A Preprocessing Stage," *IARJSET*, pp. 20–22, mar 2015. doi:10.17148/IARJSET.2015.2305.

[295] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation Measures for Models Assessment over Imbalanced Data Sets," *Journal of Information Engineering and Applications*, vol. 3, no. 10, pp. 27–38, 2013. url: http://www.iiste.org/Journals/index.php/JIEA/article/view/7633.

[296] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001. doi:10.1162/089976601750264965.

[297] H. S. Ahn, I. K. Sa, D. W. Lee, and D. Choi, "A playmate robot system for playing the rock-paper-scissors game with humans," *Artificial Life and Robotics*, vol. 16, no. 2, pp. 142–146, 2011. doi:10.1007/s10015-011-0895-y.

[298] I. Gori, S. R. Fanello, G. Metta, and F. Odone, "All gestures you can: A memory game against a humanoid robot," in *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*, pp. 330–336, IEEE, nov 2012. doi:10.1109/HUMANOIDS.2012.6651540.

[299] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, "Comparing Models of Disengagement in Individual and Group Interactions," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2015-March, no. March, pp. 99–105, 2015. doi:10.1145/2696454.2696466.

[300] M. Owayjan, A. Kashour, N. Al Haddad, M. Fadel, and G. Al Souki, "The design and development of a lie detection system using facial micro-expressions," in *2012 2nd International Conference on Advances in Computational Tools for Engineering Applications, ACTEA 2012*, pp. 33–38, IEEE, dec 2012. doi:10.1109/ICTEA.2012.6462897.

[301] K. Kobayashi and S. Yamada, "Human-Robot interaction design for low cognitive load in cooperative work," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pp. 569–574, IEEE, 2004. doi:10.1109/roman.2004.1374823.

[302] S. M. Al Mahi, M. Atkins, and C. Crick, "Learning to assess the cognitive capacity of human partners," *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 63–64, 2017. doi:10.1145/3029798.3038430.

[303] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-time eye gaze estimation in natural environments," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11214 LNCS, pp. 339–357, 2018. doi:10.1007/978-3-030-01249-6_21.

[304] M. E. Foster, B. Craenen, A. Deshmukh, O. Lemon, E. Bastianelli, C. Dondrup, I. Papaioannou, A. Vanzo, J.-M. Odobez, O. Canévet, Y. Cao, W. He, A. Martínez-González, P. Motlicek, R. Siegfried, R. Alami, K. Belhassein, G. Buisan, A. Clodic, A. Mayima, Y. Sallami, G. Sarthou, P.-T. Singamaneni, J. Waldhart, A. Mazel, M. Caniot, M. Niemelä, P. Heikkilä, H. Lammi, and A. Tammela, "MuMMER: Socially Intelligent Human-Robot Interaction in Public Spaces," *arXiv preprint arXiv:1909.06749*, 2019. url: https://gitlab.idiap.ch/software/openheadposehttp://arxiv.org/abs/1909.06749.

[305] K. Stefanov, J. Beskow, and G. Salvi, "Self-supervised vision-based detection of the active speaker as support for socially aware language acquisition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, pp. 250–259, jun 2020. doi:10.1109/TCDS.2019.2927941.

[306] J. Gonzalez, G. Belgiovine, A. Sciutti, G. Sandini, and R. Francesco, "Towards a cognitive framework for multimodal person recognition in multiparty hri," *HAI 2021 - Proceedings of the 9th International User Modeling, Adaptation and Personalization Human-Agent Interaction*, pp. 412–416, 2021. doi:10.1145/3472307.3484675.

[307] J. L. Redifer, C. L. Bae, and M. DeBusk-Lane, "Implicit Theories, Working Memory, and Cognitive Load: Impacts on Creative Thinking," *SAGE Open*, vol. 9, no. 1, 2019. doi:10.1177/2158244019835919.

[308] G. Belgiovine, F. Rea, J. Zenzeri, and A. Sciutti, "A Humanoid Social Agent Embodying Physical Assistance Enhances Motor Training Experience," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, no. ii, pp. 553–560, IEEE, aug 2020. doi:10.1109/RO-MAN47096.2020.9223335.

[309] A. Koenig, D. Novak, X. Omlin, M. Pulfer, E. Perreault, L. Zimmerli, M. Mihelj, and R. Riener, "Real-time closed-loop control of cognitive load in neurological patients during robot-assisted gait training," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 19, pp. 453–464, aug 2011. doi:10.1109/TNSRE.2011.2160460.

[310] A. Westbrook and T. S. Braver, "Cognitive effort: A neuroeconomic approach," jun 2015. doi:10.3758/s13415-015-0334-y.

[311] C. Tosone, "Living everyday lies: The experience of self," *Clinical Social Work Journal*, vol. 34, no. 3, pp. 335–348, 2006. doi:10.1007/s10615-005-0035-z.

[312] C. Stern and W. Stern, *Recollection, testimony, and lying in early childhood.* American Psychological Association, oct 2004. doi:10.1037/10324-000.

[313] V. Talwar and K. Lee, "Development of lying to conceal a transgression: Children's control of expressive behaviour during verbal deception," *International Journal of Behavioral Development*, vol. 26, no. 5, pp. 436–444, 2002. doi:10.1080/01650250143000373.

[314] A. Vrij, "Why professionals fail to catch liars and how they can improve," *Legal and Criminological Psychology*, vol. 9, pp. 159–181, sep 2004. doi:10.1348/1355325041719356.

[315] T. Brennen and S. Magnussen, "Research on Non-verbal Signs of Lies and Deceit: A Blind Alley," *Frontiers in Psychology*, vol. 11, no. December, pp. 1–4, 2020. doi:10.3389/fpsyg.2020.613410.

[316] Vimeo, "Vimeo," 2021. url: https://vimeo.com/.

[317] SurveyMonkey, "SurveyMonkey," 2021. url: www.surveymonkey.com/mp/audience.

[318] C. Chiorri, F. Bracco, T. Piccinno, C. Modafferi, and V. Battini, "Psychometric properties of a revised version of the ten item personality inventory," *European Journal of Psychological Assessment*, vol. 31, no. 2, pp. 109–119, 2015. doi:10.1027/1015-5759/a000215.

[319] T. ERIC and R. FRANKLIN, "Keys to Successful Interactive Storytelling: A Study of the Booming "Choose-Your-Own-Adventure" Video Game Industry," *i-manager's Journal of Educational Technology*, vol. 13, no. 3, p. 28, 2016. doi:10.26634/jet.13.3.8318.

[320] A.-r. Blais and E. U. Weber, "A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations," *A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations*, vol. 1, no. 1, pp. 33–47, 2006. doi:http://dx.doi.org/10.1037/t13084-000.

[321] C. Klimas, "Twine / An open-source tool for telling interactive, nonlinear stories," 2009. url: https://twinery.org/.

[322] Harlowe, "Harlowe 3.2.2," 2021. url: https://twine2.neocities.org/.

[323] Python Software Foundation, "PySide2 · PyPI," 2019. url: https://pypi.org/project/PySide2/.

[324] Sr-research, "EyeLink 1000 Plus - The Most Flexible Eye Tracker - SR Research," 2015. url: https://www.sr-research.com/eyelink-1000-plus/.

[325] Shimmer, "GSR Sensor Development Kit | GSR sensors | Galvanic Skin Response | EDA," 2017. url: https://www.shimmersensing.com/products/gsr-optical-pulse-development-kit.

[326] S. Woods, K. Dautenhahn, and C. Kaouri, "Is someone watching me? - Consideration of social facilitation effects in human-robot interaction experiments," in *Proceedings of IEEE International Symposium on Computational Intelligence in Robotics and Automation, CIRA*, pp. 53–60, 2005. doi:10.1109/cira.2005.1554254.

[327] P. Su and X. Yuan, "Are You Watching Me? A Study on Privacy Notice Design of Social Robot," in *Lecture Notes in Networks and Systems*, vol. 261, pp. 339–344, Springer Science and Business Media Deutschland GmbH, 2021. doi:10.1007/978-3-030-79760-7_41.

[328] I. D'Orta, J. Burnay, D. Aiello, C. Niolu, A. Siracusano, L. Timpanaro, Y. Khazaal, and J. Billieux, "Development and validation of a short Italian UPPS-P impulsive behavior scale," *Addictive Behaviors Reports*, vol. 2, pp. 19–22, 2015. doi:10.1016/j.abrep.2015.04.003.

[329] A. Fossati, A. Somma, K. A. Karyadi, M. A. Cyders, R. Bortolla, and S. Borroni, "Reliability and validity of the Italian translation of the UPPS-P Impulsive Behavior Scale in a sample of consecutively admitted psychotherapy patients," *Personality and Individual Differences*, vol. 91, pp. 1–6, 2016. doi:10.1016/j.paid.2015.11.020.

[330] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The Robotic Social Attributes Scale (RoSAS): Development and Validation," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. Part F1271, no. March 2017, pp. 254–262, 2017. doi:10.1145/2909824.3020208.

[331] V. V. Abeele, K. Spiel, L. Nacke, D. Johnson, and K. Gerling, "Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences," *International Journal of Human Computer Studies*, vol. 135, no. June 2019, p. 102370, 2020. doi:10.1016/j.ijhcs.2019.102370.

[332] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. Chen, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, pp. 1689–1696, feb 2021. doi:10.3758/s13428-020-01516-y.

[333] EoinTravers, "Squeak: Python library for processing and analysing data from mouse tracking psychological experiments.," 2016. url: https://github.com/EoinTravers/Squeak.

[334] J. B. Freeman and R. Dale, "Assessing bimodality to detect the presence of a dual cognitive process," *Behavior Research Methods*, vol. 45, pp. 83–97, jul 2013. doi:10.3758/s13428-012-0225-x.

[335] T. J. Morganand and K. N. Laland, "The biological bases of conformity," 2012. doi:10.3389/fnins.2012.00087.

[336] J. K. Burgoon and T. R. Levine, "Advances in deception detection," in *New Directions in Interpersonal Communication Research*, pp. 201–220, SAGE Publications Inc., jan 2010. doi:10.4135/9781483349619.n10. ISBN 9781483349619.

[337] C. M. Mills, "Knowing when to doubt: Developing a critical stance when learning from others.," *Developmental Psychology*, vol. 49, pp. 404–418, mar 2013. doi:10.1037/a0029500.

[338] C. D. Kidd and C. Breazeal, "Robots at home: Understanding long-term human-robot interaction," *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pp. 3230–3235, 2008. doi:10.1109/IROS.2008.4651113.

[339] S. P. Saunderson and G. Nejat, "Persuasive robots should avoid authority: The effects of formal and real authority on persuasion in human-robot interaction," *Science Robotics*, vol. 6, sep 2021. doi:10.1126/scirobotics.abd5186.

[340] F. Bellotti, B. Kapralos, K. Lee, P. Moreno-Ger, and R. Berta, "Assessment in and of serious games: An overview," *Advances in Human-Computer Interaction*, vol. 2013, p. 11, 2013. doi:10.1155/2013/136864.

[341] M. Guerini, O. Stock, and M. Zancanaro, "A taxonomy of strategies for multimodal persuasive message generation," *Applied Artificial Intelligence*, vol. 21, no. 2, pp. 99–136, 2007. doi:10.1080/08839510601117169.

[342] G. Soosalu, S. Henwood, and A. Deo, "Head, Heart, and Gut in Decision Making: Development of a Multiple Brain Preference Questionnaire," *SAGE Open*, vol. 9, no. 1, 2019. doi:10.1177/2158244019837439.

[343] FurhatRobotics, "Remote API - Furhat Developer Docs," 2021. url: https://docs.furhat.io/remote-api/.

[344] O. P. John and S. Srivastava, "The Big Five Trait taxonomy: History, measurement, and theoretical perspectives.," *Handbook of personality: Theory and research*, 1999. url: https://psycnet.apa.org/record/1999-04371-004.

[345] E. Childs, T. L. White, and H. De Wit, "Personality traits modulate emotional and physiological responses to stress," *Behavioural Pharmacology*, vol. 25, no. 5-6, pp. 493–502, 2014. doi:10.1097/FBP.0000000000000064.

[346] M. Naber, S. Frassle, U. Rutishauser, and W. Einhauser, "Pupil size signals novelty and predicts later retrieval success for declarative memories of natural scenes," *Journal of Vision*, vol. 13, no. 2, pp. 11–11, 2013. doi:10.1167/13.2.11.

[347] D. G. Cope, "Methods and Meanings: Credibility and Trustworthiness of Qualitative Research," *Oncology Nursing Forum*, vol. 41, pp. 89–91, jan 2014. doi:10.1188/14.ONF.89-91.

[348] P. Galdas, "Revisiting Bias in Qualitative Research," *International Journal of Qualitative Methods*, vol. 16, p. 160940691774899, dec 2017. doi:10.1177/1609406917748992.

[349] M. Belkaid, K. Kompatsiari, D. D. Tommaso, I. Zablith, and A. Wykowska, "Mutual gaze with a robot affects human neural activity and delays decision-making processes," *Science Robotics*, vol. 6, sep 2021. doi:10.1126/scirobotics.abc5044.

[350] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell, "A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception," *Computer Graphics Forum*, vol. 34, no. 6, pp. 299–326, 2015. doi:10.1111/cgf.12603.

[351] A. Moon, M. Zheng, D. M. Troniak, B. A. Blumer, B. Gleeson, K. MacLean, M. K. Pan, and E. A. Croft, "Meet me where i'm gazing: How shared attention gaze affects human-robot handover timing," in *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 334–341, IEEE Computer Society, 2014. doi:10.1145/2559636.2559656.

[352] D. Pasquali, A. Sciutti, and F. Rea, "Toward enabling iCub to detect lies in everyday life," in *I-RIM 2021*, oct 2021. doi:10.5281/ZENODO.5525169.