Individual and Cultural Differences in the Adoption of the Intentional Stance towards robots

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Science and Engineering

2021

Serena Marchesi
Department of Computer Science

Contents

Word Count: 64,894

List of Figures

List of Tables

List of publications

Publication 1: **Marchesi, S.**, Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., and Wykowska, A., (2019). Do we adopt the intentional stance toward humanoid robots?. Front. Psychol., vol. 10, no. MAR, 2019, doi: 10.3389/fpsyg.2019.00450. Materials and datasets are available at https://osf.io/pe2z3/.

Publication 2: **Marchesi, S.**, Spatola, N., Perez-Osorio, N. and Wykowska, A., (2021). Human vs humanoid. A behavioural investigation of the individual tendency to adopt the intentional stance, ACM/IEEE Int. Conf. Human-Robot Interact., pp. 332–340, 2021, doi: 10.1145/3434073.3444663. Materials and datasets are available at https://osf.io/65bn2/.

Publication 3: **Marchesi, S.**, Perez-Osorio, J., De Tommaso, D. and Wykowska, A., (2020). Don't overthink: Fast decision-making combined with behaviour variability perceived as more human-like, 29th IEEE Int. Conf. Robot Hum. Interact. Commun. RO-MAN 2020, pp. 54–59, 2020, doi: 10.1109/RO-MAN47096.2020.9223522. Materials and datasets are available at https://osf.io/3fznv/.

Publication 4: **Marchesi, S.**, Bossi, f., Ghiglino, D., De Tommaso, D. and Wykowska A., (2021). I Am Looking for Your Mind: Pupil Dilation Predicts Individual Differences in Sensitivity to Hints of Human-Likeness in Robot Behaviour, vol. 8, no. June, pp. 1–10, 2021, doi: 10.3389/frobt.2021.653537. Materials and datasets are available at https://osf.io/s7tfe/.

Publication 5: **Marchesi, S.**, De Tommaso, D., Pérez-Osorio, J. and Wykowska, A., (in press). Belief in sharing the same, phenomenological experience increases the likelihood of adopting the intentional stance towards a humanoid robot. Materials and datasets are available at https://osf.io/xnm5c/.

Publication 6: **Marchesi, S.**, Spatola, N. and Wykowska, A., (submitted, currently under revision). The mediating role of anthropomorphism in the adoption of the intentional stance towards humanoid robots. Materials and datasets are available at https://osf.io/3mzpj/.

Abstract

The influence of artificial agents on our lives is continuously changing. Recent literature shows that the presence of such agents will be more pervasive in our social environments, leaving the open question of whether humans will perceive these agents as possible social partners. Hence, a novel challenge arises in understanding whether humans will deploy socio-cognitive mechanisms similar to the ones activated in human-human interaction to understand and predict the behaviour of these new entities. The present thesis aimed at investigating how humans predict artificial agents' behaviour based on the adoption of the intentional stance, a philosophical framework developed by Daniel Dennett to explain how humans predict and interpret others' behaviour. To this aim, we first investigated how the intentional stance differs from other concepts addressing the attribution of intentionality to others. We then developed a novel tool, the InStance Test, to assess the adoption of the intentional stance towards a humanoid robot, namely the iCub robot. A variation of InStance Test was then used to explore the behavioural correlates of the adoption of the intentional stance by recording participants' response times. Next, we designed a series of more interactive experiments, where the InStance Test was administered pre-and post- interaction with the embodied humanoid robot iCub. This allowed us to assess whether observing an embodied humanoid robot exerting behaviours with different degrees of human-likeness would modulate the individual tendency to adopt the intentional stance. In addition, within these experiments we also explored the influence of individual differences on personality traits, perception of robotic agents and physiological (i.e., pupillometry) correlates of modulating the stance adoption. Finally, since humans are constantly embedded in a cultural context on both a global and a local base, growing up in different cultural environments may influence also the socio-cognitive mechanisms that we employ in social interactions. Therefore, we explored the influence that culture might have on the adoption of the intentional stance towards humanoid robots. Results from the present thesis show that humans differ in their individual tendency to adopt the intentional stance towards humanoid robots. Interestingly, this individual bias has behavioural and physiological correlates, that can predict the stance adoption and the sensitivity to human-like behaviours deployed by the iCub robot. Moreover, results show that the adoption of the intentional stance is also influenced by individual differences in cultural values, personality, and tendency to anthropomorphize non-human agents. I conclude that to allow a smoother integration of these new agents in our social environments, we should take into consideration humans' social cognition and individual differences to design robots that will socially attune with us.

Declaration of originality

I hereby confirm that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

There is no gate, no lock, no bolt that you can set upon the freedom of my mind.

A Room of One's Own (1929)
Virginia Woolf

Benjamin Jabituya: ”Who is knowing how to read the mind of a robot?”

Short Circuit (1986)

Acknowledgements

This thesis has received support from the European Research Council under the European Union's Horizon 2020 research and innovation program, ERC Starting Grant ERC-2016-StG-715058, awarded to Agnieszka Wykowska. The content of this thesis is the sole responsibility of the author(s). The European Commission or its services cannot be held responsible for any use that may be made of the information it contains.

I first decided that I would have done a Ph.D. during my Bachelor degree, in 2013. Since then, and with that aim in my mind, I worked really hard to get here. Nevertheless, there is a group of people to whom I would like to express my deep gratitude and love: please, know that without you, this path would have been impossible.

First and foremost, my gratitude goes to my parents Claudia and Sergio and to my sister Monica. Your love and support made me stronger and bolder.

The work presented in this thesis is the final step of a path that started in 2017 with a younger and more naïve version of me, who wrote an email to Prof. Agnieszka Wykowska, asking her for an internship in her research group at IIT. Thank you, Agnieszka, for believing in that version of me. Thank you for your constant (professional and personal) guidance, supervision and hard work to make me a better researcher. Thank you to Prof Angelo Cangelosi, for always being present, despite the distance, and for all your on point comments on this Ph.D. project. Thank you to Dr. Wu Yan, for all your help and efforts to make "Singapore" happen.

Thank you to my partner, Mariacarla, for being my rock, my home, and the music of my heart. You always support me, and push me to be the best possible version of myself. Without you, these years would have definitely been darker and sadder.

I would like to thank Clarissa, for being my sister of choice. I want to thank you for always being by my side throughout these crazy 12 years. Despite the geographical distance, our friendship always brings joy and love.

Thank you to Giorgia, for all the horror movies, true crimes podcasts, dinners, and chats we had in these years.

I would like to express my thanks to Georgia, for her dear love and for always having a place for me at her dinner table, since 2009.

Thank you, Sara, for your friendship and all the "piedo" videos.

My gratitude and love goes to "Le Gaglioffe" Davide, Cecilia, and Lorenzo. You are the best Ph.D. fellows one could ever wish for. Thank you for your love and for sharing the "Ph.D. Days" with me. I am humbled to call you my friends.

Thanks to Ingrid, my Habibi, for all the tandems and the chats we had.

I would like to thank all the Post-Docs of the S4HRI research group. Especially, thank you Jairo, for being my advisor; Cesco, for being my very first advisor and for having always time for a chat. Thank you, Francesco aka Doc, for everything you thought me and for your lovely friendship. Thank you to my dear Kyveli for always finding the time to share a social moment with me. Thank you, Aziz, for all the memes. Thank you, Nico, for all your kindness and knowledge. Thank you to Maria, for your silliness and for your help with LaTeX. My gratitude to Davide De Tommaso for always helping me to implement the craziest ideas. Thank you to Laura, as well, for teaching me all I know about Adobe and video recording.

Thank you to the "Manfredi Gang" and to Luca and Francesca for all the parties.

I would like to express my thanks to the Italian Crew in Singapore, for making these 8 months as much Italian as possible. My 30th birthday included. A special mention should go to Osaka, for being the cutest and the fiercer feline I know.

I beg your pardon if you happened to be in my life during this crazy journey and I forgot to mention you in this list; to you, goes my gratitude and love anyway.

Chapter 1

Thesis Structure

The present thesis is divided into five parts, and it is structured as follows:

- **Part** 1**: Attributing intentionality to robots: Dennett's stances in Human-Robot Interaction**

In **Chapter 2**, I provide an overview of the theoretical background of the present work. I introduce the problem of adopting the intentional stance in Human-Robot Interaction and the motivation of the present work. Finally, I state the research questions and the strategies I followed to address them.

- **Part** 2**: Attributing intentions to robots: the development of the InStance Test**

Chapter 3 presents a manuscript of *(Publication 1)* in which the theoretical framework has been described. Finally, I present the InStance Test, a novel tool developed to assess the adoption of the intentional stance towards a humanoid robot.

Chapter4 presents a manuscript from *(Publication 2)* which investigates the indices of the differential effect in adopting intentional stance between a human and a robot. I adapted the InStance Test to create a version where I could collect participants' response times. As a control, I created a version of the test with a human agent instead of the humanoid robot. Moreover, I empirically explored the relationship between anthropomorphism and the adoption of the intentional stance towards robots.

- **Part** 3**: How individual differences shape the adoption of the intentional stance towards an embodied humanoid robot**

Chapter 5 presents a manuscript from *(Publication 3)*, which describes the first experimental design involving the embodied robot iCub. I aimed at exploring whether it is possible to modulate the adoption of the intentional stance towards a humanoid robot by modulating the degree of human-likeness of the robot behaviour. Moreover, in this context, I explored the role of individual differences in personality traits or attitudes towards robots in the adoption of the intentional stance.

Chapter 6 presents a manuscript from *(Publication 4)* where I describe a second experiment with the embodied iCub robot. I present an experimental design aimed at investigating the physiological correlates of differential effects related to a higher or lower degree of adoption of the intentional stance, by means of different behaviours implemented on the robot (namely a human-like and a machine-like behaviour).

Chapter 7 presents a manuscript from *(Publication 5)* which reports the last series of experiments involving the embodied iCub robot. To extend the knowledge of the results reported in Chapter 7, I modulated the behaviours of the robot to the extremes of human-likeness and machine-likeness to investigate how participants' individual likelihood of adopting the intentional stance would change after having shared a familiar context with the iCub robot.

- **Part** 4**: The influence of culture on the attribution of intentions to robot**

Chapter 8 presents a manuscript from *(Publication 6)* where I describe the last experiment of this thesis. I report a study aimed at investigating the relationship between cultural differences in anthropomorphism and the adoption of the intentional stance. I administered a battery of tests and questionnaires to build a path model and a mediation model to unfold the role of these variables in the context of Human-Robot Interaction.

- **Part** 5**: Project results**

Chapter 9 concludes the present thesis and draws a summary of the overall empirical results and theoretical insights obtained from the presented works, highlighting also some limitations that hopefully will lead to future directions in the research on this topic.

The main narrative that was followed in structuring the present thesis aimed at presenting the process I followed to operationalize a philosophical concept, such as the adoption of the intentional stance. The present thesis collects the results of my journey towards this aim, and as every journey, it starts from the beginning, with the introduction of the theoretical background at the foundations of the present work. Namely, first I present Daniel Dennett's intentional stance account, its relation with other major accounts about human social cognition and why it is relevant for the field of Human-Robot Interaction (HRI, Chapter 2). After the introduction of the intentional stance theory, I report the creation and preliminary validation of a tool, the Intentional Stance Test (IST) to assess the adoption of the intentional stance towards a humanoid robot (Chapter 3), the exploration of the behavioural correlates of the adoption of such stance investigated with the IST (Chapters 4) and the empirical implementations of the IST in empirical setups to assess people's stance adoption towards a real and embodied humanoid robot (Chapters 5, 6 and 7). Finally, Chapter 8 reports a first investigation of how the IST can be used to investigate the influence of culture on the attribution of mental state to humanoid robots.

Part I

Attributing intentionality to robots: Dennett’s stances in Human-Robot Interaction

Chapter 2

Introduction

2.1 Theoretical background

As intrinsically social animals (Ebstein et al., 2010; Tomasello et al., 2005), humans developed the ability to predict, interpret and understand other humans' behaviours (Baron-Cohen et al., 1999; Gallotti and Frith, 2013). This ability is acquired during the first years of life (around 2-3 years old), and typically developed children of 5-6 years old are able to adopt a "mental perspective" of others to correctly infer motives and intentions of others' behaviours (Perner, 1991; Perner and Wimmer, 1985; Wimmer and Perner, 1983). Meltzoff suggests that humans learn to understand others through the understanding of themselves and subsequently perceiving (and explaining) others as similar to oneself - the "like me" hypothesis. The author proposes that this mechanism is the foundation of humans' social cognition. This early understanding provides infants (and, later in life, adults) a framework to interpret others' behaviours (Meltzoff, 2007). In the last five decades, philosophers and cognitive science researchers explored different approaches to investigate and explain the processes of humans' social cognition involved in the acts of predicting and interpreting other agents' behaviours. One of the main accounts is also known as Theory of Mind (ToM) or mindreading (Baron-Cohen et al., 1999 for a review see Apperly and Butterfill, 2009; H. L. Gallagher et al., 2002). In line with the ToM account, some authors posit that the most efficient strategy to predict and interpret humans' behaviour (others' and one's own) is to refer to underlying inner mental states, such as desires, intentions, and beliefs (Dennett, 1989; Fletcher et al., 1995; Frith and Frith, 2012; Gallotti and Frith, 2013). Interestingly, results from empirical investigations revealed that the ascription of mental states is possible towards humans and non-human systems (Apperly and Butterfill, 2009; Butterfill and Apperly, 2013; Happé and Frith, 1995; Heider and Simmel, 1944). Among the ToM accounts, other researchers developed different takes to human social cognition and to the "problem of the other minds", the most relevant of which are the Simulation Theory and the Phenomenological approach (S. Gallagher and Zahavi, 2020; Gallese, 2005).

The purpose of the present Chapter is to describe the major theoretical and empirical approaches to human social cognition. Once the theoretical background is presented, it is crucial to introduce and justify the relevance of Daniel Dennett's concept of "intentional stance" - the concept that lies at the foundation for this entire Ph.D. project. In the following paragraphs, I will present the Theory-Theory (TT), the Simulation Theory (SM) and the Phenomenological approach to social cognition. I will then introduce the concept of the intentional stance as described in Daniel Dennett's account. I will then discuss how the intentional stance theory relates to other accounts present in philosophy and social cognition that address the question of intentionality and, more generally, how do humans understand others. Moreover, in Part 1, Chapter 3, I will further discuss the intentional stance and differentiate it from other related concepts, namely the Theory of Mind (ToM). I will argue that these definitions are not addressing the same constructs and processes (Marchesi et al., 2019).

I will then introduce a second pivotal concept to the present thesis: anthropomorphism. Such construct is fundamental to understand how humans might attribute typically humans' characteristics to artefacts such as humanoids robots (Airenti, 2018; Epley et al., 2007; Waytz, Morewedge et al., 2010). I will present definitions of anthropomorphism that are relevant for the following chapters and discuss how and why anthropomorphism is relevant for the field of HRI.

2.2 How to approach social cognition: Theory of Mind and Phenomenological takes on understanding others

Human social cognition has been studied extensively in the last century. A number of new theories and techniques have been introduced along the path, to allow casting a light on how we, as humans, understand others. Premark and Woodruff (Premack and Woodruff, 1978) condensed the process of understanding other humans in the expression "Theory of Mind" (ToM). More specifically, philosophers and cognitive scientists use this expression to convey the human ability to attribute, interpret and explain their and others' behaviours in terms of mental states. The contemporary debate sees two major accounts that explain how these processes are possible: the Theory-Theory account (TT) and the Simulation-Theory (ST). Moreover, phenomenology attempted to address the question of "the other minds" as well, providing a major contribution to the debate about how we understand each other. In the next paragraphs, I will describe the TT and ST accounts of social cognition and mention the phenomenological approach to the same problem. Eventually, I will explain Dennett's position in the debate and justify the adoption of the intentional stance account as theoretical background for the present thesis.

2.2.1 Theory-Theory and Simulation-Theory explained

As previously mentioned, the contemporary debate about human social cognition is approached mostly from two perspectives: Theory-Theory (TT) and Simulation-Theory (ST) (and their variants and intersections). For the sake of clarity and simplicity, I will describe the prevalent and mainstream claims of TT and ST respectively.

Given the premise that ones' mental states are not directly accessible to external observers (Johnson, 2000), TT claims that to be able to understand others we appeal to the so-called “folk psychology” (Goldman, 1993; Stich and Nichols, 2003), the common sense that can provide explanations of why people are behaving the way they are. Thus, the way we explain others is driven by causal theories that we build making inferences based on the available information (Leslie, 1987; Baron-Cohen, 1999; for a review see Apperly, 2008). Although this assumption is shared by all theory-theorists, there are differences on how the causal theory is generated: some authors (e.g., Carrutherts, Barron-Cohen, Fodor) have a top-down approach and, thus, think that it is innate, modularized and develops along the course of the toddler development; others (e.g., Gopnik, Wellman, Meltzoff) argue that it is the results of a bottom-up learning and acquired via everyday experience since the day we are born (Apperly, 2008; S. Gallagher and Zahavi, 2020). Nonetheless, it is important to highlight that the appeal to folk-psychology and to the theoretical knowledge about others' minds is an automatic process that does not necessarily require to emerge at the conscious level.

Albeit agreeing on the premise that mental states are not accessible, ST argues that the biological evolution of the human mind implies that involving mental states are similar across minds (see Apperly, 2008), thus we do not need an organized and pseudo-scientific theory to understand others, instead we can use our mind as a model to simulate what “the other minds are like”. Following this, ST theorists are divided on how the simulation should take place: either explicitly or implicitly. Goldman (Goldman, 2005) argues that to understand others, we use our imagination to put ourselves in the others' shoes, and simulate “how it would be like to be them”. Here's how Goldman describes the explicit simulation process:

> ”First, the attributor creates in herself pretend states intended to match those of the target. In other words, the attributor attempts to put herself in the target's “mental shoes”. The second step is to feed these initial pretend states [e.g. beliefs] into some mechanism of the attributor's own psychology [...] and allow that mechanism to operate on the pretend states as to generate one or more new states [e.g. decisions]. Third, the attributor assigns the output state to the target [...].” (Goldman, 2005, pp. 80-81)

This approach has been criticized with two main arguments: first, the idea of and explicit and conscious simulation process: some authors (e.g., Dennett, 1989) argue that the simulation must be organized into constructs that can be defined as “theories”. By putting ourselves into the “mental shoes” of the others, we would be somehow “theorizing” about them, making the claims of explicit ST closer to the ones made by TT theorists. A second criticism is that an explicit process of simulation could allow adopting a “first-person” imaginative process, therefore we would not be really understanding others, instead we are projecting ourselves onto the other and, thus, understanding ourselves in their situation (S. Gallagher and Zahavi, 2020).

Recently, the neuroscientific description of the mirror neuron system lead some authors to consider a more implicit version of the Simulation Theory. Empirical results showed activation of the motor systems in our brains when 1 – we are engaging in the action; 2 – when we are observing the action; 3 – when we are imagining ourselves or others engaging in the action; 4 – when we plan to imitate the action (Butterfill and Sinigaglia, 2014; P. F. Ferrari and Rizzolatti, 2014; Grezes and Decety, 2001; Rizzolatti and Craighero, 2004; Rizzolatti et al., 2009). In this version, ST postulates that the simulation is an enactive process (i.e, sensorimotor) that happens at a sub-personal, implicit and almost automatic level. Thus, the neural activations are the covert simulation of the other's action (Gallese, 2001, 2005). The first objection to the implicit version of the ST, is that it still implies a first-person perspective. The simulation that happens in my motor system is not extendable to what is happening in the motor system of another person. Nevertheless, some authors argue that since the motor neuron system activates regardless of the agent performing an action (myself or another person), the simulation is about the action per se, not the agent performing it. Gallagher and Zahavi (2020) argue that since simulation is an enactive process, the motor resonance is not the initiation of the simulation, rather than a part of the enactive, intersubjective process of seeing (i.e., perceiving) the action as the others' meaningful action. That is, we do not have to simulate the action, we directly perceive it.

Although initially TT and ST were considered mutually exclusive, recently some authors have made attempts to integrate the two frameworks (Currie, Ravenscroft et al., 2002; Nichols and Stich, 2003). The integrative approaches argue that experimental evidence are in favour of both TT and ST, leading to the conclusion that both are necessary to attribute, read and interpret the mental state of others (for a review see Apperly, 2008; Spaulding, 2015).

2.2.2 The Phenomenological approach explained

Although TT and ST are debating about the architecture underlying the sub-personal processes involved in humans' social cognition (i.e., theories or simulation about others' mental states), most of the accounts would agree that the processes related to mindreading may occur both subconsciously or emerge to the conscious level, are pervasive and the primary way we understand others (Spaulding, 2015; Carruthers, 2009; Gopnik, 1999). This crucial point always excluded phenomenology from the debate about social cognition, as by definition, phenomenology investigates only the conscious experience from a "first-person" perspective, and thus it cannot really contribute to the debate about others' minds (Dennett, 1989; Spaulding, 2015). In particular, Dennett argues that when we investigate social cognition applying the phenomenological methodologies, we are actually investigating our own mental states rather than investigating the process of understanding others', falling in a methodological solipsism (Dennett, 1989; S. Gallagher and Zahavi, 2020). It is important to highlight that almost every TT, ST and integrative approaches to mindreading do not postulate that neither the theories nor the simulations about others' mental states are necessarily consciously accessible, and consequently, not they cannot be accessible to the phenomenological experience (Spaulding, 2010, 2015). It follows, according to mindreading proponents, that the phenomenology cannot be considered as a reliable (and, thus, relevant) methodology to the social cognition debate (Spaulding, 2015).

Nonetheless, some authors argue for a relevant role of phenomenology in humans' social cognition (De Jaegher and Di Paolo, 2007; Fuchs, 2013; S. Gallagher, 2012; Ratcliffe, 2006; Zahavi, 2011). According to these authors, there is phenomenological evidence to support (or go against) other theories of social cognition. The most relevant phenomenological accounts of social cognition are the Interaction Theory, Embodied Cognition and Enactive Cognition. Although they vary in the definition of some notions such as "social interaction", they all agree on the main argument that what should be highlighted is a non-mentalizing, embodied and perceptual approach to the problem (Gallagher and Zahavi, 2020). Gallagher (2012, 2011, 2001) argues that, although phenomenology cannot directly discuss the sub-personal processes of social cognition, it can be an indirect evidence of such processes. That is because, Gallagher (and most phenomenologists) posits that our social interactions do not happen in a "third-person" of view (i.e., theorizing about their mental states), nor are detached from our bodies. We understand others by seeing them and directly perceiving their actions through our bodies and our self experience (Becchio et al., 2018; Fuchs and De Jaegher, 2009; S. Gallagher, 2012; S. Gallagher and Zahavi, 2020). For example, Zahavi (Durt et al., 2017; Kriegel, 2020) introduced the notion of "minimal experiential selfhood" as the primitive perspective we adopt to interpret our experience and others. As the most extensive experience I have in life is to be myself, the minimal experience selfhood is the pre-reflective awareness of being conscious that lies at the foundation of every experience I have in life. This construct is inertly subjective and, according to Zahavi, is the precondition for any type of experience. As a result, the social dimension of experience (and thus the understanding of others) has to emerge from the minimal experiential selfhood of each individual.

Finally, the last account that I will present as relevant for the present thesis is the "we-intentionality" (WI). The WI account considers that humans are intrinsically social animals (Ebstein et al., 2010; Tomasello et al., 2005). More specifically, according to Tomasello, what makes humans able to coordinate and create sociality far more complex than other primates, is the ability to jointly direct goals, actions, and intentions. This ability to form a shared agency lets us collectively intend to reach a common goal (Ciaunica and Fotopoulou, 2017; Dewey et al., 2014; Knoblich and Sebanz, 2006; Pacherie, 2014; Sebanz et al., 2006; Tomasello and Rakoczy, 2003). Higgins (2020) proposes to include, along with the primitive of the minimal experience selfhood discussed by Zahavi, the "minimal relational self". Higgins argues that social interaction is one of the first experiences we have in life, and that this develops along with the internal perception about ourselves and our consciousness. Therefore, according to Higgs, we use social interaction to define ourselves through others, as much as we do when we use our own perception about the experience (Higgins, 2020).

Although the philosophical and psychological accounts that investigate social cognition are numerous, the previous paragraphs aimed at introducing the most relevant ones and present an overview of the theoretical discussion about how humans understand each other so well. Each account attempted to find empirical evidence to support their claims, and to some extent there is a growing corpus of experimental results that supports each account. Although no definitive result can be drawn at this moment regarding which of the presented accounts for empirical results best (for a thorough review, see Apperly, 2008; Becchio et al., 2018; S.

Gallagher and Zahavi, 2020; Schurz et al., 2021), they all contribute to the field by addressing and operationalize complex philosophical problems. All in all, that is already an achievement in itself that will help to understand which empirical methodologies can help us to unravel what is human social cognition, how it works and how (if at all) it changes when we meet different social agents.

In the next paragraph, I will introduce the concept of the intentional stance, as described by Daniel Dennett (Dennett, 1971, 1981). I will then introduce the concept of anthropomorphism and, finally, justify why this, out of all the accounts in social cognition, lies at the basis of the present Ph.D. project.

2.3 Dennett's stances

Philosopher Daniel Dennett describes three main strategies, or "stances", that humans might adopt to explain and predict the observed behaviour of a system. Each stance refers to different levels of abstraction: 1 – the first stance, called physical stance, makes reference to the physical domain. In this case, the observer would predict the behaviour by means of the laws of physics (i.e., the prediction of the trajectory of a ball). 2 – The second stance, called the design stance, makes reference to how the system was built to function. The interpreter here will use their knowledge about the system's functional design to predict the behaviour. For example, I can predict my car to stop if I push the brake. 3 – The third stance, called the intentional stance, makes reference to the above-mentioned mental states and beliefs of the agents. The interpreter in this case will ascribe to the system desires, beliefs, and propositional states in general to predict the observed behaviour. For example, I can predict that Davide wants to visit Tenerife because he thinks it is a beautiful island (intentional stance). Dennett highlights how the three stances have different epistemological requirements. In fact, while the first two can be applied to any system, regardless of any rationality assumption, the third is stricter on this claim. Dennett (1989) describes the process of adopting the intentional stance as follows: initially, the observer decides to treat the agent as rational. Then, the observer interprets the agent's mental states (i.e., desires or beliefs that the agent might have). Finally, based on these assumptions, the observer predicts that the agent will act pursuing its goals based on its mental states. Thus, by adopting the intentional stance, we assume that the system's behaviour is the most rational one that the agent can exert in that context, given the ascribed beliefs, desires, and constraints. In other words, we have to assume that we are facing a rational system (Dennett, 1989), but crucially, Dennett highlights that any system can be treated as rational and intentional. However, only for truly intentional agents (that Dennett calls "true believers"), the intentional stance reveals to be the most efficient strategy. For other types of agents or systems, switching to a different stance (i.e., the design or the physical stance) would be more efficient.

From this description, it is clear that Dennett rejects the phenomenological accounts of social cognition in favour of an approach that falls closer to the Theory-Theory explanations of the attribution of mental states to others. Nevertheless, Dennett's stance can be distinguished from other accounts in Theory-Theory (such as the Theory of Mind (ToM). Specifically, Dennett's intuition on the intentional stance merges two epistemological views: Dennett is a *realist* about beliefs, meaning that he considers beliefs as objective phenomena. He also argues that beliefs can only be interpreted and confirmed by "the eye of the beholder" who adopts a predictive strategy (i.e., one of the stances). This latter formulation, makes Dennett an *interpretationist* (Dennett, 1989). In other words, although you do have mental states, to understand them, I observe your behaviour by adopting a successful predictive strategy (that, Dennett argues, in the case of humans is the intentional stance). Crucially, in this case, the false belief attribution leads the strict realist approach to attribute a false belief to the agent (Baron-Cohen, 1997), while the interpretationist approach allows adopting the intentional without making assumptions on their epistemological status. Further discussion about the comparison between ToM and intentional stance can be found in Chapter 3.

Lastly, it is worth to mention that recently Robbins and Jack attempted to expand the discussion about the intentional stance by introducing the "phenomenal stance" (A. I. Jack and Robbins, 2012; Robbins, 2006). According to the authors, Dennett's intentional stance does not address the hard problem of consciousness raised by David Chalmers (Chalmers, 2007), avoiding answering how do we attribute consciousness, emotions, and inner experience to another agent. To address this problem, Robbins and Jack argue that we adopt the phenomenal stance, and that psychopathy and neurodevelopmental disorders may represent a double dissociation between phenomenal (in the first case) and intentional (in the second case) stances. Interestingly, one could hypothetize that, while in the case of other human agents the dissociation might be present only in specific cases (i.e., psychopathy), in the case of an artificial agent it might be more prominent (Huebner, 2010).

Before discussing the relevance of the intentional stance account for Human-Robot Interaction, it is crucial to introduce a second pivotal construct for the present Ph.D. thesis: anthropomorphism. Thus, in the following paragraph, I will introduce the most pertinent definition of anthropomorphism and its declinations.

2.4 Defining anthropomorphism

Anthropomorphism is a broad concept that has been applied to different research fields (from literature to natural sciences, from art to psychology). Although addressing different research problems, anthropomorphism can be generally defined as the attribution of human traits, emotions, or intentions to non-human entities (Epley et al., 2007; Waytz, Cacioppo et al., 2010). Epley and colleagues (2007) argue that anthropomorphism can be elicited by three conditions:

1. **Elicited agent knowledge**: the availability of characteristics that activates knowledge that we have about humans (i.e., physical characteristics);
2. **Sociality**: humans' social need for connection and interaction;

3. **Effectance**: individual traits and differences in interacting with and understanding one's environment.

There is evidence from developmental psychology that humans show this tendency since early age, associated with a high tendency to apply egocentric reasoning (i.e., taking for granted the others have the same experiences, perceptions, and thoughts as we do) (Barrouillet, 2015; McLeod, 2007). Thus, in the context of cognitive and developmental psychology, it is generally agreed that anthropomorphism is a pervasive cognitive bias (Airenti et al., 2019; Andrews and Huss, 2014; Dacey, 2017; Epley et al., 2007; Jones, 2021), and evidence from neuroscience seems to confirm that anthropomorphism emerges as an automatic process (Mitchell et al., 1997; Urquiza-Haas and Kotrschal, 2015).

Given this definition, for the purposes of the present thesis, anthropomorphism has been investigated in two slightly different ways:

1. **Anthropomorphism as a general cognitive bias**: given the results presented in Chapter 2 concluding that anthropomorphism could play a role in the adoption of the intentional stance towards humanoid robots, I decided to investigate this role as an individual cognitive bias in Chapter 3. Within this Chapter, the investigation revolved around the definition of anthropomorphism as a general cognitive bias towards non-living entities. To do so, I administered the Individual Differences in Anthropomorphism Questionnaire (IDAQ) developed by Waytz, Cacioppo and Epley (Waytz, Cacioppo et al., 2010). The IDAQ questionnaire asks to attribute a given set of capacities to different non-humans agents, such as cows, trees, and clouds. As a result, the IDAQ explores the tendency to attribute human-like characteristics to a variety of non-human agents, ranging from living to non-living systems.

2. **Anthropomorphism as an actionable concept towards robots**: in Chapters 5, 7 and 8 I decided to investigate the concept of anthropomorphism applied to Human-Robot Interaction. The major difference compared to the previous definition is that in this latter case, I explored humans' anthropomorphic attributions directed onto a specific artificial agent, such as a humanoid robot. That is because, as Epley and colleagues argued, one of the three factors that influence the tendency to attribute human characteristics to non-human agents is the availability of physical features that elicit the knowledge we have about humans. Thus, one can hypothetise that we might find it easier to anthropomorphise humanoid robots due to their shapes (Epley et al., 2007; Wiese et al., 2017). For this reason, I employed two different questionnaires specifically developed to investigate the anthropomorphic bias towards robot: 1- the Robotic Social Attributes Scale (RoSAS, Carpinella et al., 2017) and 2- the Human–Robot Interaction Evaluation Scale (HRIES, Spatola and Wykowska, 2021). Although both the RoSAS and the HRIES are multi-component approaches to the investigation of anthropomorphic attribution to robots, the RoSAS questionnaire emphasises the social aspect of the interaction with the robot, while the HRIES aims at assessing the whole quality of the interaction, including, for example, the level of perceived agency of the robot during the interaction.

The investigation of the influences of anthropomorphic attributions (both at the general and particular level) on the adoption of the intentional stance towards humanoid robots have implications that can span from the theoretical implications (i.e., a better understanding of the relationship between these two concepts) to applicative suggestions to engineers (i.e., suggestions on how to design robots that can evoke more or less intentional stance depending on their purpose). Thus, in the next paragraph, I will explain the relevance of these concepts to the field of HRI.

2.5 Dennett's stances and the human-model applied to artificial agents

The motivations of this Ph.D. project are to examine factors that contribute to a complex phenomenon such as the adoption of the intentional stance towards a new type of social system: artificial agents. More specifically, the factors of interest are: 1- the behaviours of the agents, and how they contribute to the adoption of the intentional stance and 2- the influence of cultural and individual differences of the observers in the likelihood of adopting the intentional stance.

Artificial agents, such as humanoid robots, will soon populate our society, and a portion of them will be designed to engage in social interaction with us (Prescott and Robillard, 2021; Samani et al., 2013). Humanoid robots are of particular interest because they represent a "middle point" in the continuum between physically embodied humans and disembodied artificial agents such as computer algorithms. In fact, in the context of investigating the adoption of intentional stance towards artificial agents, special interest has been given to humanoid robots, since they represent entities that are somewhat "in-between", as hypothesized by the New Ontological Category theory (NOC, Kahn et al., 2011; Kahn and Shen, 2017). On the one hand, the design of humanoid robots can evoke and fulfil the human need for "socialness", especially if they are designed as social companions in different contexts of our lives, from elderly care, to educational context and healthcare applications (Epley et al., 2007, Birks et al., 2016; Tapus et al., 2007; for a review see Wykowska, 2020). Indeed, their embodied (and potentially social) presence might evoke anthropomorphism (Airenti, 2018; Epley, Akalis et al., 2008; Epley et al., 2007; Złotowski et al., 2014). As previously introduced, anthropomorphism, in the definition of Epley and colleagues (2007) is the process of ascribing human characteristics to non-human agents. According to the authors, this process can be elicited by three factors: 1- the physical characteristics of the agent that can activate knowledge and heuristics related to humans; 2- the possibility of fulfilling the human need for connection and sociality; 3- individual traits of the human observer. The first two factors depend on how much the non-human agent will evoke the so-called "human model" (Wiese et al., 2017). On the other hand, these agents are merely artefacts, such as a car or a fridge. This awareness may evoke the adoption of the design stance. In addition to the characteristics of the humanoid robots, individual traits of the observer play a role in anthropomorphism (Spatola and Wykowska, 2021). Several studies in the field of Human-Robot Interaction (HRI) showed that humans tend to anthropomorphize agents, by attributing them different characteristics that are typically human (Bartneck et al., 2009; Dacey, 2017; Fink, 2012; Złotowski et al., 2014, for a meta-analysis on the effectiveness of anthropomorphism in HRI see (Roesler et al., 2021)). Moreover, Ciardo and colleagues (Ciardo et al., 2022) reported that coordination in joint task with an embodied robot is affected by the human-likeness expressed by the agent both in the behaviours and in the appearance. Additionally, the authors report also that these factors have an influence on the social inclusion of a humanoid robot. Hence, if we consider humanoid robots as potentially social systems to which we can attribute humans' characteristics, it becomes crucial to understand under which circumstances (if any) humans would spontaneously adopt the intentional stance towards humanoid robots and to what extent. That is because, although they are just artefacts, the elicited anthropomorphism might facilitate the adoption of the intentional stance and, therefore, the interpretation of the robot behaviour with reference to mental states.

2.5.1 The influence of cultural differences in Human-Robot Interaction

In the context of analysing the factors that might play a role in the adoption of the intentional stance towards robots, it is important to point out that humans are embedded in different cultural environments. This aspect can contribute to a different perception of robots, and therefore to a different likelihood of adopting the intentional stance towards them. Indeed, Varnum and colleagues report that people from the Eastern part of the world adopt a more holistic approach and are more prone to cooperate compared to people from the Western world (Varnum et al., 2010). Other authors showed differences in cultural values, such as collectivistic and individualistic values, between Eastern and Western countries, leading to the conclusion that Collectivism and Individualism are "Cultural Syndromes" (for a review see Triandis, 1993). Cultural Syndrome can be defined when the underlying constructs (such as collectivism/individualism) are: 1 – organized around a theme, a value; 2 – the differences within-culture are smaller compared to the ones across cultures, and 3- a pattern is found between the emergence of these constructs and geographic regions (not necessarily defined by the borders of countries). Therefore, this shows that individualism and collectivism can coexist in the same country and be applied differently upon the situation. Cultural Syndromes are internalized throughout our development, leading individuals to use them as biases and heuristics in their day-by-day life. Each person has an individual tendency to be prone towards these constructs that influences their interactions and decisions; in other words, Cultural Syndromes shape the way we present ourselves to others (Bandura, 2002; Markus and Kitayama, 1991; Vygotsky, 1980). Given that literature proves that cultural differences influence humans' social cognition, researchers in Human-Robot Interaction investigated whether this influence would extend to robots (for a review, see Lim et al., 2020).

The present Ph.D. project will attempt to disentangle the relationship between anthropomorphism and the adoption of the intentional stance in contexts where the continuum between collectivistic and individualistic varies. More specifically, recent literature on cultural differences in HRI (Lim et al., 2020; Marchesi, Roselli et al., 2021) showed that both differences in the tendency to attribute a mind to a robot and in being more prone to collectivism influences the social inclusion of robots. Chapter 8 of the present thesis will explore the hypothesis that the higher tendency to have collectivistic values, the higher the anthropomorphic attributions, the higher the adoption of the intentional stance towards a humanoid robot.

2.6 Aims and Objectives

With these considerations, the present work aimed at investigating whether and when humans would adopt the intentional stance towards a humanoid robot. To this aim, I considered the cultural and individual differences in personality traits and anthropomorphism as possible influencing factors. I aimed at disentangling the theoretical and empirical relationship between these constructs by designing experiments tailored to unpack progressively the facets of each of the above-mentioned constructs.

2.6.1 Empirical strategy and research questions

To address these aims, I: **a.** conducted en extensive research literature to explore and extend the theoretical background on the adoption of the intentional stance towards robots; **b.** a novel developed tool to assess the adoption of the intentional stance was developed and compared it with pre-existing tools to assess the related concepts of Theory of Mind (ToM); **c.** I conducted experiments with the iCub robot to test possible modulation of the adoption of intentional stance in real scenarios and which factors could contribute to this modulation (i.e., different robot's behaviours, individual differences in personality, anthropomorphism); **d.** I further explored the theoretical framework by building a path model to include the role of culture in the adoption of intentional stance towards a humanoid robot.

Driven by the key aims of this project, the present work addresses the following research questions:

RQ1 - As intentional stance is a philosophical concept, how can We operationalize it and design empirical tests of the intentional stance? Is it possible to identify neural and behavioural markers of adoption of the intentional stance?

RQ2 – Can We modulate the likelihood of adopting the intentional stance towards a humanoid robot by manipulating the robot's behaviour?

RQ3 - Is the adoption of the intentional stance towards a humanoid robot be modulated by cultural differences and the individual tendency to anthropomorphize robots?

2.7 Research Plan

To address RQ 1, we proceeded as follows:

1- We investigated how the intentional stance differs from other similar (but, crucially, not identical) concepts, namely the Theory of Mind (Baron-Cohen et al., 1999) and anthropomorphism (Epley et al., 2007). This step allowed us to understand the relationships between these constructs and to formulate clearer experimental hypothesis;

2- Building on the knowledge obtained from point 1, we created a novel tool to assess the adoption of the intentional stance, the InStance Test (IST). The IST allowed us to investigate empirically the adoption of the intentional stance towards a humanoid robot;

3- As the final step, we investigated the behavioural and neurophysiological correlates of the adoption of the intentional stance with the help of measures from cognitive psychology (response times, RTs) and neuroscience (electroencephalography, EEG);

To address RQ 2 we considered the "second person neuroscience" framework (Schilbach et al., 2013). Schilbach and colleagues argue that humans' social cognitive mechanisms are activated differently when participating in a real-time social interaction, relative to only observing social stimuli. In other words, to understand the mechanism of social cognition, it is pivotal to observe them in action (Bolis and Schilbach, 2020; Redcay and Schilbach, 2019). Therefore, we designed three different experiments where participants engaged with a humanoid robot to various degrees. We followed the steps below:

1- We designed an experiment where participants were exposed to different sets of behaviours exerted by the robot. As measures, we administered the IST and a battery of questionnaires to assess our participants' personality traits and attitudes towards the robot. With this experiment, we aimed at exploring whether only observing a robot displaying different degrees of a thought process (hesitating over a decision vs. prompt, seemingly pre-programmed decisions) would modulate the adoption of the intentional stance. Moreover, we hypothesized that individual differences in personality traits and attitudes towards robots would correlate with the adoption of the intentional stance;

2- Building on the knowledge acquired from RQ 1 and the first step in RQ 2, we designed an experiment where participants would face a humanoid robot performing behaviours with different levels of human-likeness. After each session with the robot, participants completed the IST while we recorded their pupillary response. This experiment allowed us to investigate whether participants' individual differences in the likelihood of adopting the intentional stance towards the robot would lead to a different allocation of the cognitive resources when facing a humanoid robot and whether this different allocation of mental resources would modulate the sensitivity to hints of human-likeness;

3- Lastly, we designed an experiment in which participants truly participated in a shared social context with the robot (rather than only observing its behaviour). In this experiment, participants would watch some videos with the robot. From time to time, the robot would either emotionally react contingently to the events on the screen (humanlike condition) or would exert simple machinelike movements (machinelike condition). Before and after the videos' session, participants would complete the IST and a battery of questionnaires to assess their individual differences in personality, anthropomorphism, and attitudes towards robots. We then compared the two experimental conditions on their tendency to adopt the intentional stance towards a robot. Moreover, we report correlations between the IST and the individual differences battery of questionnaires.

Finally, the studies designed to address RQ3 were most affected by the Covid-19 pandemic. In collaboration with Dr. Wu Yan, from I2R, A*STAR, we aimed at addressing the question with an experimental design involving the embodied robot and participants from Singapore. Despite our strenuous efforts in conducting the study (see Covid-19 statement for further details), it was not possible to conclude the data collection. As a contingency measure, we decided to conduct an online study, in order to examine the relationship between cultural differences, individual tendency to anthropomorphize, and to adopt the intentional stance towards humanoid robots. The experiment consisted in a battery of tests and questionnaires, involving a large number of participants from all over the world. We modelled our data with a path model, hypothesizing the influence of cultural differences on anthropomorphism and the adoption of the intentional stance. In addition, as a second step, we further proceed to explore the causal effect of collectivism on the adoption of the intentional stance, mediated by the individual tendency to anthropomorphize a robot.

2.8 Candidate's contribution

As a Ph.D. candidate, I contributed to all major aspects of the present Ph.D. project and to the publications that are presented in it.

Below I will report my contribution to each paper collected in the present Ph.D. thesis:

Publication 1: I conceptualised the design of the experiment with Dr. Davide Ghiglino, Dr. Francesca Ciardo, Dr. Ebru Baykara and Prof. Agnieszka Wykowska. I developed the experimental materials with Dr. Jairo Perez-Osorio, and planned statistical analysis with Dr. Ebru Baykara, Dr. Francesca Ciardo and Dr. Davide Ghiglino. I wrote the manuscript with Dr. Davide Ghiglino, Dr. Francesca Ciardo, Dr. Jairo Perez-Osorio and Prof Agnieszka Wykowska.

Publication 2: I conceptualised the design of the experiment with Prof. Agnieszka Wykowska, I developed the experimental materials and collected the data. I analysed the data with Dr. Nicolas Spatola and discussed the data and wrote the manuscript with Dr. Nicolas Spatola, Dr. Jairo Perez-Ososrio and Prof. Agnieszka Wykowska.

Publication 3: I conceptualised and designed the experiment with Dr. Jairo Perez-Ososrio and Prof. Agnieszka Wykowska. I collected the data. I analysed, discussed the data and wrote the manuscript with Dr. Jairo Perez-Osorio and Prof. Agnieszka Wykowska. Dr. Davide De Tommaso programmed the robot behaviours.

Publication 4: I conceptualised the design of the pupillometry task with Prof. Agnieszka Wykowska, I collected the data with Dr. Davide Ghiglino. I analysed the data with Dr. Francesco Bossi. I wrote the manuscript with Prof. Agnieszka Wykowska. Dr. Davide Ghiglino and Prof. Agnieszka Wykowska designed the observational task. Dr. Davide De Tommaso programmed the robot behaviours.

Publication 5: I conceptualised the design of the experiment with Dr. Jairo Perez-Osorio and Prof. Agnieszka Wykowska, I created the behaviours of the robot with Dr. Davide De Tommaso. I collected and analysed the data. I discussed the data and wrote the manuscript with Prof. Agnieszka Wykowska. Dr. Davide De Tommaso implemented the behaviours on the robot.

Publication 6: I conceptualised the design of the experiment with Prof. Agnieszka Wykowska. I collected the data. I analysed the data with Dr. Nicolas Spatola. I and Prof Agnieszka Wykowska discussed the data and wrote the manuscript with Dr. Nicolas Spatola and Prof. Agnieszka Wykowska.

Part II

Attributing intentionality to robots: the development of the InStance Test

Chapter 3

Publication 1: Do we adopt the intentional stance towards humanoid robots?

This chapter presents the published manuscript from the following publication:

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., and Wykowska, A., (2019). Do we adopt the intentional stance toward humanoid robots?. Front.Psychol.,vol.10,no.MAR,2019, doi:10.3389/fpsyg.2019.00450. Materials and datasets are available at https://osf.io/pe2z3/

Authors Contributions SM, DG, FC, EB, and AW designed the study. SM and JP-O developed the experimental materials. EB, FC, and DG analyzed the data. SM, DG, FC, JP-O, and AW wrote the manuscript

3.1 Abstract

In daily social interactions, we need to be able to navigate efficiently through our social environment. According to Dennett (Dennett, 1971), explaining and predicting others' behaviour with reference to mental states (adopting the intentional stance) allows efficient social interaction. Today, we also routinely interact with artificial agents: from Apple's Siri to GPS navigation systems. In the near future, we might start casually interacting with robots. This paper addresses the question of whether adopting the intentional stance can also occur with respect to artificial agents. We propose a new tool to explore if people adopt the intentional stance toward an artificial agent (humanoid robot). The tool consists in a questionnaire that probes participants' stance by requiring them to choose the likelihood of an explanation (mentalistic vs. mechanistic) of a behaviour of a robot iCub depicted in a naturalistic scenario (a sequence of photographs). The results of the first study conducted with this questionnaire showed that although the explanations were somewhat biased toward the mechanistic stance, a substantial number of mentalistic explanations were also given. This suggests that it is possible to induce adoption of the intentional stance toward artificial agents, at least in some contexts.

3.2 Introduction

Over the last decades, new technologies have entered our houses inexorably, becoming an integral part of our everyday life. Our constant exposure to digital devices, some of which seemingly "smart," makes the interaction with technology increasingly more smooth and dynamic, from generation to generation (Baack et al., 1991; Gonzàles et al., 2012; Zickuhr and Smith, 2012). Some studies support the hypothesis that this exposure is only at its beginning: it seems likely that technologically sophisticated artefacts, such as humanoid robots, will soon be present in our private lives, as assistive technologies and housework helpers (for a review see Stone et al., 2016). Despite the fact that we are becoming increasingly habituated to technology, little is known about the social cognitive processes that we put in place during the interaction with machines and, specifically, with humanoid robots. Several authors have theorized that humans possess a natural tendency to anthropomorphize what they do not fully understand. Epley et al. (2007) for instance, defined anthropomorphism as the attribution of human-like characteristics and properties to non-human agents and/or objects, independently of whether they are imaginary or real. The likelihood of spontaneous attribution of anthropomorphic characteristics depends on three main conditions (Epley et al., 2007; Waytz, Morewedge et al., 2010): first, the availability of characteristics that activate existing knowledge that we have about humans; second, the need of social connection and efficient interaction in the environment; and, finally, individual traits (such as the need of control) or circumstances (e.g., loneliness, lack of bonds with other humans). In the Vol. I of the Dictionary of the History of Ideas, Agassi (Agassi, 1973) argues that anthropomorphism is a form of parochialism, allowing projecting our limited knowledge into a world that we do not fully understand. Some other authors claim that we, as humans, are the foremost experts in what it means to be human, but we have no phenomenological knowledge about what it means to be non-human (Gould, 1996; Nagel, 1974). For this reason, when we interact with entities for which we lack specific knowledge, we commonly choose "human" models to predict their behaviours.

A concept similar, but not identical, to anthropomorphism is the intentional stance. Intentional stance is a narrower concept than anthropomorphism, as the latter seems to be involving attribution of various human traits, while adopting the intentional stance refers more narrowly to adopting a strategy in predicting and explaining others' behaviour with reference to mental states. The concept of intentional stance has been introduced by Daniel Dennett, who proposed that humans use different strategies to explain and predict other entities' (objects, artefacts, or conspecifics) behaviours (Dennett, 1971; Dennett, 1989). Dennett defines three main strategies or "stances" that humans use. Consider chemists or physicists in their laboratory, studying a certain kind of molecules. They try to explain (or predict) the molecules' behaviour through the laws of physics. This is what Dennett calls the physical instance. There are cases in which laws of physics are an inadequate (or not the most efficient) way to predict the behaviour of a system. For example, when we drive a car, we can fairly predict that the speed will decrease if we push the brake pedal, since the car itself is designed this way. To make this kind of prediction, we do not need to know the precise physical mechanisms behind all atoms and molecules in the braking system of the car, but it is sufficient to rely on our experience and knowledge of how the car is designed. Dennett describes this as the design stance. Dennett proposes the existence of also a third strategy, the intentional stance. Intentional stance relies on the ascription of beliefs, desires, intentions and, more broadly, mental states to a system, in order to explain and predict its behaviour: "(. . .) there is yet another stance or strategy that one can adopt: the intentional stance. Here is how it works: first you decide to treat the object whose behaviour is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many – but not in all – instances yield a decision about what the agent ought to do; that is, what you predict the agent will do" (Dennett, 1989, p. 17). Please note that the concept of intentional stance, following Dennett's description, can be distinguished from the concept of Theory of Mind (ToM) in the sense in which it has been used in developmental psychology (e.g., Baron-Cohen, 1997). Although the two concepts are very tightly linked, and often subsumed under a common conceptual category (e.g., Baron-Cohen, 1997), they do differ with respect to the context in which they have been introduced in literature, and, what follows, in the empirical ways of addressing the concepts. As described above, the intentional stance has been introduced by Dennett in the context of two other stances (or strategies) that allow predicting and explaining behaviour of an observed system. Therefore, if an empirical test is set out to examine whether one adopts the intentional stance toward a system, the contrasting conditions should be either the design stance or the physical stance. On the contrary, the ToM has been introduced (Baron-Cohen, 1997; Leslie, 1994) to denote a capacity of understanding mental states of other humans that explain and predict their behaviour, but that might be different from one's own mental states (perspective taking) and might misrepresent reality (false beliefs). In this context, empirical tests that address the concept of ToM will not contrast the ToM condition with design or physical stance, but rather will make a contrast between different mental states (e.g., true vs. false beliefs). This implies that if one has a ToM of another human's behaviour, one has adopted the intentional stance, but not necessarily vice versa. In the "Sally and Anne" test (Baron-Cohen et al., 1985; Wimmer and Perner, 1983), I can adopt the intentional stance toward Sally (explain her behaviour with reference to mental states) but attribute incorrect mental states to her (that is, not understand that her perspective is different from mine, or from reality). So I can fail in ToM test, but I can still adopt the intentional stance. In this way, with reference to human agents, the intentional stance is a necessary condition for ToM, but not a sufficient one.

Even though we distinguish intentional stance from ToM, the idea of adopting the intentional stance toward others shares with the ToM the reference to mental states during social cognition processes. Therefore, it might fall under the same criticism as the ToM accounts of social cognition. In literature, there has been a heated debate regarding ToM accounts of social cognition (e.g., Gopnik and Wellman, 1992). Some authors (e.g., S. Gallagher, 2001; Zahavi and Gallagher, 2008) proposing the "direct perception" account deny the core assumption of ToM accounts, which is based on the implication: from the observed behaviour of others, we infer the unobservable mental states to explain and predict the behaviour. The authors propose that there is no distinction between observable behaviour and unobservable mental states, as the observed behaviour contains already cognitive/mental processes, and is perceived as such. Another line of criticism is embedded in the enactivist or interactionist accounts of social cognition (De Jaegher et al., 2010; Michael, 2011; Zahavi and Gallagher, 2008). The criticism of ToM based on those approaches is that ToM account is grounded too much in spectatorial, individualist, and cognitivist assumptions, which rely on the observer passively viewing behaviours of others and making inferences about mental states. The interactionists propose that social cognition is rather a participatory and interactive process allowing humans to understand behaviour of others without mindreading, through "making sense of the situation together" (Bohl and van den Bos, 2012, p. 3), based on the interaction processes themselves, at the supra-individual level (e.g.,Reddy and Morris, 2004, see also Michael, 2011; Bohl and van den Bos, 2012 for a review). Some authors (e.g., Bohl and van den Bos, 2012) postulate that neither of the accounts (neither ToM nor the interactionist/enactivist accounts) is sufficient to explain all domains of social cognition. This is because they address different type of processes: while interactionists accounts are better suited to explain Type I (fast, efficient, stimulus-driven and inflexible) processes of social cognition, ToM accounts for Type II (slow, cognitively laborious, flexible, and often conscious) processes (Bohl and van den Bos, 2012, p. 1). In either case, the aim of this paper is not to defend ToM accounts of social cognition. In our view, addressing the question of whether humans adopt the intentional stance toward artificial agents is of interest independent of the debate regarding various accounts of social cognition. This is because, even if ToM accounts cannot explain various aspects of social cognition, humans do sometimes use mentalistic vocabulary when describing behaviour of others. Except for radical interactionists, it is rather agreed upon that ToM accounts for at least a set of processes occurring during social interaction. Bohl and van den Bos (2012), for example, point out that interactionism cannot explain situations when interactions are not going smooth, when we try to explain others' behaviour with reference to mental states because of competition, disagreement, or conflict. Therefore, as long as one does not postulate that ToM is the foundation for all social cognitive processes, it remains justified to propose ToM as one aspect of social cognition. In this context, it is theoretically interesting to ask whether humans could potentially also use mentalistic vocabulary toward artificial agents, and if so, under what conditions. Perhaps this is not the only factor that would have an impact on social engagement or interaction with those agents, but certainly one that does play a role, along with other factors. Furthermore, it is an interesting question from the point of view of artificial intelligence, as adopting the intentional stance toward artificial agents is some form of the Turing test (Turing, 1950). Therefore, given the above considerations, we set out to explore whether humans can adopt the Intentional Stance toward artificial agents, by contrasting mentalistic interpretations of behaviour with mechanistic ones.

In this context, it needs to be noted, however, that adopting the intentional stance toward an artefact (such as a humanoid robot) does not necessarily require that the artefact itself possesses true intentionality. Adopting the intentional stance might be a useful or default way to explain a robot's behaviour. Perhaps we do not attribute mental states to robots, but we treat them as if they had mental states (Thellman et al., 2017). Breazeal and Scassellati (Breazeal and Scassellati, 1999) highlighted that this process does not require endowing machines with mental states in the human sense, but the user might be able to intuitively and reliably explain and predict the robot's behaviour in these terms. Intentional stance is a powerful tool to interpret other agents' behaviour. It leads to interpreting behavioural patterns in a general and flexible way. Specifically, flexibility in changing predictions about others' intentions is a pivotal characteristic of humans. Adopting the intentional stance is effortless for humans, but of course, it is not the perfect strategy: if we realize that this is not the best stance to make predictions, we can refer to the design stance or even the physical stance. The choice of which stance to adopt is totally free and might be context-dependent: it is a matter of which explanation works best. Let us consider our interactions with smartphones. The way we approach them is fluid: we adopt the design stance when we hear a "beep" that notifies us about a text message or an e-mail, but we might switch to the intentional stance when voice recognition such as Apple's Siri is not responding to us adequately, and gives us an incorrect answer to our question. In fact, it might even happen that we become frustrated and ask our smartphone a rhetorical question "why are you so stupid?". In this context, it is important to consider also cultural differences: the likelihood of adopting the intentional stance toward artefacts might differ from culture to culture (Dennett, 1989). Taken together, it is intriguing whether humans do sometimes adopt the intentional stance toward humanoid robots, at least in some contexts.

3.3 Aim of the study

In order to address the question of whether humans adopt the intentional stance toward a robot, we created a tool (the Intentional Stance Questionnaire, IST) that should probe the adoption of intentional stance toward a specific robot platform, the iCub robot (Metta et al., 2010; Natale et al., 2017). The aims of this study were the following:

3.3.1 Aim 1

Developing a tool that would allow for measuring whether humans would sometimes (even if only in some contexts) adopt the intentional stance toward a robot. The study reported in this paper aimed at providing a baseline "intentional stance" score (ISS). As such, it could subsequently serve as a score against which other experimental conditions may be compared (in future research). When other conditions in which participants' likelihood of adopting the intentional stance are manipulated experimentally (for example, through the robot's appearance, behaviour, a specific mode of interaction, etc.) a given ISS measured with IST, might then be compared to the baseline ISS reported here. This should allow for evaluating whether the experimental factors up- or down-modulate baseline ISS. Please note that the study reported here focused only on the baseline score.

3.3.2 Aim 2

Exploring if humans would sometimes adopt the intentional stance toward robots. We were interested in whether some contexts can evoke mentalistic explanations of behaviour of a humanoid robot. Please note that we focused on only one type of mental states (and, what follows, intentionality) attributed to artificial agents, namely the more explicit, propositional-attitudes mental states carrying the inherent "aboutness" (e.g., belief that. . . , desire that. . . , see Brentano and Heidegger, 1874; Churchland, 1981). Our study did not address more implicit forms of adopting the intentional stance. As argued above, our aim was not to defend the ToM accounts of social cognition. Furthermore, we also did not aim at determining how the appearance of a robot influences adoption of the intentional stance, or whether the degree of adoption of the intentional stance is smaller or larger as compared to other agents (humans or non-anthropomorphic robots). On the contrary, we intended a more modest aim: our goal was only to establish a baseline ISS toward a specific humanoid robot, iCub. We created 34 fictional scenarios, in which iCub appeared (in a series of photographs) engaged in different activities in a daily life context. Each scenario consisted of three aligned pictures, depicting a sequence of events. For each scenario, participants had to rate (by moving a slider on a scale) if iCub's behaviour is motivated by a mechanical cause (referring to the design stance, such as malfunctioning, calibration, etc.) or by a mentalistic reason (referring to the intentional stance, such as desire, curiosity, etc.).

3.4 Materials and Methods

3.4.1 Sample

One hundred and six Italian native speakers with different social and educational backgrounds (see Table 3.1 for demographical details) completed our InStance Test. Data collection was conducted in accordance with the ethical standards laid down in the Code of Ethics of the World Medical Association (Declaration of Helsinki), procedures were approved by the regional ethics committee (Comitato Etico Regione Liguria).

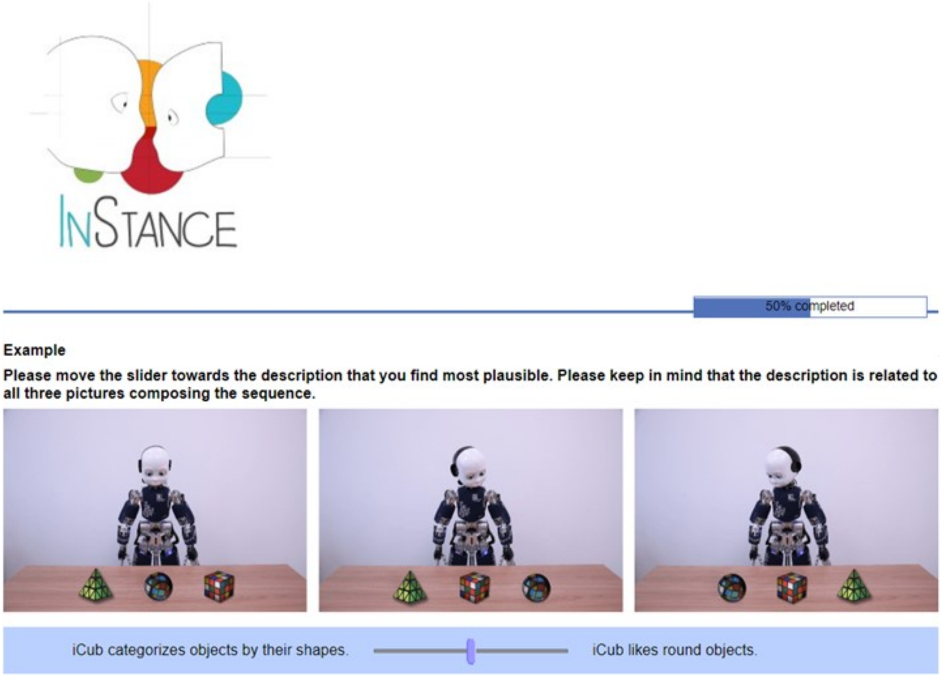
	Demographic characteristics
Age(year), mean (SD) [min - max]	$33.28(12.92)[18-72]$
Female, n (%)	$68(64.2\%)$
Education (years), mean (SD) [min - max]	$16.43(3.04)[8,24]$

Table 3.1. Demographic details of the sample (N= 106).

3.4.2 Questionnaires

3.4.3 The InStance Test

Each item of IST was composed of a scenario and two sentences with a bipolar scale and a slider between the two sentences (one of the sentences was positioned on the left, and the other one on the right extreme of the scale), see Figure 3.1 for an example. For a complete list of images and sentences included in the IST see Appendix A (English version), or the following link (for English and Italian version): https://instanceproject.eu/publications/rep.



Example

Please move the slider towards the description that you find most plausible. Please keep in mind that the description is related to all three pictures composing the sequence.

iCub categorizes objects by their shapes.

iCub likes round objects.

Figure 3.1. Screenshot from the InStance Test in English.

We created 34 scenarios depicting the iCub robot interacting with objects and/or humans. Each scenario was composed of three pictures (size 800 × 173.2 pixels). Out of the 34 scenarios, 13 involved one (or more) human interacting with the robot; 1 scenario showed a human arm pointing to an object; 20 scenarios depicted only the iCub robot. Ten scenarios included pictures digitally edited (Adobe Photoshop CC 2018). The types of action performed by iCub depicted in the scenarios were: grasping, pointing, gazing, and head movements. Each item included two sentences, in addition to the scenario. One of the sentences was always explaining iCub's behaviour referring to the design stance (i.e., mechanistic explanation), whereas the other was always describing iCub's behaviour referring to mental states (i.e., mentalistic explanation). Mentalistic and mechanistic sentences were equally likely to appear either on the left or on the right side of the scale, as the mapping between type of sentence and position was counterbalanced across items. Moreover, we kept iCub's emotional expression constant across the scenarios to avoid bias toward mentalistic explanations. In order to be certain that the sentences were describing the action in a mechanistic or mentalistic way, prior to data acquisition, we distributed the sentences (only sentences alone, with no associated pictures) to 14 volunteers who had a degree in philosophy (all Italian native speakers) to rate on a 10-

point Likert scale how much they understood each sentence as mechanistic or mentalistic (0 = totally mechanistic, 100 = totally mentalistic). As no scenario was presented with the sentences, the raters were not informed that the sentences were created to describe the behaviour of a humanoid robot. In addition, the subject of each sentence was named as a general "Agent A" to avoid any bias arising from the name of the robot. The mean score given was 8.2 for the mentalistic sentences and 4.3 for the mechanistic sentences. Based on the responses in this survey, we modified the sentences that were not matching our intent. In particular, we modified sentences that obtained an average score between 4 and 5 (meaning that they were not clearly evaluated as mechanistic or mentalistic), using as cut-off 4.3 since most of the critical sentences were from the mechanistic group (14 out of 15 sentences). We modified 15 out of the 35 initial pairs of sentences to match our intended description (mentalistic or mechanistic).

3.4.4 Other Questionnaires

In addition to the InStance Test, we administered the Italian version of the Reading the Mind in the Eyes (Baron-Cohen et al., 1999; Baron-Cohen et al., 2001; Serafin and Surian, 2004) and the ToM subscale of the questionnaire developed by Völlm et al. (Völlm et al., 2006). These tests were used as a control to check for outliers in ToM abilities.

3.4.5 Reading the Mind in the Eyes

The Reading the Mind in the Eyes test was developed by Baron-Cohen et al. (1999, 2001) to test the abilities of infering other's mental states through only looking at others' eyes. In this questionnaire, participants are asked to view 36 photos of people's eyes. Below the photographs of the eyes, four adjectives are presented. Participants are asked to choose one adjective that describes best the photograph they are looking at. We selected this test, as it is one of the most used tests to measure ToM abilities.

3.4.6 A test of Völlm et al. (2006)

Völlm et al. (2006) developed a questionnaire to evaluate the neural correlates of ToM and empathy in an fMRI study. In this test, the presented stimuli are non-verbal cartoon stripes of three frames: participants are instructed to look at the stripes and choose the frame that, according to them, would best conclude the depicted story. For the purpose of this study, we presented only the 10 items included in the ToM subscale of the test. We selected this test as a non-verbal test of mentalising abilities, complementary to the Reading the Mind in the Eyes test.

3.4.7 Data acquisition for the InStance Test

All the questionnaires were administered via SoSci survey. Participants received the URL addresses of all the questionnaires. Participants were asked to fill out the questionnaires in the order they were provided: a generic information questionnaire, IST, the Mind in the Eyes and finally the Völlm et al. (2006) questionnaire. The generic information questionnaire collected demographic information of participants (see Table 3.1) and whether they were familiar with robots or not. The IST was composed of 34 items and 1 example item. Only one item at a time was presented (Figure 3.1). In each item, participants were explicitly instructed to move a slider on a bipolar scale toward the sentence that, in their opinion, was a more plausible description of the story depicted in the scenario. As illustrated in Figure 3.1, the two statements (mentalistic and mechanistic) were placed at the two bonds of the scale. The cursor was initially always placed at the center of the scale (i.e., the null value). For 50% of the items, the mechanistic sentence was presented on the left side of the slider, while the mentalistic was presented on the right side. For the other 50% the location of mechanistic and mentalistic sentences was reversed. The order of presentation of the items was randomized.

3.5 Data Analysis and Results

All statistical analyses were performed in R (version 3.4.0, available at http://www.rproject.org).

Data analysis was conducted on a sample including responses collected from 106 participants. For each participant, we calculated the InStance Score (ISS). To this end, we converted the bipolar scale into a 0–100 scale where 0 corresponded to completely mechanistic and 100 to a completely mentalistic explanation. The null value of the scale, i.e., the starting position of the slider that was equally distant from both the two limits, corresponded to the 50. The ISS score was computed as the average score of all questions. Scores under 50 meant the answer was mechanistic', scores above 50 meant they were 'mentalistic.'

The overall average score for the IST was 40.73 (SD= 10.1) (with 0 value indicating the most mechanistic score and 100 indicating the most mentalistic score). We tested the distribution of ISS for normality with the Shapiro–Wilk test. Results showed that Average ISS were distributed normally, $W = 0.99, p > 0.05$. We analyzed scores of each item of the IST for outliers, values that lie outside 1.5 $*$ Inter Quartile Range (IQR, the difference between 75th and 25th quartiles), and did not find any outlier item. In order to compare if the average ISS significantly differed from a completely mechanistic bias, we run one-sample t-tests against a critical value of 0 (i.e., the value corresponding to a mechanistic bond). Results showed that the average ISS significantly differed from 0, t(105) = 28.80, p < 0.0001.

The average ISS for each item are summarized in Table 3.2; the ISS distribution of the individual averages is reported in Figure 3.2.

Two scenarios that scored highest in mentalistic descriptions (78.16 and 68.83 on average, respectively) are presented in Figure 3.3; two scenarios with the most mechanistic scores (10.07 and 11.97 on average, respectively) are presented in Figure 3.4. In order to test internal consistency of the responses to the IST, we calculated Cronbach's alphas, which yielded a

Item	N° humans in the scenario	Mean	SD	Item	N° humans in the scenario	Mean	SD
1	0	32.25	39.72	18	1	38.09	38.33
2	0	46.37	41.91	19	0	24.71	31.91
3	1	50.28	41.71	20	1	56.58	39.13
4	1	31.51	35.74	21	2	24.42	33.80
5	0	31.45	38.51	22	0	34.32	37.02
6	0	43.46	40.89	23	1	52.42	42.11
7	2	33.27	38.11	24	1	68.83	38.41
8	0	22.84	30.21	25	1	78.16	30.59
9	0	55.72	41.15	26	1	57.61	38.94
10	1	32.35	37.37	27	0	45.08	40.81
11	1	66.05	39.47	28	0	10.07	20.62
12	0	28.56	35.14	29	0	48.55	41.41
13	0	42.65	40.26	30	0	25.10	34.34
14	0	41.85	40.47	31	0	11.97	25.22
15	0	33.79	38.01	32	0	34.21	38.33
16	0	33.28	36.97	33	0	46.40	41.53
17	1	65.04	38.54	34	0	34.40	36.91

Table 3.2. Average score and standard deviation for each item of the IST $(N = 106)$.

result of 0.83 for 34 items, indicating high internal consistency of the IST items. To evaluate the contribution of each item to the internal consistency of the IST, we run an item analysis (Ferketich, 1991). Thus, we re-estimated the α coefficient when an item was deleted. If the value of the alpha coefficient increases after the exclusion of the item, this means that the item is inconsistent with the rest of the test (Gliem and Gliem, 2003). As reported in Table 3.3, results of the Item Analysis clearly indicate that none of the items is inconsistent with the rest of the questionnaire to explore possible latent traits underlying the variance of the ISS, we conducted a principal-components analysis (PCA, varimax method). We assumed that the number of humans present in the scenario (one, two, or zero) might have introduced a latent factor explaining the variability of the ISS.

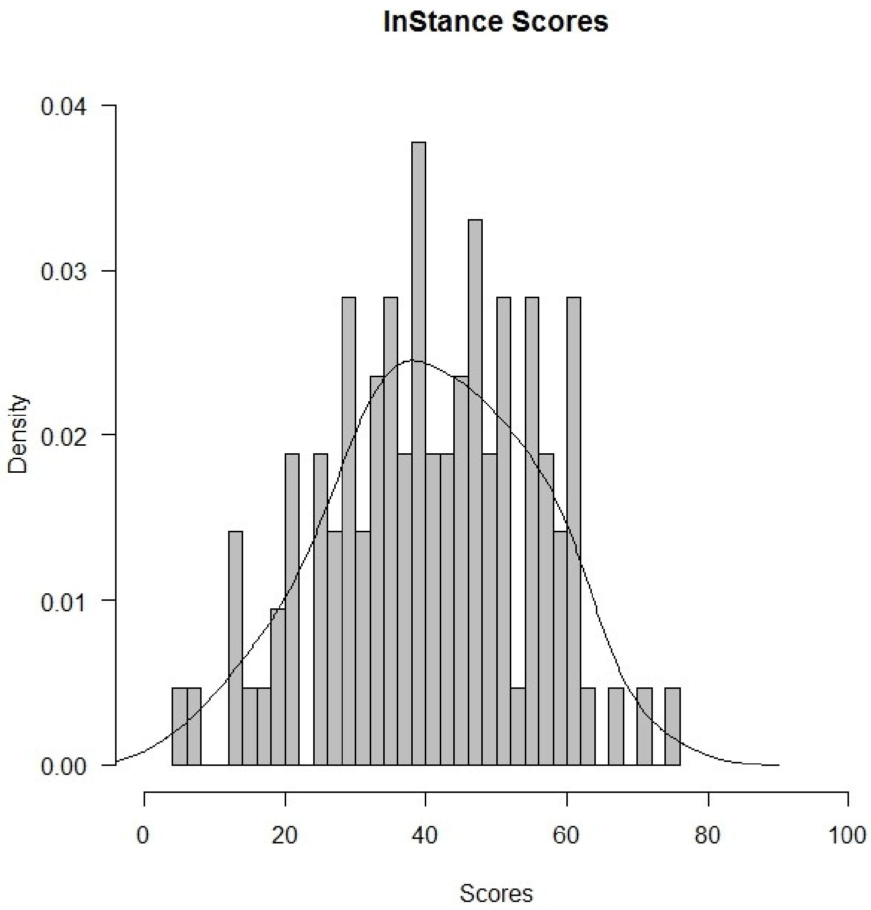


Figure 3.2. ISS distribution (individual averages, N = 106).

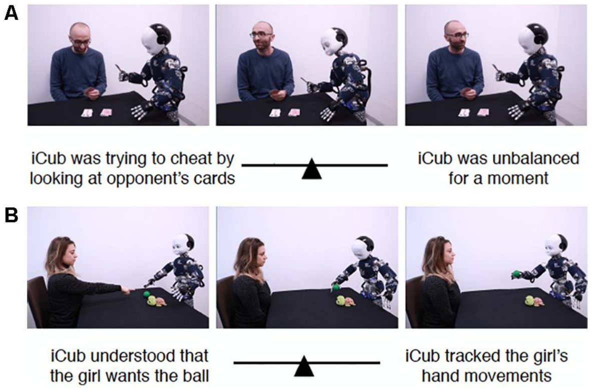

A
iCub was trying to cheat by looking at opponent's cards
iCub was unbalanced for a moment
B
iCub understood that the girl wants the ball
iCub tracked the girl's hand movements

Figure 3.3. Two scenarios with the highest mentalistic scores. (A) Shows Item 25 which received the score 78.16 on average, while (B) depicts Item 24 which received the score of 68.83 on average.

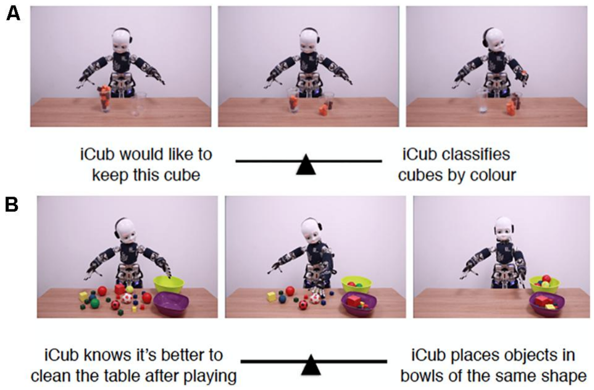

A
iCub would like to keep this cube
iCub classifies cubes by colour
B
iCub knows it's better to clean the table after playing
iCub places objects in bowls of the same shape

Figure 3.4. Two scenarios with the highest mechanistic scores. (A) Shows Item 28 which received the score 10.07 on average, while (B) depicts Item 31 which received the score of 11.97 on average.

To this end, the number of components to be extracted was limited to three. If the number of humans depicted in the scenarios represented a latent trait of the variance, then items depicting the same number of humans were expected to be significantly correlated ($r < -0.30$ or $r > 0.30$) with the same component. Results showed that the three components together accounted for only 30.34% of the variance. The variance accounted for by each component was only 13.74, 9.88, and 6.71% for component 1, 2, and 3 respectively. Moreover, as reported in Table 3.4 items involving the same number of humans (e.g., Items 3, 4, 10, 11, 18, 20) did not significantly score $r < -0.30$ or $r > 0.30$ on the same component.

Thus, the three components are not related to the number of humans depicted in each item, and the presence or absence of humans depicted in the items is not a latent factor underlying the variance of our data. We evaluated the associations between the ISS scores and gender, age, education, number of children, and siblings of the respondents using multiple linear regression. Results showed no significant associations between the ISS and the respondents' characteristics of age, gender, education, number of children, or siblings (all ps > 0.32). To assess the relationship between the ISS and scores in Reading the Mind in the Eyes and Völlm questionnaires, Pearson product-moment correlation coefficients were calculated. Results showed no correlation between the ISS and the scores in the Reading the Mind in the Eyes or Völlm questionnaires (all ps > 0.06). From 106 respondents, 84% reported that they were completely unfamiliar with robots (N = 89) while the remaining 16% reported various levels of familiarity (N = 17). In order to check whether the ISS were associated with the degree of familiarity of the respondents with humanoid robots, we performed linear regression analysis. No effect of familiarity with robots on the ISS was observed, $p = 0.21$. The average scores for the IST were 36.69 and 41.50 for familiar and not familiar with robots respondents, respectively. The average ISS significantly differed from the critical value of 0 (i.e., the value associated with a completely mechanistic bias), $t(16) = 9.79$, $p < 0.0001$ and $t(88) = 29.30$, $p < 0.0001$, for the familiar and not familiar with robots respondents, respectively.

Item	Scale Mean if Item Deleted	Scale Variance when item Deleted	Cronbach's Alpha when Item Deleted
1	1349.41	229722.72	0.82
2	1338.29	224066.30	0.82
3	1334.38	224588.31	0.82
4	1353.15	238581.33	0.83
5	1353.21	234277.14	0.83
6	1341.20	225788.88	0.82
7	1351.39	229548.98	0.82
8	1361.82	237512.03	0.83
9	1328.94	226282.11	0.82
10	1352.31	229355.11	0.82
11	1318.61	227443.13	0.82
12	1356.10	234238.61	0.83
13	1342.01	224117.95	0.82
14	1342.81	236687.81	0.83
15	1350.87	232598.90	0.83
16	1351.38	232130.07	0.83
17	1319.62	228192.90	0.82
18	1346.57	234567.60	0.83
19	1359.95	233798.25	0.83
20	1328.08	235541.72	0.83
21	1360.25	230697.02	0.82
22	1350.31	234925.42	0.83
23	1332.24	226723.12	0.82
24	1315.83	229252.24	0.82
25	1306.50	236004.61	0.83
26	1327.05	232772.24	0.83
27	1339.58	22.3211.94	0.82
28	1374.59	241566.89	0.83
29	1336.11	236128.94	0.83
30	1359.56	241206.73	0.83
31	1372.69	236999.57	0.83
32	1350.45	231757.64	0.83
33	1338.26	229756.04	0.82
34	1350.26	236259.80	0.83

Table 3.3. Results of the item analysis.

	Component					Component		
Item	1	2	3		Item	1	2	3
1	−0.393	0.448	0.451		18		0.402	
2	0.348	−0.686			19	0.369		
3			−0.534		20	0.437		
4	0.318	−0.318			21	0.278	0.377	
5	0.512				22		0.261	
6	0.582				23	0.363		
7	0.268	0.381	−0.400		24		0.359	0.450
8	0.360				25	0.431	0.278	
9	0.266				26	0.345		
10	0.440	0.251			27	0.524	−0.380	
11	0.338	0.319	−0.288		28	0.524	−0.262	
12	0.413				29	0.482		
13	0.311				30	0.339	0.259	
14	0.605				31		0.677	0.310
15			0.506	−0.463	32	0.265		0.549
16	0.628				33		0.252	0.432
17		0.434	−0.255		34		0.385	

Table 3.4. Correlations between instance Test scores and other questionnaires.Extraction Method: Principal Component Analysis, Varimax Rotation. Scores between −0.25 and 0.25 are not reported.

3.6 Analyses on the Group Not Familiar With Robots

In addition to the analyses of data from the entire sample, we conducted further analyses on data from participants who reported no familiarity with robots. As outlined above, our main aim was to investigate the likelihood of adopting the intentional stance for a given humanoid iCub in the general population – that is, people with no previous experience with robots. For the group of respondents who were not familiar with robots, the ISS were also distributed normally (Figure 3.5), Shapiro–Wilk test: W = 0.99. p > 0.05.Results also showed no significant associations between the ISS and the respondent characteristics of age, gender, education, number of children, and siblings, for the group of respondents who were not familiar with robots (all ps > 0.16). Similarly, no correlations between the ISS and the scores in the Reading the Mind in the Eyes and Völlm et al. (2006) questionnaires were found (all ps > 0.16). In order to explore if responses to the questionnaire were polarized, we conducted a polynomial curve-fitting analysis on the density of the raw scores. Results showed a significant quadratic trend (t = 5.59; p < 0.001, R2 = 0.83), supporting the polarization hypothesis (Figure 3.6).

In addition, we estimated the percentage of participants who attributed 'mechanistic' or 'mentalistic' descriptions according to their average ISS. Participants were classified into two groups according to their ISS. Participants who scored below 50 (0 – 50 in our scale) were assigned to the Mechanistic group (N = 62), whereas participants with an ISS above 50 (50

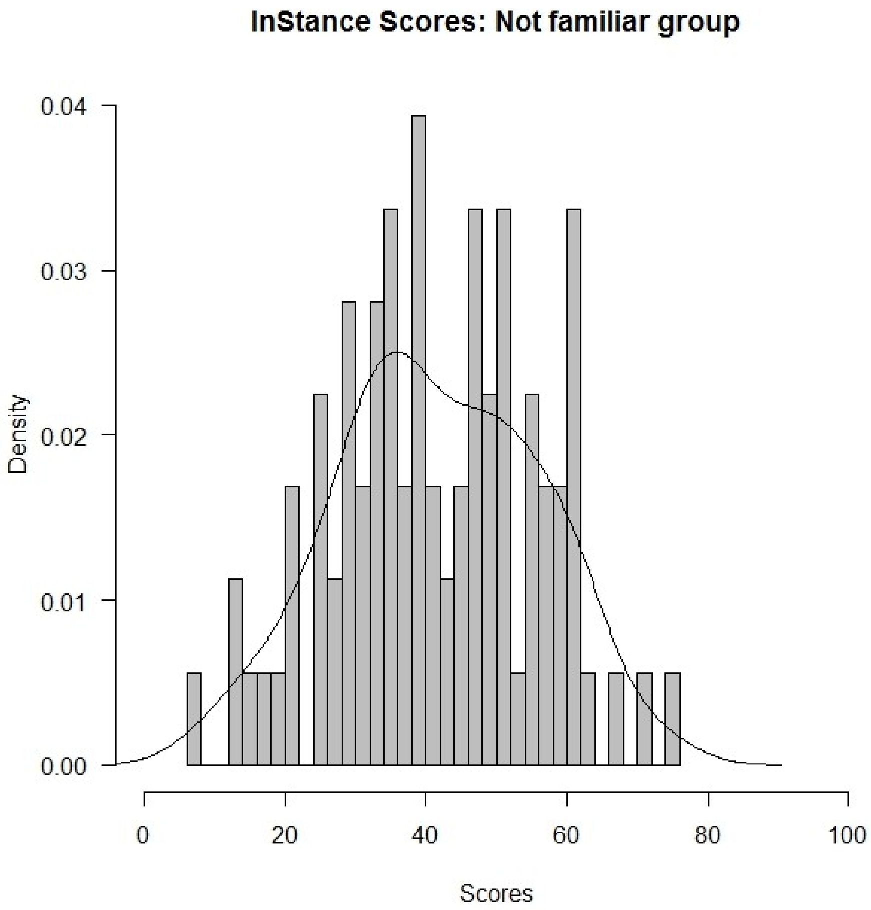


Figure 3.5. ISS (individual averages) distribution for the not familiar group (N = 89).

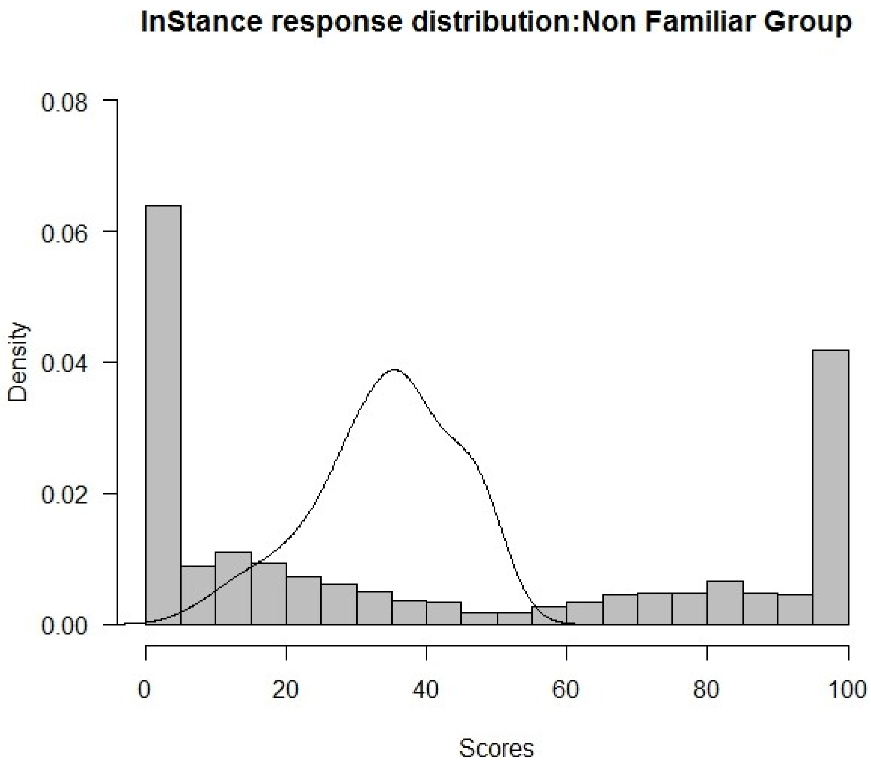


Figure 3.6. Plot of raw data for the not familiar group $(N = 89, Shapiro–Wilktest : W = 0.81p < 0.001)$.

– 100) were classified as the Mentalistic group (N = 27). To check whether the percentage of respondents in the Mechanistic and Mentalistic group differed from chance level (i.e., expected frequency of 0.5), we performed a chi-square test. Results revealed that the frequency of participants who scored Mechanistic (69.7%) and the frequency of participants who scored Mentalistic (30.3%) were both different from the chance level, $\chi^2(1.\ N = 89) = 13.76.\ p < 0.001$, Figure 3.7. In order to compare if the mean ISS of the two groups (Mechanistic and Mentalistic) significantly differed from the null value of our scale (i.e., 50, which corresponded to the position at which the slider was equally distant from both statements), we run one-sample t-tests against a critical value of 50 (i.e., the null value of our scale). Results showed that the mean ISS significantly differed from the null value of 50 both for the Mechanistic [M = 34.19, SEM = 1.28, t(61) = −12.33, p < 0.0001] and the Mentalistic group [M = 58.27, SEM = 1.18, t(26) = 7.03, p < 0.0001].

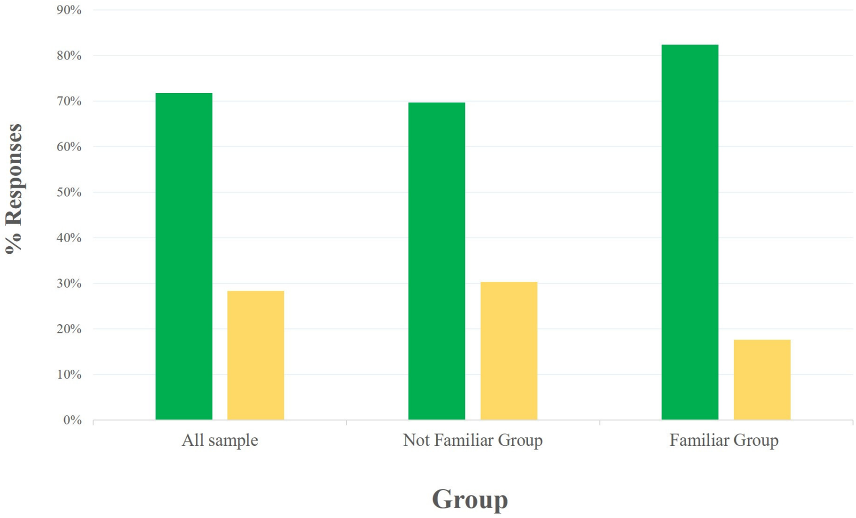


Figure 3.7. Percentage of Mechanistic (green bars) and Mentalistic (yellow bars) respondents for the entire sample (N = 106) on the left, the Not Familiar group (N = 89), and the Familiar group (N = 17) on the right.

3.7 General discussion

The aim of the present study was to develop a method for assessing whether humans would sometimes (at least in some contexts) adopt the intentional stance toward the humanoid robot iCub. To meet this aim, we developed a questionnaire (Intentional Stance Questionnaire, IST) that consisted of 34 items. Each item was a sequence of three photographs depicting iCub involved in a naturalistic action. Below each sequence, there was a slider scale on which participants could move the cursor in the direction of one of the extremes of the scale. On one of the extremes, there was a mentalistic description of what was depicted in the photographs, on the other extreme there was a mechanistic description. We administered the IST online and received responses from 106 participants. The high internal consistency (α = 0.83) of mean scores (Intentional Stance Scores, ISS) of the items revealed that the questionnaire we developed was uniform. This is consistent also with the result of the principal component analysis (PCA). Specifically, no differences were found between “social” (iCub interacting with one or two other agents) and “individual” (iCub alone) scenarios. Thus, we can conclude that our tool is reliable.

No correlation was found between sociodemographic characteristics of participants and ISS. Adopting the intentional or design stance was not affected by socioeconomic factors (i.e., education) for our questionnaire. Furthermore, we found no correlation with the two tests used for the assessment of ToM abilities, namely the Reading Mind in the Eyes (Golan et al., 2006) and the Völlm et al. (2006) comic-strip questionnaire. The main reason behind this might be that error rates in the healthy adult population in these two tests are typically very low, and, in our sample, participants were at ceiling performance and there was almost no variance in terms of accuracy. Arguably, testing a psychiatric population could lead to a higher variance of accuracy between participants, and could provide additional information regarding the relationship between ToM abilities and adoption of intentional stance toward non-human agents. Importantly for our purposes, however, using these tests showed that we did not have any outliers in terms of ToM abilities.

Overall, our results indicate that participants showed a slight bias toward a mechanistic explanation when they were asked to evaluate robot actions (mean overall score = 40.73). This might be due to the fact that participants were observing a robot agent. This is also in line with previous studies emphasizing that a lower degree of intentionality is attributed to machines when explaining their actions, while a higher degree is attributed when perceiving conspecifics' behaviours (Chaminade et al., 2010; Krach et al., 2008; for review, see Wiese et al., 2017). The results are also in line with the essence of the concept of the intentional stance (Dennett, 1989). We can speculate that, in order to understand a behaviour of a robot, humans are more prone to adopt the design, rather than the intentional, stance – and this is in spite of the natural tendency to anthropomorphize unknown entities (Epley et al., 2007). Perhaps nowadays, humanoid robots are not unknown entities anymore: in the Western world, some of them are already used at airports, shops or cultural events to provide information to visitors (Aaltonen et al., 2017; for a review see Mubin et al., 2018).

However, and interestingly, the design stance descriptions were not always chosen in order to explain iCub's actions, as the average score was significantly different from zero. This clearly indicates that participants have at times also chosen mentalistic explanations of the given scenarios. This choice could have depended on specific scenarios (some were more likely to be interpreted mentalistically than mechanistically), or also on individual differences among participants (there was a sub-sample that was more likely to adopt the intentional stance, and a sub-sample that was more likely to adopt the design stance). This shows that in principle, it might be possible to induce the adoption of intentional stance toward artificial agents. The likelihood of adopting the intentional stance might depend on the context in which the robot is observed, its behavioural characteristics (e.g., contingency of its behaviour on participant's behaviour, cf. Willemse et al., 2018), cultural background (attitude toward humanoid robots is strongly associated with culture (for a review see Haring et al., 2014) and also individual differences of participants.

In order to further investigate potential factors contributing to the choice between the mentalistic vs. mechanistic rating, we asked participants whether they had previous experience with robots. Our data did not show any effect of familiarity with robots ($p > 0.05$) on the ratings, but it is important to point out that the majority of our sample consisted of people not familiar with this kind of technology ($N = 89$). To analyze a more homogenous sample with respect to familiarity with robots, we conducted follow-up analyses only on the sample not familiar with robots.

By analyzing the differences in scores between items in the "non-familiar" group, we noticed that some scenarios strongly elicited a mentalistic explanation. Interestingly, when we examined mentalistic ($M = 58.27$) and mechanistic ($M = 34.19$) evaluations between participants, we found that both types of scores were significantly different from the null value of our scale (for both, one-sample t-test revealed significant difference from 50, $p < 0.001$). Together with the results implying polarization in the scores (Figure 3.6), this suggests that participants were clearly choosing for each scenario either a mechanistic or a mentalistic explanation, neither answering randomly nor relying on the middle point of the scale. Davidson (1999) has proposed one possible explanation for this phenomenon. The author pointed out that humans possess many types of vocabulary for describing the nature of mindless entities, and for describing intentional agents, but might lack a way of describing what is between the two (Davidson, 1999). This is also in line with Dennett's proposal (Dennett, 1981) of intentional stance: when we find a model that is most efficient to explain a behaviour, we take its prototypical explanation, and do not necessarily search for explanations that are in the fuzzy zones of in-between models. In each scenario of our questionnaire, the robot was always the same, but the action and the environment/context around it changed across conditions, modulating participants' ratings. However, the rating became quite polarized once there was a bias toward either the mentalistic or the mechanistic explanation. This might be a consequence of a general tendency of people to form discrete categories rather than continuous fuzzy concepts (Dietrich and Markman, 2003), but it does also suggest the existence of a certain degree of flexibility that allows humans to shift between such categories. Based on our results, we can argue that the adoption of such mentalistic or mechanistic models does not rely only on intrinsic properties of the agent, but also on contingent factors (Waytz, Morewedge et al., 2010) and on the observer's dispositions (Dennett, 1990), in a similar way as it occurs for human-likeness and anthropomorphism (Fink, 2012).

3.8 Limitations

Despite the novelty of our questionnaire, some level of caution needs to be assumed when interpreting the results. The main aim of our study was to develop a tool for exploring whether humans would sometimes adopt the intentional stance toward a specific robot iCub. Therefore, for each scenario, we created two alternative explanations equally plausible (and equally ambiguous so that one would not be chosen over the other due to its higher level of subjectively perceived accuracy in description). We asked 14 volunteers (with philosophy background) to rate whether the sentences we created were falling in the mechanistic or mentalistic category, as we intended them to be. The final version of our questionnaire demonstrated high internal consistency, but a deeper analysis of the effective ambiguity between the proposed sentences is needed. When dealing with questionnaires, items are often open to multiple interpretations, which might change from one individual to another. Future studies should address how individual differences affect the interpretation of our items. Similarly, some scenarios we created might appear difficult to interpret. It could therefore be argued that the coherence of the story lines might affect participants' ratings. However, the item analysis demonstrated that there is no significant difference between items in terms of reliability, thereby reassuring that items were uniformly coherent. Nevertheless, a future investigation on internal coherence of the scenarios might be needed in order to apply the IST to social robotics research.

One further limitation could be the lack of possibility to design a proper control condition. A control condition with a human agent is almost impossible, as mechanistic descriptions of a human behaviour are very unnatural. It is very strange to provide participants with mechanistic descriptions of human behaviour, for example, descriptions such as "calibration of motors," and it might even contradict the key concept of intentional stance. On the other hand, modifying descriptions to make them more plausible for human agents would not provide a proper control. In fact, our attempt of designing a control condition with a human agent (see Supplementary Materials) showed that participants found the task very strange, and they experienced the agent as robotic and unnatural (open comments after completion of the questionnaire). What has presumably happened in this case is that the depicted human agent became dehumanized due to the restrictions on the expressiveness and posture (to match the robot condition) and due to the mechanistic descriptions.

On the other hand, a non-anthropomorphic control condition would not be viable, as many mentalistic descriptions of IST actually refer to the behavioural repertoire that is human-like (gaze direction, being "surprised to see"). A non-anthropomorphic comparison should not have any human-like features (e.g., eyes). Without eyes, however, such descriptions are senseless.

However, comparing the intentional attributions toward a humanoid robot to a human agent or to a non-anthropomorphic robot was not the aim of the present study. We did not intend to ask the question of whether the degree of intentional stance adopted to a humanoid robot would be comparable with respect to another type of agent (human or non-anthropomorphic robot). The aim of our study was to examine if people sometimes adopt the intentional stance toward humanoid robots, and the results of our questionnaire showed that this is indeed the case.

A first step toward a comparison between a robot and a human agent in terms of adoption of intentional stance was made by Thellman et al. (2017). The authors designed a study in which they presented a series of images and verbal descriptions of different behaviours exhibited either by a person or by a humanoid robot. Verbal statements described an outcome, event, action or state either in positive or in negative terms (i.e., "Ellis makes a fantastic cake" or "Ellis burns the cake"). Participants were asked to rate the intentionality, controllability and desirability of the behaviour, and to judge the plausibility of seven different types of explanations (derived from a psychological model of lay causal explanation of human behaviour, Malle and Knobe, 1997). Results showed that the scores were very similar for humans and robots, meaning, people explained the behaviour of both agents in terms of mental causes. Interestingly, participants were also asked to rate how confident they were with their score. The confidence ratings revealed lower confidence when rating robot behaviour relative to rating human behaviour. This shows that despite explaining the behaviour of the robot in intentional terms, such interpretation perhaps caused some degree of cognitive dissonance. Further studies might need to explore more in depth on the origins of such effects.

Our study is, however, somewhat different from the study of Thellman et al. (2017), as it provided participants with the direct choice of mentalistic and mechanistic explanations of robot behaviour, thereby probing the adoption of intentional stance, and contrasting it with design stance more directly. Moreover, as argued above, we did not aim to compare the degree of adoption of the intentional stance toward a humanoid robot, with respect to another human. Finally, we also did not aim at answering the question of whether anthropomorphic physical appearance of the robot plays a role in adoption of intentional stance (in this case, a comparison with a non-anthropomorphic robot would be needed). Our aim was solely to present a tool that we developed, and explore whether humans would at times choose a mentalistic description/interpretation of behaviour of a humanoid robot.

3.9 Future Directions

In general, it is of interest – not only for theoretical but also practical reasons – to ask whether adoption of intentional stance is a factor in social acceptance. Robots are considered as future assistive technologies, with potential applications from healthcare to industry. However, some studies brought to light potential issues with acceptance of artefacts in the human social environments (Bartneck and Reichenbach, 2005). Furthermore, pop-culture has induced a substantial amount of skepticism toward robots, associating them with threats for humanity (F. Kaplan, 2004). For these reasons, it is crucial to understand what factors contribute to acceptance of robots in human environments, and whether adoption of the intentional stance is one of those factors. It has been argued that robots, and specifically humanoids, have the potential to trigger attribution of mental states, as long as they display observable signs of intentionality, such as human-like behaviours (Wykowska, Kajopoulos, Ramirez-Amaro et al., 2015; Wykowska, Kajopoulos, Obando-Leiton et al., 2015; Wykowska et al., 2014) and appearance (i.e., biological motion or human-like facial traits, Chaminade et al., 2007; Chaminade et al., 2012; Özdem et al., 2017; Wiese et al., 2017; Wiese et al., 2012).

3.10 Conclusion

In summary, the present study used a novel method to explore whether the intentional stance is at times adopted toward a humanoid robot iCub. Our results show that it is possible to induce adoption of the intentional stance toward the robot at times, perhaps due to its human-like appearance. Further research needs to explore what are the exact factors (individual difference, cultural context, specific characteristics of robot appearance or behaviour) that influence the adoption of intentional stance.

Chapter 4

Publication 2: Human vs Humanoid. A Behavioural Investigation of the Individual Tendency to Adopt the Intentional Stance

This chapter presents the published manuscript from the following publication:

Marchesi, S., Spatola, N., Perez-Osorio, N. and Wykowska, A., (2021). "Human vs humanoid. A behavioural investigation of the individual tendency to adopt the intentional stance," ACM/IEEE Int. Conf. Human-Robot Interact., pp. 332–340, 2021, doi: 10.1145/3434073.3444663. Materials and datasets are available at https://osf.io/65bn2/

Author Contributions: SM and AW designed the study. SM developed the experimental materials and collected the data. SM and NS analyzed the data. SM, NS, JP-O, and AW discussed the data and wrote the manuscript.

4.1 Abstract

Humans interpret and predict behaviour of others with reference to mental states or, in other words, by adopting the intentional stance. The present study investigated to what extent individuals adopt the intentional stance towards two agents (a humanoid robot and a human). We asked participants to judge whether two different descriptions fit the behaviours of the robot/human displayed in photographic scenarios. We measured acceptance/rejection rate of the descriptions (as an explicit measure) and response times in making the judgment (as an implicit measure). Our results show that at the explicit level, participants are more likely to use mentalistic descriptions for the human agent and mechanistic descriptions for the robot. Interestingly, at the implicit level, we found no difference in response times associated with the robotic agent. We argue that, at the implicit level, both stances are processed as "equally likely" to explain the behaviour of a humanoid robot, while at the explicit level there is an asymmetry in the adopted stance. Furthermore, cluster analysis on participants' individual differences in anthropomorphism likelihood revealed that people with a high tendency

to anthropomorphize tend to accept faster the mentalistic description. This suggests that the decisional process leading to adoption of one or the other stance to adopt is influenced by individual tendency to anthropomorphize non-human agents.

4.2 Introduction

Humans predict others' actions based on the assumption that the observed behaviour can be explained with reference to internal mental states (i.e., desires, beliefs and goals). Given the covert nature of these internal states, we have learned to detect and interpret social signals in order to infer the mental states and predict upcoming actions (Baron-Cohen et al., 1999). Daniel Dennett [1989] defined this strategy of predicting other agents' behaviour with reference to their mental states as the intentional stance. We adopt the intentional stance towards other humans quite effortlessly and ubiquitously. Research already demonstrated that humans can sometimes adopt the intentional stance also towards artificial agents (Perez-Osorio and Wykowska, 2020 for review), a phenomenon related to the process of anthropomorphism, i.e., attributing human properties to non- human agents. To what extent humans readily adopt the intentional stance toward robots and how to measure this phenomenon remain unclear, although some research has addressed this problem and tried to operationalize such a philosophical concept (Marchesi et al., 2019; Thellman et al., 2017). The present paper addresses these issues by investigating whether humans would adopt the intentional stance in order to predict and explain the actions of humans and robots, at the explicit level (acceptance/rejection of a description that uses mechanistic/mentalistic vocabulary), and at the implicit level (participants' response time in the task). In addition, this paper investigates the relation between likelihood of anthropomorphism and intention attribution and attempt to delineate phenotypes of the tendency towards adopting intentional stance.

4.2.1 The Intentional Stance

According to Dennett (Dennett, 1971; Dennett, 1989), in predicting and explaining behaviour of other systems, humans employ the strategy that is most efficient. For example, we expect a kite to fly, as it is designed to do so. If one of its surfaces is broken, we expect that the kite would not rise high to the sky. Predicting behaviour of a system using the designed function of the object is what Dennett defined as adopting the design stance. Alternatively, when it comes to other humans, the most efficient stance is the intentional stance, i.e., predicting behaviours based on inferred mental states. It has been extensively demonstrated that healthy adults spontaneously adopt the intentional stance towards other humans [for a review, see 3]. Interestingly, humans also have the tendency to adopt intentional stance to non-human objects (Abu-Akel et al., 2020; Happé and Frith, 1995; Heider and Simmel, 1944; Zwickel, 2009). Nevertheless, it is yet to be understood whether humans employ a similar social cognitive mechanism during interaction with robots, and what conditions determine whether humans would choose one over the other stance.

4.2.2 Adopting the intentional stance towards robots and the Intentional Stance Test (IST)

In recent years, multiple studies have attempted to operationalize the adoption of intentional stance towards humanoid robots (see Perez-Osorio and Wykowska, 2020 and Schellen and Wykowska, 2019, for review). For example, Thellman et al., (2017) evaluated whether people would use the same causal models to explain the behaviour of a human and a humanoid robot. The authors reported that participants used (to a certain degree) similar intentional explanations for behaviour from both agents. They also showed that observers were less confident assigning intentional explanations to the robot. In a similar fashion, Marchesi and colleagues (2019) proposed a test – the InStance Test (IST) - to evaluate whether people employ mentalistic or mechanistic explanations to describe the behaviour of a humanoid robot. IST consisted of pictures (34 scenarios) showing the iCub robot (Metta et al., 2010) performing some daily-life actions, each one with a sequence of events depicted in three images (See figure 4.1). Participants were asked to select the description (mentalistic or mechanistic) that, according to them, better explained the behaviour of the robot. Authors reported that although participants had a bias toward the mechanistic explanations, in some circumstances, they opted for mentalistic explanations. Importantly, the authors suggest that adopting the intentional stance towards artificial agents does not imply that observers assume that those agents have indeed mental states that govern their behaviour (see also Nass and Moon, 2000). Instead, people might predict artificial agents' behaviour "as if" they would have mental states. Therefore, people might switch between adopting the intentional and design stances when they try to explain the behaviour of an (artificial) agent. Interestingly, the results of Marchesi and colleagues (2019) study revealed a binomial distribution on the responses, as participants tended to "polarize" in the preferences for either mentalistic or mechanistic explanations of the robot actions. Furthermore, the human tendency to ascribe human-like characteristics, such as intentionality or emotions, to inanimate entities (known as anthropomorphism, Airenti, 2018; Cullen et al., 2013), might also contribute to the attribution of human mental capabilities to robots (Epley et al., 2007; Kiesler et al., 2008; Spatola, 2019). Arguably, the attribution of human-like characteristics activates the general knowledge we have about humans (Airenti, 2018; Airenti et al., 2019; Cullen et al., 2013) and the activation of neural representation related to the perception of other humans (Chaminade et al., 2012; H. L. Gallagher et al., 2002; Özdem et al., 2017). For example, factors like appearance (Martini et al., 2016), social gaze (Wiese et al., 2018; Wykowska, Kajopoulos, Obando-Leiton et al., 2015), expectations and individual differences (Ghiglino, De Tommaso et al., 2020; Perez-Osorio et al., 2019), variability of behaviour (Marchesi et al., 2020), and perception of human likeness (Willemse and Wykowska, 2019), might also impact the level of attributed anthropomorphism and therefore help to induce the adoption of the intentional stance towards a humanoid robot. Finally, the adoption of intentional stance and anthropomorphic attributions may be influenced by individual characteristics (e.g., dispositional, developmental and cultural) and the context in which individuals are observing an agent (Epley et al., 2007). In sum, it is pivotal to investigate the individual tendency of humans to adopt the intentional stance, to take into account the “intentional stance tendency phenotype” of each individual

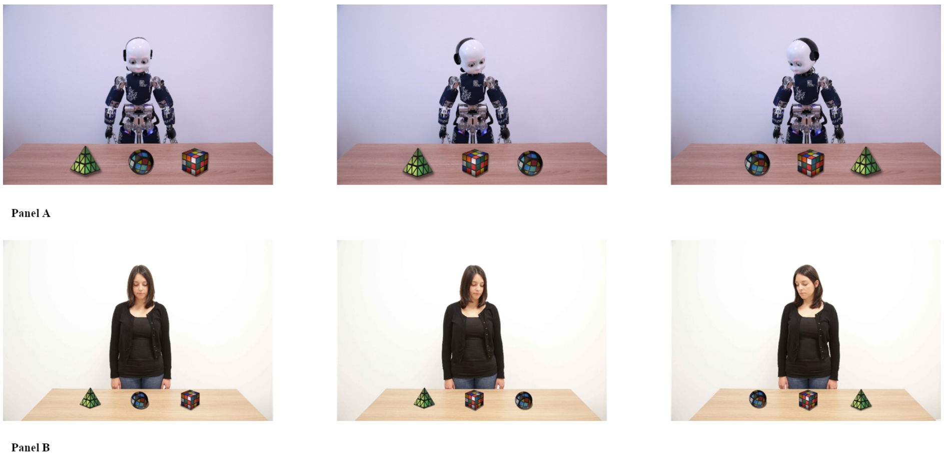


Figure 4.1. Examples of items depicting either the iCub robot (Panel A, top) or the human character, Sara (Panel B, bottom) as protagonists of the scenarios.

4.2.3 Task adaptation

In the original IST of Marchesi et al. (2019) participants were asked to provide their subjective rating (on a continuous scale) regarding the descriptions of the scenarios they viewed. To have a further insight into the processes leading to the adoption of the intentional stance, we adapted Marchesi et al. task to a design with a two-alternative-forced-choice task (2AFC) (Naefgen et al., 2018). Such adaptation grants for collecting participants’ response times (RTs), a classical mental chronometry measure of cognitive processes used in experimental psychology and HRI (Donders, 1969; Jensen, 1987; Kompatsiari, Pérez-Osorio et al., 2018; Kompatsiari, Ciardo et al., 2018a, 2018b; Luce et al., 1986; Posner, 1978; Spatola, 2019; Stenzel et al., 2012). Therefore, our adaptation makes it possible to evaluate the adoption of the intentional stance at a more implicit level. Explicit attitudes operate on a conscious level and are generally measured through explicit self-reports (e.g., questionnaires), while implicit attitudes rely on unconscious and automatic processes, and are typically assessed via implicit measures (e.g. response time paradigms, implicit association test) (De Houwer et al., 2009). Research suggests that implicit attitudes might constitute better predictors of future intentions and behaviours (Kurdi et al., 2019), and thus be more representative of real attitudes than explicit declarations. Implicit measures have also proved to be well-equipped to predict the behavioural consequences of individuals’ implicit representations (Friese et al., 2008; Kurdi et al., 2019).

4.2.4 Aims and hypotheses

The aims of the present study were to (i) test whether participants’ response times would differ while processing the two stances and (ii) replicate and validate the results of Marchesi et al. (2019). Furthermore, we included a control condition in which a human actor was presented in the same actions as the robot agent, allowing us to compare participants’ behaviour between the two agents, thereby further validating the IST. We hypothesized that participants would (1) accept more often (explicit level) and have faster response times (implicit level) for mentalistic statements in human scenarios. We expected that participants would be less likely to accept (explicit level) and have slower response times (implicit level) for mentalistic statements in the scenarios with iCub. The secondary aim of the study was to (2) investigate the inter-individual differences in the tendency to anthropomorphize by measuring how the likelihood to anthropomorphize robots may predict the likelihood of adopting the intentional stance towards a humanoid robot.

4.3 Materials and Methods

Forty-one participants were recruited for the experiment and received a monetary compensation of 15€. One participant was excluded from the analysis due to a technical problem in one of the blocks, leading to a final sample of N = 40 (mean age: 24.77; SD: 4.01; 25 women). All participants reported no history of psychiatric or neurological diseases and signed the informed consent before the beginning of the experimental session. The study was approved by the local Ethical Committee (Comitato Etico Regione Liguria) and was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). All participants were naïve to the purposes of the study but were debriefed after the experiment.

4.3.1 Experimental design

The new version of the IST measured the ratio of acceptance/rejection of mentalistic/mechanistic statements and response times of participants' choices. That is, two major modifications were made to the original IST. First, only one statement (mechanistic or mentalistic) was presented at a time for each scenario; second, instead of a slider, participants were asked to accept or reject each single description (mentalistic/mechanistic) associated with each scenario. This would result in a two-alternative forced-choice task. This design provides measures of participants' performance on two levels: an implicit level (response times) and an explicit level (the acceptance/rejection rate). The test had two blocks, one with iCub was the main character (robot block) and another one in which a human (named Sara – the human block). Each trial consisted in a scenario taken from the original IST depicting the character performing an action and a sentence describing the agents' behaviour. The description presented below the scenarios was either mechanistic ("iCub/Sara was unbalanced for a moment") or mentalistic ("iCub/Sara was trying to cheat by looking at opponent's cards"). Each one of the original IST 34 scenarios from [5] was shown twice: once with the mechanistic description and once with the mentalistic description. This led to 68 items per block (robot and human). The items for the robot block were taken from the InStance Test (IST, Marchesi et al., 2019); while items for the human consisted of photos of equivalent scenarios but depicting a human character (Sara) instead of the iCub (see Fig 4.1). Note that the name of the character (iCub or Sara) contains an equal number of letters, in order to avoid biases during the reading process. The descriptions were identical for both agents for each corresponding scenario. The presentation of the items within the block was randomized, and the order of the human and robot blocks was counterbalanced across participants.

4.3.2 Apparatus and procedure

The experiment was performed inside a dimly lit room. Items were presented on a 27 LCD screen (resolution: 1920 × 1080). Participants were seated approximately at 62 cm distance from the screen. After participants read and signed the consent form, they were instructed about the task. They were told that their task would be to judge whether the sentence presented below a scenario plausibly explained what was depicted in the scenario. Participants were asked to respond by pressing two keys on the numeric pad of the keyboard (8 for "yes" and 2 for "no"). Before each block, participants had a practice trial with the agent corresponding to the respective block. The trial procedure was the following: 1- a fixation dot appeared at the center of the screen for 500 ms; 2- the sequence of three images was presented at once at the center of the screen for 3000 ms. 3- the sentence appeared below the scenario with response options "yes" or "no" and remained until the keypress. Participants' response times were measured from the onset of the sentence to the keypress. Before the beginning of the human block, we asked participants to complete the Individual Differences in Anthropomorphism Questionnaire (IDAQ), developed by Waytz, Cacioppo and Epley (Waytz, Cacioppo et al., 2010. The reason to present the IDAQ to the participants before the human block was two folded: (1) to avoid any anthropomorphism bias towards the robot; (2) since observing a human may already induce anthropomorphism in the subsequent questionnaire, affecting the IDAQ score, a safe solution was to administer it before any exposure to the human agent.

4.3.3 The Individual Differences in Anthropomorphism Questionnaire (IDAQ)

In order to evaluate whether the individual differences in the general tendency towards anthropomorphism influences the attribution of mental states to artificial agents, we administered the Individual Differences in Anthropomorphism Questionnaire (IDAQ), by Wayts, Cacioppo and Epley (2010). The IDAQ investigates participants' stable and general individual differences in anthropomorphism with several items on various agents. The scale of the IDAQ items ranges from 0 to 10, with a Cronbach alpha ≥ .82. Examples of the items are: "To what extent does the average insect have a mind of its own?", "To what extent do cows have intentions?" and "To what extent does the average fish have free will?". The authors argue that the mental state attributes used in the items (i.e., "have a mind of its own", "have intentions", and "have free will") have been proven to be reliable characteristics of human-uniqueness and higher order cognitive processes (Demoulin et al., 2004; Haslam et al., 2005; Kozak et al., 2006). According to the authors, individual scores in IDAQ allow predicting three major factors related to anthropomorphism that might have consequences in everyday life: 1- the moral care and concern towards other agents; 2- the trust and responsibility attributed to other agents and 3- the extent to which an agent is considered as root of social influence on the self. Therefore, the IDAQ can be a useful tool to investigate at the individual level, how the likelihood of anthropomorphism can relate to the adoption of the intentional stance strategy towards a robotic agent.

4.4 Results

Statistical analysis were conducted in R Studio (version $4.0.2$, Team R) using the package lme4 (Bates, 2018)

4.4.1 Acceptance/Rejection of mentalistic description

To investigate the probability of acceptance or rejection of a description, we performed a Generalized Linear Mixed Model. We considered the choice (acceptance/rejection) as variable of interest, while the type of description (mechanistic vs. mentalistic) and the agent (human vs. robot) as fixed within- participants factors. Participant and scenario were considered as random factors. Results showed a significant interaction effect between type of description*agent: b = -2.30, z (5433) = -19.03, p = $< .001$, CI95% $[-2.53; -2.06]$ (comparison with null model: $\chi^2(1)$ = 381.03, p = $< .001$). As visible in Figure 4.2, participants showed an inverse pattern of mentalistic/mechanistic description acceptance for the two agents. When asked explicitly, they accepted more often the mentalistic description for a human, while they accepted more often the mechanistic description. We further investigated the contrast between mechanistic and mentalistic description with planned post-hoc pairwise comparisons (Tukey's HSD correction for multiple comparisons): human agent: b= -1.70, p= $< .001$; robot agent: b= $.60$, p= $< .001$.

4.4.2 Response Times results

The Linear Integrated Speed-Accuracy Score (LISAS) approach

Trials with RTs slower or faster than 3 standard deviations per condition (mechanistic vs. mentalistic; human vs. robot) for each participant were considered outliers and then removed from analyses. This resulted in the exclusion of 1.27% (69) of the trials (this filtering procedure has the advantage of taking out extreme values without affecting the data of one specific condition or of one participant in particular). To investigate participants' response times during the task, we analyzed the relation between response times, agent presented and descriptions, correcting participants' response time using the Linear Integrated Speed-Accuracy Score (LISAS, Vandierendonck, 2017).

This transformation allows considering a linear relationship between response times (RT) and the proportion of decision (PD) toward a response 0 and an alternative response 1. This makes it possible to evaluate participants' response bias, including both the decision type and the decision speed. The main advantage of this analysis method (compared to standard RT analysis) consists in including and integrating the type of choice and response time for

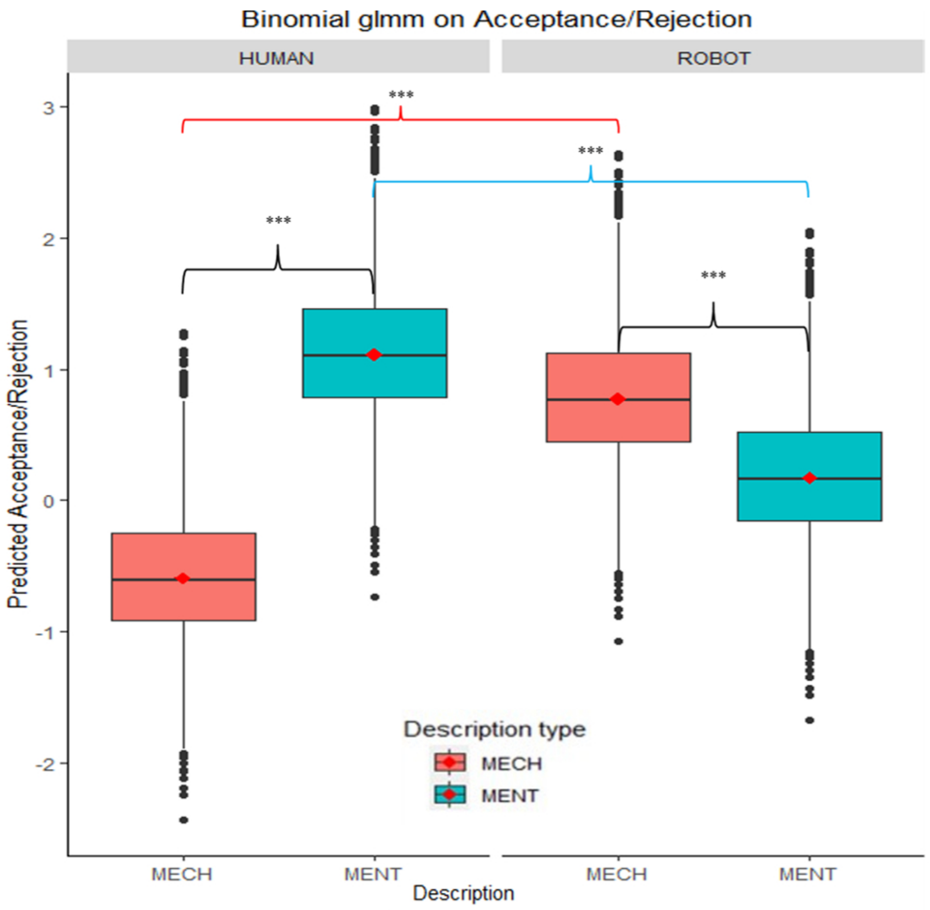


Figure 4.2. interaction effect in GLMM between description*agent. The y-axis reports the predicted probability of acceptance/rejection on a binomial scale: positive numbers represents the probability of acceptance (yes) and the negative numbers represent the probability of rejection (no). * = p $< .05$, ** = p $< .01$, *** = $p < .001$

each condition (and their respective variance) in a single score. We considered the proportion of choice (acceptance/rejection) associated with the proportion of mentalistic description choices (i.e., 1) contrasted to the mechanistic description choices (i.e., 0).

$$\text{LISAS} = \text{RT}_x + \frac{\text{S}_{RT}}{\text{S}_{PD}} \times \text{PD}_x, \tag{4.1}$$

Where RTx is the mean response time in condition x, PDx is the proportion of choice (1 vs 0) in condition x and SRT and SPD is the overall standard deviation for response times and proportion of choices, respectively. Two additional participants that exclusively chose mechanistic or mentalistic responses in all trials in at least one condition were excluded from the analysis, since their data are to be considered not reliable. The remaining sample size was equal to 38 (N = 38). We conducted a linear mixed effect model with agent and description as fixed factors and RTs -LISAS corrected as a variable of interest. Participant was considered as random factor. We found an interaction effect between agent and description (b = -2467.2, t (38) = -3.94, p = $< .001$, CI95% [-3688.84; -1245.60]) (Fig. 4.3), showing that participants were faster in accepting a mentalistic description for the human agent, relative to the robotic agent. Contrasts with Bonferroni correction showed that participants were faster in choosing the mentalistic descriptions in the human compared to robot trials, F(1, 37) = 10.34;

p = .003; CI95% [495.16; 2182.41]. Conversely, they were faster choosing the mechanistic descriptions in robot compared to human trials, F(1, 37) = 6.93; p = .012; CI95% [259.66; 1997.22]. The difference between response times between mentalistic and mechanistic was significant only in human trials, F(1, 37) = 74.46; p = < .001; CI95% [1998.01; 3224.26], with faster responses for mentalistic compared to mechanistic descriptions.

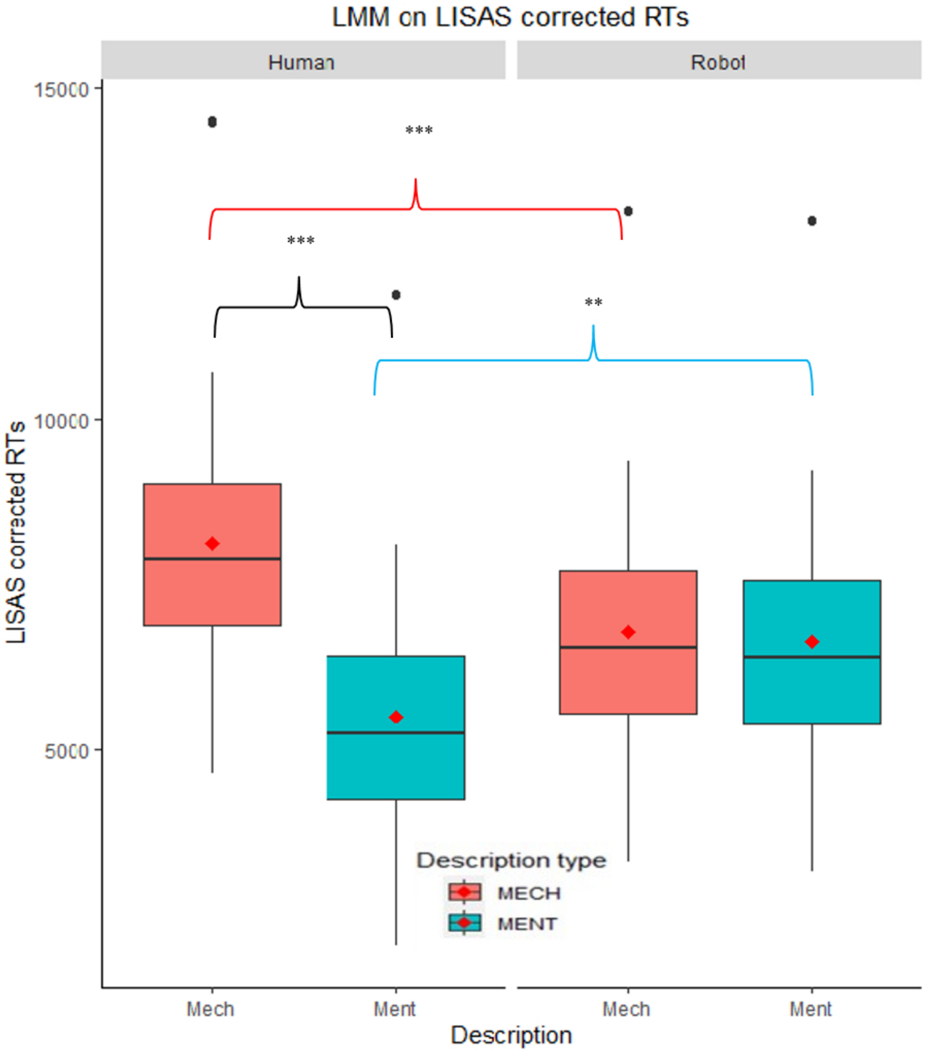


Figure 4.3. Interaction effect on LISAS scores agent*description.* = p < .05, ** = p < .01, *** = p< .001.

Clustering approach on the individual tendency to accept a mentalistic description for a robot and the IDAQ score

To evaluate whether we could categorize participants in order to predict their tendency for anthropomorphism, we first processed a two-step clustering using the mentalistic (vs. mechanistic) bias of participants measured by LISAS score (Vandierendonck, 2017). Given the way LISAS score are computed, they contain information about the relationship between the response times and the proportion of choice (acceptance/rejection) associated with the mentalistic and mechanistic descriptions. The result is a score that represents the mentalistic bias of each participant. We preferred the 2 step-clustering approach (compared to correlation) due to the high inter-individual variability in RTs range, that biases the correlation results with the normative IDAQ scale. As we used LISAS score and only have one data point for each participant for each condition (4 in total), a normalization would be highly sensitive to this inter-individual variability. We thus preferred the cluster approach that provides a common structure to all participants. Before proceeding with the clustering, we first calculated the difference between the LISAS scores of the mentalistic option in the human and the robot block. For detailed description of the clustering process, see Appendix B, section B.2 The clustering proposed a solution with a 2 clusters' matrix with a 1.71 ratio of sizes and a cluster silhouette quality = 0.7. Cluster silhouette measures the cohesion and separation of clusters purple and their partitioning quality (good fit) (for similar procedure, see Spatola and Urbanska, 2020)). According to cluster silhouette and cluster comparison, analyses argue for a high tendency vs. low tendency group in mentalistic attributions toward robot. Detailed descriptive metrics of the clusters are reported in Appendix B. Indeed, participants in the first cluster chose the mentalistic description for the robot more rapidly compared to participants in the second cluster (Cluster 1 mean_LISAS: 1544.41; Cluster 2 mean_LISAS: -2687.60), F(1, 37) = 56.94, $p < .001$, CI95% $[-4105.78; -1239.28]$. Interestingly, the clusters also predicted participants' tendency to anthropomorphize (IDAQ scores). Participants in the cluster 1 scored higher on the IDAQ than participants in the cluster 2, F(1, 37) = 4.28, p = .046, CI95% [.29, 30.34]

4.5 Discussion

The aims of the present paper were to: (i) test whether participants' response times would differ while deciding to adopt one of the two stances towards a human or the robot and (ii) replicate and validate the results of Marchesi et al. (2019). In addition, we aimed at examining the relationship between likelihood of adopting the intentional stance and individual tendency to anthropomorphize. In line with our hypothesis, at the explicit level (Acceptance/Rejection of mentalistic description), we found that mechanistic descriptions were attributed to the robot to a higher extent than mentalistic descriptions. The inverse pattern was found for the human, where participants were more likely to attribute a mentalistic description to the depicted human. Indeed, these results replicate the results of Marchesi et al. (2019) with a different design and measure. In Marchesi et al. (2019), the authors reported an average IST score of 40.73 on a scale from 0 to 100 (where 0 would be totally mechanistic and 100 totally mentalistic), showing a bias towards the mechanistic description for the robot. The present study made it possible to use objective measure such as response times as a probe for adopting the intentional stance. The results showed that participants' response times were faster in mentalistic descriptions for a human compared to a robot, but also compared to mechanistic descriptions. Furthermore, participants were faster in responding in mechanistic descriptions for the robot, as compared to the human. Interestingly, when looking only at the robotic agent condition, no significant difference in the response times between mechanistic and mentalistic descriptions was observed. This is in contrast to the human condition, where the mentalistic descriptions were evaluated faster than mechanistic descriptions. This pattern suggests that the intentional stance might be the default strategy to explain other humans' behaviours, both at the explicit and implicit level. When it comes to a robotic agent, participants might not find any of the descriptions as unlikely as the mechanistic descriptions for the human. Meaning that both stances are compatible with participants' mindset, at least at the implicit level. However, at the explicit level, participants are more likely to accept mechanistic descriptions of the robot actions. This might reflect that both the intentional and design stance are strategies that could be adopted to explain an artificial agent's behaviour. Some authors assume that intentional stance would be the default stance in general (relying on the social cognition system) (A. I. Jack, Dawson and Norr, 2013; Schilbach et al., 2008). Indeed, neural areas (e.g., the superior temporal sulcus, lateral fusiform gyrus, medial prefrontal cortex, posterior cingulate, insula and amygdala) associated with mentalizing (i.e., the attribution of mental states to others), belong to the default mode network (Spreng and Andrews-Hanna, 2015), suggesting that this key social cognition mechanism is the default mode of functioning of the brain (A. I. Jack, Dawson, Begany et al., 2013; Spreng and Andrews-Hanna, 2015; Spreng et al., 2016). However, although a default system, it can be controllable [53], this would imply that at the earlier (implicit) stage of processing, individuals may adopt the intentional stance towards a robot, but upon deliberate reflection, they would quickly identify the robot as an artefact, and they would reject the mentalistic descriptions. Results from cluster analyses revealed that participants differed in the tendency to adopt the intentional stance, showing the presence of individual differences in explaining a humanoid robot's behaviour with a mentalistic vocabulary (Nass and Moon, 2000). Moreover, the results showed that those participants that were faster in choosing a mentalistic description for the robot, also scored higher in the IDAQ questionnaire, as compared to those who were slower in selecting mentalistic descriptions. This is in line with previous literature (Waytz, Morewedge et al., 2010) and shows that a high level of anthropomorphism is associated with a high tendency to adopt the intentional stance. Moreover, recent literature reports that participants' tendency to adopt the intentional or design stance, could be influenced by individual differences (Bossi et al., 2020; Ghiglino, De Tommaso et al., 2020; Spatola, 2019), the need for cognition (Cacioppo and Petty, 1982; Cohen et al., 1955) or the expectations about robots (Perez-Osorio et al., 2019) but also the embedded cultural values and other factors (Epley et al., 2007). In line with these previous studies (Epley et al., 2007), our results argue that it might be useful to delineate personality phenotypes to describe clusters of individual characteristics, such as the tendency to anthropomorphize, since they might be predictive of future interactions in HRI.

4.6 Limitation and future work

The present adaptation of Marchesi et al. (2019) material, despite using RT, remains a choice task between two options about a complex semantic sentence. According to the RT that are longer (overall LISAS mean: 6749.77)) than usual implicit test using implicit association (RT mean = 2500-3000) (De Houwer et al., 2009; Greenwald et al., 2002), we cannot rule out that the decision could suffer from biases inherent to explicit reasoning engaged by the participants. Moreover, although we cannot exclude that varying confederates across different items may affect participants' performance, we highlight that in (Marchesi et al., 2019), the authors showed that humans included in scenarios do not cause substantially different scores relative to non-human scenarios. Future, tailored, work should investigate these aspects.

4.7 Conclusions

Our findings indicate that it is possible to design a test probing adoption of intentional stance with objective measures, such as response times. Furthermore, we show that the individual bias in the adoption of the intentional stance towards humanoid robots should be investigated both on the explicit and implicit level, as those levels are not necessarily always in accordance (see also Ghiglino, De Tommaso et al., 2020). Additionally, the evaluation of individual tendencies to adopt the intentional stance and to anthropomorphize allowed exploring more in depth the connection between individual preferences for intentional stance towards artificial agents and anthropomorphism.

Part III

How individual differences shape the adoption of the intentional stance towards an embodied humanoid robot

Chapter 5

Publication 3: Don't overthink: fast decision-making combined with behaviour variability perceived as more human-like

This chapter presents the accepted version of the manuscript from the following publication:

Marchesi, S., Perez-Osorio, J., De Tommaso, D. and Wykowska, A., (2020). "Don't overthink: Fast decision making combined with behaviour variability perceived as more human-like," 29th IEEE Int. Conf. Robot Hum. Interact. Commun. RO-MAN 2020, pp. 54–59, 2020, doi:10.1109/RO-MAN47096.2020.9223522. Materials and datasets are available at https://osf.io/3fznv/

Author Contributions: SM, JP-O and AW designed the study. SM collected the data. SM and JP-O analyzed the data. DD programmed the robot behaviours. SM, JP-O, and AW discussed the data and wrote the manuscript.

5.1 Abstract

Understanding the human cognitive processes involved in the interaction with artificial agents is crucial for designing socially capable robots. During social interactions, humans tend to explain and predict others' behaviour adopting the intentional stance, that is, assuming that mental states drive behaviour. However, the question of whether humans would adopt the same strategy with artificial agents remains unanswered. The present study aimed at identifying whether the type of behaviour exhibited by the robot has an impact on the attribution of mentalistic explanations of behaviour. We employed the Instance Test (IST) pre and post-observation of two types of behaviour (decisive or hesitant). We found that decisive behaviour, with rare and unexpected "hesitant" behaviours, lead to more mentalistic attributions relative to behaviour that was primarily hesitant. Findings suggest that higher expectations regarding the robots' capabilities and the characteristics of the behaviour might lead to more mentalistic descriptions.

5.2 Introduction

Understanding human cognitive processes involved in interactions with artificial agents is crucial for designing robots that are supposed to enter our daily lives. Many studies have focused on how the attribution of human characteristics, like emotions, intentions, or mental states to non-human agents, plays a role in social human-robot interaction (Epley et al., 2007). More recently, several studies have examined whether humans can adopt the intentional stance towards humanoid robots (Marchesi et al., 2019; Thellman et al., 2017; Thellman and Ziemke, 2020). According to Dennett (Dennett, 1971; Dennett, 1989), adopting the intentional stance is a strategy to predict others' behaviour with reference to mental states. When we use this strategy, we understand the other agents have their understanding of the world, separate from ours. Importantly, we also assume that those mental states drive people's actions. Dennett distinguishes the intentional stance from the design stance. We use the design stance to understand and predict a device's behaviour based on its purpose (Dennett, 1971; Dennett, 1989). For instance, a kite will fly depending on the strength and direction of the wind, not depending on whether it has the intention to fly. Humans attribute emotions, intentions, and mental states to agents that show some level of autonomous agency. This tendency to anthropomorphize non-human agents has been considered a default and automatic psychological process (Mitchell et al., 1997). According to Fisher (Fisher, 1991), anthropomorphizing could be either interpretative or imaginative. That is, attributing human-like characteristics could be produced by observation of behaviour (interpretative anthropomorphism) or by endowing human capabilities to non-human agents by imagining what they can or cannot do (imaginative anthropomorphism). In this context, it is plausible to think that not only experiences or direct observation can produce attribution of human-like characteristics, but also ideas about the capabilities of agents. For example, users can expect a humanoid robot to behave in a humanlike manner even if they have never interacted with a robot. What this theory highlight is that attribution of human-like characteristics to artificial agents might be the result of multiple factors. The physical appearance of the robot, designed to interact with humans, the context, the expectations, and assumptions that one might have regarding the robot and the previous experiences generate attributions of various traits towards the artificial agent. Social (especially humanoid) robots are an interesting case for examining factors contributing to adopting the intentional stance towards artificial agents. Humanoid robots are machines, but they have a human-like shape, and they are designed to interact with humans. This makes them agents in-between categories, potentially able to elicit the adoption of the intentional stance (for a review see (Perez-Osorio and Wykowska, 2020)), even though they are just machines. A recent study by Marchesi et al. (Marchesi et al., 2019) designed the InStance Test (IST), a novel tool to assess the adoption of the intentional stance towards social robots. The questionnaire was administered online. Respondents observed 34 scenarios depicting the iCub robot (Metta et al., 2010) and had to choose between two options on how to interpret iCub's behaviour. One option described the behaviour of the robot with reference to mental states and using a mentalistic vocabulary (intentional stance). The second one, explained the behaviour of the robot in mechanical terms using a technical vocabulary (design stance). Results showed a general tendency to prefer the mechanistic descriptions; however, some mentalistic explanations have also been chosen. Interestingly, some scenarios elicited more mentalistic descriptions than mechanistic ones. The authors concluded that the adoption of the intentional or design stance is context-dependent and might be associated with the depicted behaviour. Importantly, these results showed that if people choose to describe the robot behaviour in a mentalistic manner in some cases, it means that they are capable in general of adopting the intentional stance towards artificial agents. Knowing that people can, in some cases, adopt the intentional stance toward robots, it is important to understand what factors contribute to it. In other words, what are the conditions for adopting the intentional stance. Previous studies have theorized whether people might adopt the intentional stance towards humanoid robots (Chaminade et al., 2012; H. L. Gallagher et al., 2002; Perez-Osorio and Wykowska, 2020) and even investigated the causality approach towards robot behaviour (Thellman et al., 2017). For example, (Thellman et al., 2017) found that participants used mentalistic explanations to describe the behaviour of the robot. Nonetheless, the confidence scores of mentalistic explanations were lower compared to the confidence scores of human actions. This indicates that some degree of cognitive dissonance in describing robots with mentalistic terms. In the case of (Marchesi et al., 2019), results showed that some participants were more likely to choose mechanistic scores, while others tended to give more mentalistic explanations. It might be that some participants had certain expectations regarding the human capabilities of the robot and used predominantly mentalistic explanations. In contrast, some other participants had much more mechanistic pre- assumptions. Perez-Osorio et al. (Perez-Osorio et al., 2019) explored this alternative, creating a questionnaire that evaluates the expectations of participants regarding iCub. Results showed that people who had high expectations about iCub's behaviours were more prone to explain the behaviour of the robot with mentalistic explanations, as compared to those who had lower expectations. Participants' expectations and individual differences are one potential factor contributing to the likelihood of adopting the intentional stance. Robot appearance and behaviour are obviously other key factors. Ghiglino et al. (Ghiglino, Willemse, De Tommaso et al., 2020) asked participants to score the human likeness of iCub when it moved its eyes between two fixation points. Slow but variable profile movements yielded higher humanness scores relative to constant speeds. Moreover, human-like eye movements were more engaging and evoked spontaneous attentional following. These findings suggest that the attribution of human-like characteristics might depend on the combination of external factors (subtle variations in social signals) and internal states of the observer, that is, for example, higher-level expectations regarding the behaviour of artificial agents. In the present study, we aimed at evaluating whether the attribution of mechanistic/ mentalistic explanations of robot behaviour might be influenced by the type of behaviour the robot exhibits, not at the low-level of movement characteristics, but at a level, which might manifest various complex cognitive processes. We designed a paradigm in which two different groups of participants observed a robot performing a cognitive task. Importantly, the robot showed either a predominantly "hesitant" or a predominantly "decisive" behaviour during a decision-making task. The hesitant behaviour took longer to respond and exhibited more body and eye movements compared to the other condition. With the present study, we examined whether being exposed more frequently to a hesitant behaviour in a humanoid robot would modulate adoption of the intentional stance. We examined whether the frequency of the hesitant behaviour would modulate adoption of the intentional stance. In order to measure the effects of the manipulation, we divided the Instance Test (IST (Marchesi et al., 2019)) into two parts and administered the first half before the observation of the robot and the other half right after the session with the robot. Additionally, with multiple questionnaires, we explored possible correlations between the IST scores, attitudes toward robots, personality traits, and expectations.

5.3 Methods

5.3.1 Sample

Forty-two participants (mean age = 24.72, SD = 3.337, 19 men, 4 left-handed) took part in the experiment and received monetary compensation (15€). None of the participants reported a clinical history of psychiatric or neurological diseases. All participants were naïve to the purpose of the experiment and signed written informed consent. At the end of the experimental session, each participant was debriefed by the experimenter. The study was approved by the local Ethical Committee (Comitato Etico Regione Liguria), and it was conducted following the Code of Ethics of the World Medical Association (Declaration of Helsinki).

5.3.2 Procedure

During the experiment, participants first read and signed the written informed consent. Then, they completed the first group of items (17 out of the 34) from the InStance Test before the interaction (IST PRE), and the RobEx questionnaire (Perez-Osorio et al., 2019), Frankenstein Syndrome Questionnaire, FSQ (Nomura et al., 2015), the Negative Attitude Towards Robots questionnaire, NARS (Nomura, 2014), and the Robotic Social Attributes Scale, RoSAS (Carpinella et al., 2017). At this point, participants moved inside a cabin where they shared the space with the robot and sat down in front of iCub. Two 21" screens were placed on the desk between paticipants and the robot, one facing the robot and the other facing the participants (see Figure 6.1). The screens were placed at their lowest height and tilted so that participants could easily see both the screen and the robot in front of them. Participants were instructed to observe the behaviour of the robot performing a cognitive task. More in detail, participants were told that the iCub was required to find similarities between three objects presented on the screen. The similarities could be either feature characteristics (like shape or colour) or semantic links (same category, like animals). During this phase of the task, iCub observed two first items presented on the screen and then "had to select" one out of three additional items. The other items were presented vertically and perpendicularly on the right side of the screen (see figure 6.1, for example). After the robot made a choice, the participants were prompted to respond whether they thought the robot chose the item based on a feature-

or category-based link among the items. Through the decision phase, the robot could exhibit either a hesitant behaviour or an automatic behaviour. Two groups were created based on the frequency of hesitation displayed by the robot robot behaviour was selected randomly before the beginning of the experiment. Group 1 observed mostly hesitant behaviour of the robot (hesitation in 80% of the total trials) while Group 2 observed mainly automatic behaviour (hesitation in 20% of the total trials). Video showing the experimental setup and the trial sequence of events are available at: https://osf.io/3fznv/. After 40 trials, participants proceeded to complete the second half (17 items) from the InStance Test (IST POST), Godspeed questionnaire (Bartneck et al., 2009) and Big Five Inventory (Goldberg, 1993). Finally, participants were debriefed. Regarding the InStance Test, participants responded using a slider. On each extreme there was a statement, either mentalistic or mechanistic. Participants moved the slider towards one of the sides and were asked to indicate how much that statement explains the robot behaviour. The slider scores were converted to a single score, InStance Score (ISS) from 0 to 100, being 0 the most mechanistic and 100 the most mentalistic. All the scores in the present study are ISS. The 34 scenarios were randomly selected for every participant and divided into two halves at the beginning of the experiment.

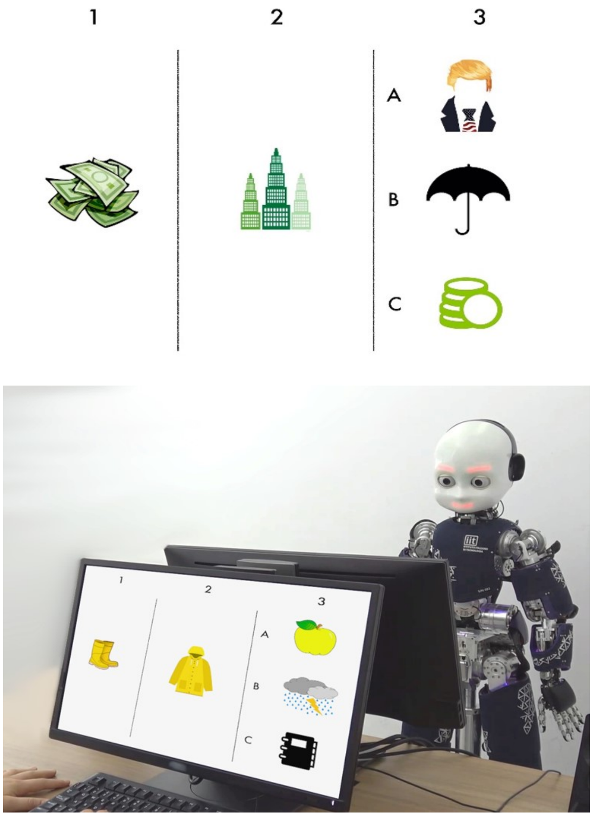


Figure 5.1. Stimulus example (upper frame) and setup with the iCub robot. (40 stimuli sets were used in the experiment. The participants saw on the screen the same images presented to the robot)

5.3.3 Apparatus and stimuli

The experiment required two 21’’ screens placed in the middle of the desk, one facing the robot and the second towards the participant. The displays were tilted back with an angle of 12° to allow participants to see the robot (see Fig. 6.1). Participants responded using a standard keyboard. The robot and the experimental presentation of stimuli and data collection were connected to a single workstation (SSD 250 Gb, 32 Gb RAM, Intel i7 7700CPU@3.60 GHz 4 cores, GPU GeForce GTX 1060 3Gb, running Ubuntu 16.04 LTS) through a peer-to-peer Ethernet connection and to both screens with DP cables holding a 100Hz refresh rate. The robot and the experiment were controlled using OpenSesame (Mathôt et al., 2012), an open-source graphical experimental builder based on Python programming language. Specifically, the iCub robot behaviours have been implemented using the YARP (Metta et al., 2006) Python wrappers. To command the robot to look at a specific location in space, a 6-DOF gaze controller has been used (Roncone et al., 2016). The iKinGazeCtrl controller uses inverse kinematics to find eyes and neck positions to make the robot look at specific 3D Cartesian coordinates. Target locations were predefined so that the poses of the robot were comparable across trials. Regarding the other movements of the robot, specifically the ones related to the hip, we used the YARP position controller that generates joints trajectories following the minimum jerk profile. For creating the movement, we added to the predefined positions a uniform random generated displacement useful for making the robot movements slightly different in every trial. The items displayed on the screen among which iCub was supposed to choose matching objects were validated in a sample of 86 participants online through the SoSci Survey service. Combinations of items were selected a priori. Participants first saw two items, and from an arrangement of 3 items presented in the right of the screen, they had to choose one that shared either a feature or a categorical connection with the other two, similarly to the task presented in Figure 6.1. Based on this, we selected the items combinations that were linked by category or by feature by the majority of the participants. We performed a Chi - squared analysis on the frequencies of judgment for the category and feature-based links, excluding the catch items, for each combination. Then, we modified the ones that fail to reach significance (p = < .05), meaning links that were not categorical nor feature-based. Fourteen out of the first 40 combinations, did not reach the significance level and were replaced.

5.4 Results

To measure the impact of the behaviour exhibited by the robot on the attribution of mentalistic/mechanistic explanations to the robot, we measured the Instance Test (IST) pre and post-observation in two groups. Results showed that, in general, participants tended to interpret the behaviour of the robot in more mechanistic terms (M = 36.71, SD = 14.40). However, the comparison between conditions and groups showed an underlying difference. One way ANOVA revealed no significant difference between the groups on the IST-PRE (G1-PRE: M = 33.42, SD = 11.34; G2-PRE: M = 35.71, SD = 15.28); F(1,40)= .330,p = .585. However, the type of robot behaviour had a significant effect on IST-POST (G1- POST: M = 34.00, SD = 14.83; G2-POST: M = 43.66, SD = 14.52); F(1,40) = 4.554, p = .039 (see Figure 6.2).

This suggests that participants who observed iCub with predominantly automatic behaviour had a higher score in the IST-POST, compared to the group who observed mainly hesitant behaviours. Within the groups, there were no differences between pre and post (p > .159).

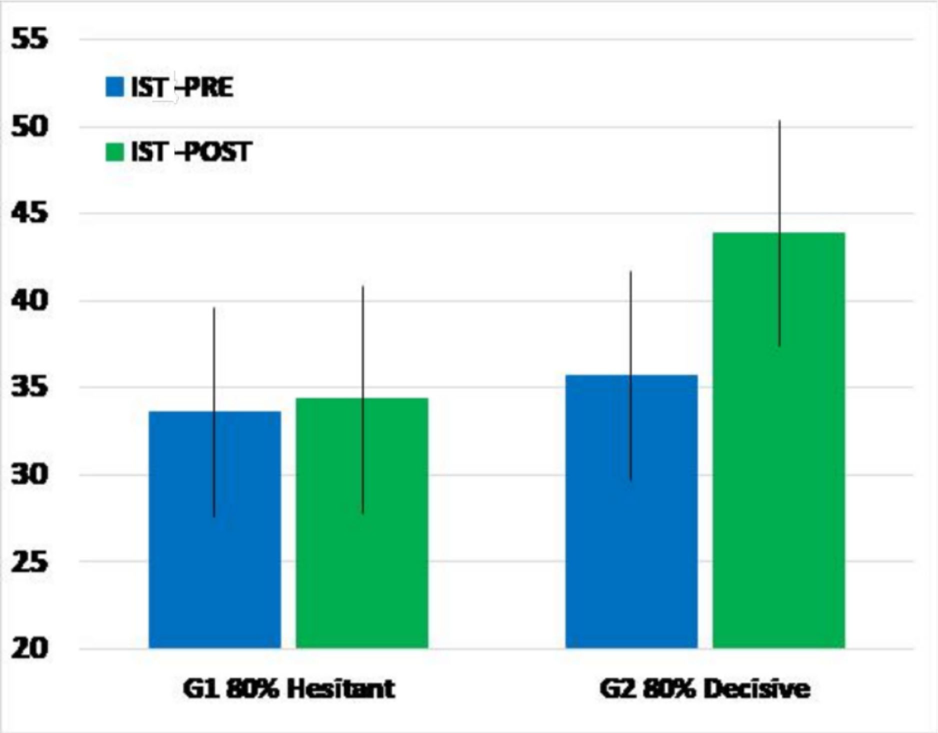


Figure 5.2. Instance Test scores for both groups. (The difference between groups on the POST measure was significant. Error bars correspond to the standard error of the mean (SEM)).)

Analyses of the additional questionnaires failed to reveal any difference between the groups (p > .124). Additionally, we conducted exploratory correlations among the different questionnaires subscales and the IST-PRE and –POST scores within each group. For Group 1 and Group 2, we found no correlations between IST-PRE and the subscales. Correlations were only observed with IST-POST. For Group 1, only a positive correlation between IST-POST and Godspeed Anthropomorphism scale [r(21) = .550, p = .020] was observed. However, for Group 2, several correlations were found for the IST-POST: on the positive direction with GS Anthropomorphism subscale [r(21) = .496, p = .022], GS Animacy subscale [r(21) = .473, p = .030], RoSaS Competence subscale [r(21) = .490, p = .024], BFI Conscientiousness subscale [r(21) = .436, p = .048], FSQ expectations subscale [r(21)=.534, p =.013] and finally a negative correlation for the RoSaS discomfort subscale [r(21) = -.509, p = .018] (see Table 6.1).

Interestingly, among the subscales, we found that the FSQ Expectations score had a strong positive correlation with GS Anthropomorphism subscale [r(21) = .730, p > .000], GS Animacy subscale [r (21) = .760, p > .000], RoSaS Competence subscale [r(21) = .624, p= .003]. We conducted a post-hoc ANCOVA analysis to compare the IST-POST whilst controlling for expectations. Levene's test and normality checks were carried out, and the assumptions met. There was a significant difference between groups [F(1,39) = 5.047, p = .03, partial Eta squared .115]. The effect of FSQ expectations scores was also significant [F(1,39) = 8.202, p = .007, partial Eta squared .174] This suggests that differences observed between groups are influenced, but not entirely, depending on participants expectations.

	Group 1: more hesitant				Group 2: more decisive			
	PRE	p	POST	p	PRE	p	POST	p
IST-PRE	1		.307	.175	1		.399	.073
IST-POST	.307	.175	1		−.399	.073	1	
GS Ant	.274	.229	.503*	.020	.065	.779	.496∗	.022
GS Ani	.223	.332	.278	.222	.151	.514	.473*	.030
GS Like	−.029	.900	.385	.085	.093	.687	.256	.262
FSQ Exp	.366	.103	.280	.219	−.028	.905	.534*	.013
FSQ Anx	−.432	.051	.049	.832	−.150	.517	−.215	.349
FSQ SocRis	−.288	.206	−.188	.415	−.238	.298	−.163	.480
FSQ Trust	.097	.676	−.195	.397	−.159	.490	.369	.100
NARS Emo	.132	.568	.070	.762	−.256	.262	.381	.088
NARS Sit	−.202	.381	.206	.371	−.140	.544	−.425	.055
NARS Soc	−.429	.052	.220	.339	−.131	.571	−.113	.626
RoSAS Com	.212	.356	.030	.898	−.165	.474	.490*	.024
RoSAS Dis	−.291	.200	−.058	.802	−.176	.445	−.509*	.018
RoSAS War	.275	.227	.431	.051	−.138	.552	.353	.166
BFI Extra	−.040	.865	−.116	.617	.419	.059	.211	.358
BFI Agree	−.055	.811	−.181	.433	.308	.174	.177	.443
BFI Cos	.247	.280	.060	.797	−.251	.273	.436*	.048
BFI Neuro	−.051	.828	.117	.614	−.372	.096	−.264	.248
BFI Open	−.032	.890	−.333	.140	−.097	.676	.410	.065

Green color highlights significant correlations.

Table 5.1. Exploratory correlations between InStance Test scores and other questionnaires. p-values were not subject to multiple comparison correction

5.5 Discussion

The present study aimed at evaluating whether a behaviour manifesting cognitive processes, exhibited by a humanoid robot has an impact on the attribution of mentalistic/ mechanistic explanations. We found that scores for the group of participants who observed a robot exhibiting a decisive behaviour on a decision-making task were higher (more mentalistic) relative to the scores of the group that observed the hesitant behaviour. Similar to previous findings, and despite the difference in the scores between groups, all the participants judged the behaviour of the robot mainly in mechanistic terms with tendencies towards the middle of the scale. Additionally, we found significant correlations between multiple subscales of various questionnaires and the POST-IST score including: Godspeed Anthropomorphism and Animacy, RoSaS competence, FSQ Expectations subscale, and BFI Conscientiousness, and an additional inverse correlation with RoSaS discomfort. Our results suggest that participants were sensitive to robot behaviour, which made a difference on the attribution of mechanistic explanations on the InStance Test after observation of behaviour. Interestingly, people were more likely to attribute mental capabilities to iCub performing a predominantly

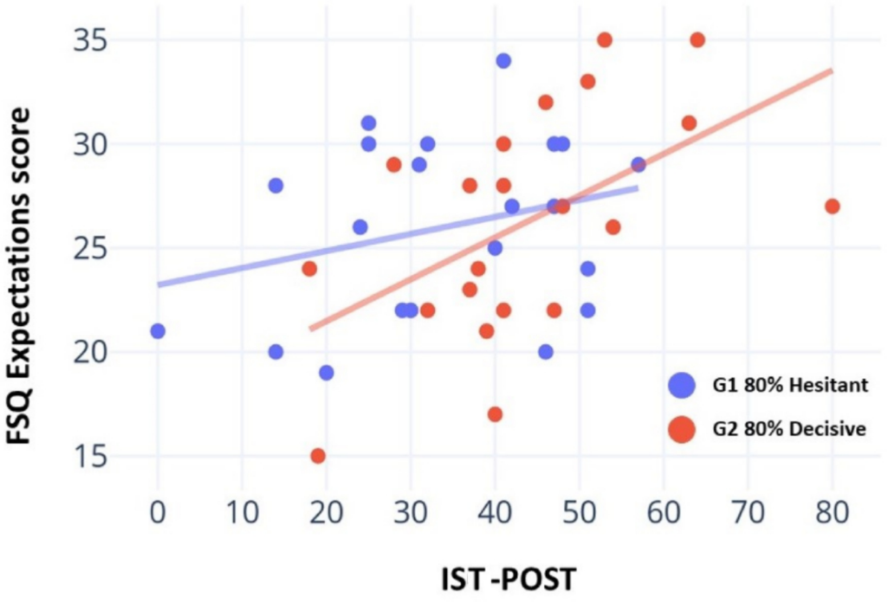


Figure 5.3. Scatter plot and trend lines between FSQ expectations score and IST-POST (Participants that had higher scores on the FSQ subscale tended to score higher on IST-POST. This correlation was significant for Group 2 [r(21)=.534, p =.013] in red but not for the Group 1 [r(21)=.280, p =.219] in blue.)

decisive behaviour and a small percentage of hesitations. These results might seem at first counter-intuitive, as we might expect that less variable behaviour would elicit less mentalistic explanations. However, it is expected that decisive behaviour and some degree of variability (hesitant behaviour) might have been perceived as more human-like than the robot that displayed consistently hesitant actions. Literature has shown that people expect robots to be precise, efficient, competent, and reliable with some degree of social skills (Horstmann and Krämer, 2019). This was also revealed on the scales we applied before the experiment. Thus, participants' expectations fit better with a fast responding and competent robot. Simultaneously, the lower probability of occurrence of the hesitant behaviour, and thus high degree of unexpectancy, might have surprised the observers and made a more human-like impression, translating on higher scores for the group that observed this behaviour. We argue that when hesitant behaviour was "the rule", rather than "the exception", it lost the surprise factor and became a delayed and not efficient behaviour and was not interpreted as the result of mental capacities but rather as a hindrance to performance. Therefore, such a robot might evoke fewer impressions of an intentional agent. In addition, the small amount of hesitant behaviours might have been interpreted as a vulnerability, revealing that the robot was not able to complete the task 100% of the time. Studies have shown that when people perceive the robot as vulnerable, they feel more comfortable during interactions and tend to include the robot more in-group dynamics, relative to robots that do not recognize having difficulties in completing tasks (Strohkorb Sebo et al., 2018). This might have made the robot appear more human-like, and thus described in terms that are more mentalistic. However, this effect might break when vulnerability becomes a rule and impedes the robot to perform a task. Regarding the results from the questionnaires, it is worth mentioning that both groups had very similar profiles before and after the experiment in terms of expectations and attitudes towards robots. However, after observation, the only observable difference between groups was the

IST-POST, which suggests that in order to elicit mentalistic explanations to artificial agents' behaviour, a general positive predisposition towards robots might be required, but it is not sufficient. IST-POST scores were also correlated to Conscientiousness scores on the Big Five Inventory. This subscale evaluates the personality trait linked to being organized, efficient, and diligent. Our results might suggest that an artificial agent that reflects certain personality characteristics might be perceived as more human-like. In our case, additionally to the mentioned factors, the observed behaviour might have resonated with peoples' preferences and thus was judged as a result of internal states. Individual differences might be crucial when adopting the intentional stance. Further studies should address this phenomenon.We observed that the likelihood of using mentalistic explanations to describe iCub's behaviour was influenced principally by the type of behaviour observed. However, we investigated whether this difference could have been attributed to participants' expectations regarding robots' general capabilities, according to the Expectations scores on the FSQ questionnaire. We found that although expectations play a role in the difference of scores, the main effect was associated with the observation of a decisive behaviour. Some studies have pointed out that the discrepancy between expectations and reality might affect human-robot interaction (Chaminade et al., 2012; Edwards et al., 2016; Komatsu et al., 2012; Perez-Osorio et al., 2019; Schramm et al., 2020) [13, 14, 25, 26, 27]. These studies indicate that human-like robots elicit more elaborated expectations regarding competence, cognition, social interaction skills and emotions, relative to more machine-like platforms. Therefore, design and behavioural elements should be aligned with the agent's performance. We speculate that expectations may also play a role in the attribution of mental states to artificial agents during social interaction. If those expectations match observed behaviour, humans might be more likely to use mentalistic explanations. Future studies should deepen the knowledge regarding the link between expectations and the attribution of mental states to artificial agents.

5.6 Conclusion

The present paper addressed the question of relationship between observed behaviour and attribution of mental states to a humanoid robot. We suggest that variations of behaviour reflecting cognitive processes exhibited by a humanoid robot modulate the use of mentalistic explanations. Interestingly, decisive behaviour with a sporadic and unexpected "hesitant" behaviour was related to mentalistic attributions. Although expectations regarding the performance of the robot play a role in the mental attribution, the kind of behaviour exhibited has more considerable weight on the adoption of the intentional stance. Our results also suggest that individual differences might be taken into account on the interpretation of human-robot interaction findings.

Chapter 6

Publication 4: I Am Looking for Your Mind: Pupil Dilation Predicts Individual Differences in Sensitivity to Hints of Human-Likeness in Robot Behaviour

This chapter presents the published manuscript from the following publication:

Marchesi, S., Bossi, f., Ghiglino, D., De Tommaso, D. and Wykowska A., (2021). "I Am Looking for Your Mind: Pupil Dilation Predicts Individual Differences in Sensitivity to Hints of Human-Likeness in Robot Behaviour," vol. 8, no. June, pp. 1–10, 2021, doi: 10.3389/$frobt$.2021.653537. Materials and datasets are available at https://osf.io/s7tfe/

Author Contributions: SM and AW designed the pupillometry task. DG and AW designed the observational task. DD programmed the behaviours of the robot. SM and DG performed data collection. SM and FB analyzed the data. SM and AW wrote the manuscript. All authors contributed to reviewing the manuscript and approved it.

6.1 Abstract

The presence of artificial agents in our everyday lives is continuously increasing. Hence, the question of how human social cognition mechanisms are activated in interactions with artificial agents, such as humanoid robots, is frequently being asked. One interesting question is whether humans perceive humanoid robots as mere artefacts (interpreting their behaviour with reference to their function, thereby adopting the design stance) or as intentional agents (interpreting their behaviour with reference to mental states, thereby adopting the intentional stance). Due to their humanlike appearance, humanoid robots might be capable of evoking the intentional stance. On the other hand, the knowledge that humanoid robots are only artefacts should call for adopting the design stance. Thus, observing a humanoid robot might evoke a cognitive conflict between the natural tendency of adopting the intentional stance and the knowledge about the actual nature of robots, which should elicit the design stance. In the

present study, we investigated the cognitive conflict hypothesis by measuring participants' pupil dilation during the completion of the InStance Test. Prior to each pupillary recording, participants were instructed to observe the humanoid robot iCub behaving in two different ways (either machine-like or humanlike behaviour). Results showed that pupil dilation and response time patterns were predictive of individual biases in the adoption of the intentional or design stance in the IST. These results may suggest individual differences in mental effort and cognitive flexibility in reading and interpreting the behaviour of an artificial agent.

6.2 Introduction

Artificial agents are becoming increasingly present in our daily environment. From vocal assistants to humanoid robots, we are observing a change in the role played by these new entities in our lives (Samani et al., 2013). However, it is still a matter of debate as to whether humans perceive embodied artificial agents, such as humanoid robots, as social and intentional agents or simple artefacts (Hortensius and Cross, 2018; Wykowska et al., 2016). Several researchers have investigated whether humans would deploy similar sociocognitive mechanisms when presented with a novel type of (artificial) interaction partner (i.e., humanoid robots) as they would activate in an interaction with another human (Cross et al., 2019; Saygin et al., 2012; Wykowska, 2020).

In this article, we report a study in which we investigated whether robot behaviour—by being humanlike or mechanistic—can modulate the likelihood of people adopting the intentional stance (Dennett, 1971). The study also addressed the question of whether pupil dilation—a marker of cognitive effort—can predict the type of stance people would adopt toward the robots, and how all these factors are related to individual "mentalistically inclined" or "mechanistically inclined" biases.

According to Dennett (1971), the intentional stance is a strategy that humans spontaneously adopt to interpret and predict the behaviour of other humans, referring to the underpinning mental states (i.e., desires, intentions, and beliefs). The intentional stance is an efficient and flexible strategy, as it allows individuals to promptly interpret and predict others' behaviour. However, when interacting with non-biological systems, humans might adopt a different strategy, which Dennett describes as the design stance. According to the author, we deploy this strategy when explaining a system's behaviour based on the way it is designed to function. The intuition behind Dennett's definition is that humans would adopt the stance that allows them to predict and interpret the behaviour of a system in the most efficient way. Thus, the adoption of either stance is not predefined; on the contrary, if the adopted stance is revealed as inefficient, one can switch to the other stance.

Several authors have demonstrated that people tend to spontaneously adopt the intentional stance toward other human and nonhuman agents (Abu-Akel et al., 2020; Happé and Frith, 1995; Heider and Simmel, 1944; Zwickel, 2009; see also Perez-Osorio and Wykowska, 2019 and Schellen and Wykowska, 2019 for a review). However, it is not yet entirely clear which of the two aforementioned stances humans would adopt when interacting with humanoid robots. On the one hand, humanoid robots present humanlike characteristics, such as physical appearance (Fink, 2012). Hence, it is possible that these characteristics elicit representations and heuristics similar to those that we rely on when interacting with humans (Airenti, 2018; Dacey, 2017; Waytz, Cacioppo et al., 2010; Złotowski et al., 2014). This might trigger the neural representations related to the adoption of the intentional stance (Chaminade et al., 2012; H. L. Gallagher et al., 2002; Özdem et al., 2017; Spunt et al., 2015). Indeed, the presence of humanlike characteristics is one of the key factors that, according to Epley et al. (2007), contribute to anthropomorphism toward artificial agents, facilitating the adoption of the intentional stance. On the other hand, humanoid robots are man-made artefacts, and therefore, they might evoke the adoption of the design stance, as they can be perceived simply as machines (Wiese et al., 2017).

Recent literature has addressed the issue of adopting the intentional stance toward robots. For example, Thellman et al. (2017) presented a series of images and explicitly asked their participants to rate the perceived intentionality of the depicted agent (either a human or a humanoid robotic agent). The authors reported that participants perceived similar levels of intentionality behind the behaviour of the human and the robot agents. Marchesi et al. (2019) investigated the attribution of intentionality to humanoid robots, developing a novel tool, the InStance Test (IST). The IST consists of a series of pictorial "scenarios" that depict the humanoid robot iCub (Metta et al., 2010) involved in several activities. In Marchesi et al. (2019), participants were asked to choose between mentalistic and mechanistic descriptions of the scenarios. Interestingly, individuals differed with respect to the likelihood of choosing one or the other explanation. Such individual bias in adopting one or the other stance toward humanoid robots called for examining whether it is possible to identify its physiological correlates. In fact, Bossi et al. (2020) examined whether it is possible to relate individual participants' EEG activity in the resting state with the individual likelihood of adopting the intentional or design stance in the IST. The authors found that resting-state beta activity differentiated people with respect to the likelihood of adopting either the intentional or the design stance toward the humanoid robot iCub. Recently, Marchesi et al. (Marchesi, Spatola et al., 2021) have identified a dissociation between participants' response time and the stance adopted toward either a human or a humanoid robot. Moreover, the individual bias emerged as being linked to participants' individual tendency to anthropomorphize nonhuman agents.

Since the literature presents evidence for various individual tendencies to adopt either the design or the intentional stance, in the present study, we aimed at using pupil dilation as a marker of individual bias and cognitive effort invested in the task of describing a robot's behaviour, by adopting either stance. In addition, we were interested in finding out whether observing different types of robot behaviour (humanlike or mechanistic) would have an impact on adopting the two different stances, taking into account individual biases.

6.2.1 Pupillometry as an Index of Cognitive Activity

We focused on pupil dilation, as pupillary response is a reliable psychophysiological measure of changes in cognitive activity (for a review, see Larsen and Waters, 2018; Mathôt, 2018). Literature reports show that the pupils dilate in response to various cognitive activities. Previous studies have investigated the mechanisms underpinning pupil dilation, such as emotional and cognitive arousal (how much activation a stimulus can elicit) and cognitive load (the mental effort put into a task) (Larsen and Waters, 2018; Mathôt, 2018). de Gee et al.(De Gee et al., 2014) reported that, in a visual detection task, pupil dilation was greater for participants with a tendency to stick to their decisional strategy (defined as "conservative participants") who made a decision not in line with their individual bias in the task. This result shows that pupil dilation can be considered as a marker of conflict between participants' individual bias and the decision they take. Moreover, it has been shown that the variation in pupil size is linked to the activity in the locus coeruleus (Jackson and Sirois, 2009) and to the noradrenergic modulation (Larsen and Waters, 2018), and thus, greater pupil size can be considered as an indicator of general arousal and allocation of attentional resources. Other studies have used pupil dilation as an indicator of cognitive load and mental effort. For example, Hess and Polt (Hess and Polt, 1964) reported that pupil dilation is closely correlated with problem-solving processes: the more difficult the problem, the greater the pupil size. Moreover, the recent literature (Pasquali et al., 2020; Pasquali et al., 2021) assessed the use of pupillometry in real and ecological scenarios where participants interacted with the iCub robot. The authors show that pupillometry can be a reliable measure to investigate cognitive load in the context of human–robot interaction. Overall, these studies provide evidence that pupillometry is an adequate method to study individual tendencies and how they are related to resources allocated to a cognitively demanding task (for a comprehensive review, see also Mathôt, 2018). Here, we consider pupil dilation as a measure of cognitive effort related to the activation of one or the other stance in the context of one's individual biases.

6.2.2 Aims of the Study

The aims of the present study were to 1) examine whether observing an embodied humanoid robot exhibiting two different behaviours (a humanlike behaviour and a machine-like behaviour) would modulate participants' individual bias in adopting the intentional or the design stance (assessed with the IST) and 2) explore whether this modulation would be reflected in participants' pupil dilation, which is considered as a measure of cognitive effort. More specifically, we explored whether observing a humanoid robot behaving either congruently or incongruently with respect to participants' individual tendency to adopt the intentional stance would lead them to experience different levels of cognitive effort in the InStance Test. That is because we expected participants to experience an increase in cognitive effort due to the dissonance between their individual tendency in interpreting the behaviour of a humanoid robot and the need for integrating the representation of the observed behaviour manifested by the embodied robot.

6.3 Materials and Methods

6.3.1 Participants

Forty-two participants were recruited from a mailing list for this experiment (mean age: 24.05, SD: 3.73, females: 24) in return for a payment of 15€. All participants self-reported normal or corrected-to-normal vision. The study was approved by the local Ethical Committee (Comitato Etico Regione Liguria) and was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Each participant provided written informed consent before taking part in the experiment. All participants were naïve to the purpose of this experiment and were debriefed upon completion. Five participants were excluded from data analysis, due to technical problems occurring during the recording phase. Three participants were excluded due to insufficient amount of valid pupil data (<60%). A total of 34 participants were included in the data analysis.

6.3.2 Pupil-Recording Apparatus, Materials, and Procedure

In a within-subject design, participants first attended, in a dimly lit room, the robot observation session, where they were positioned in front of the embodied iCub and observed it exhibiting a humanlike or a machine-like behaviour. Right after this session, the participants were led to a different room (dimly lit) where they were instructed to sit down and position their head on a chinrest. They were then presented with the IST. The procedure would then be repeated for the second behaviour of the robot. Choosing a within-participants design, and exposing participants to both behaviours of the robot, allows for a higher control of their previous knowledge and experience related to the iCub robot.

Items from the IST were presented on a 22- LCD screen (resolution: $1,680 \times 1,050$). A chinrest was mounted at the edge of the table, at a horizontal distance of 62 cm from the screen. The monocular (left eye) pupil signal was recorded using a screen-mounted SMI RED500 eyetracker (sampling rate of 500 Hz). The dim illumination of the room was kept constant through the whole duration of the experimental sessions. The IST items were displayed through Opensesame 3.2.8 (Mathôt et al., 2012).

Robot Behaviour

Before taking part in the IST, the participants were asked to observe the embodied iCub robot, which was programmed to behave as if it was playing a solitaire card game on a laptop positioned in front of it. From time to time, the robot was turning its head toward a second monitor, located on its left side, in the periphery. On this lateral monitor, a sequence of videos was played for the entire duration of this session. The behaviours displayed by the robot, in terms of eye and head movements, were manipulated between two experimental conditions. One condition involved the robot displaying a humanlike behaviour, which was a replica of the behaviour recorded in a previous attentional capture experiment from a human participant (detailed description of the robot behaviours is beyond the scope of this article; for details, see Ghiglino et al., 2018). It is important to point out that the behaviour displayed by the robot in this condition fully embodied the variability and the unpredictability of the behaviour displayed by the human when the recording was first made. As a contrast condition, we programmed the robot to display another behaviour, which was extremely stereotypical and predictable, defined as "machine-like" behaviour. While the "humanlike" behaviour consisted of several patterns of neck and eye movements, the "machine-like" behaviour consisted of just one pattern of neck and eye movements. In other words, the "machine-like" behaviour was generated in order to display no variability at all. The order of presentation of these two behaviours was counterbalanced across participants.

InStance Test Stimuli and Task

After the observation session, the participants performed a 9-point calibration, and they were then presented with the IST (Bossi et al., 2020; Marchesi et al., 2019; Figure 7.1). The instructions in each trial were as follows: (i) first, look freely at the baseline image ($1,000$ ms), (ii) freely explore the presented item ($5,000$ ms), (iii) listen to the two sentences ($5,000$ ms Sentence A and $5,000$ ms Sentence B), and finally, (iv) choose the description that you think better explains the presented scenario by moving a cursor on a slider (until click) (Figure 7.2). The presentation order of mechanistic and mentalistic sentences was counterbalanced. Presentation of items was randomized. The IST was split into two subsets 17 of items, with one half (one subset, 17 items) presented after one observation session and the other half (17 items) after the second observation session (the order of presentation of the subsets was counterbalanced). An example of the mentalistic sentences is "iCub pretends to be gardener"; an example of a mechanistic sentence is "iCub adjusts the force to the weight of the object" (Figure 7.2).

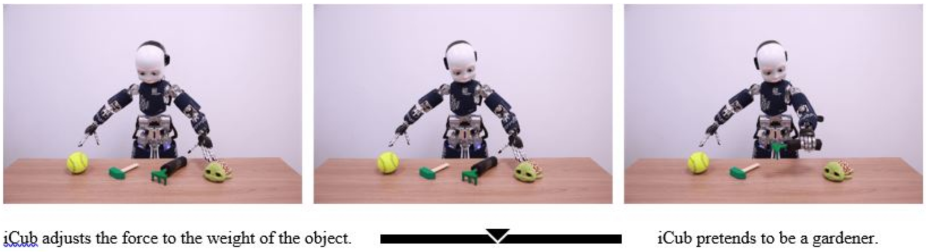

iCub adjusts the force to the weight of the object.
iCub pretends to be a gardener.

Figure 6.1. Exemplification of the IST items with exemplification of Sentence A and Sentence B (Marchesi et al., 2019).

The complete list of mechanistic and mentalistic sentences, associated with the corresponding scenarios, is reported in Marchesi et al. (2019) Appendix A. To avoid eye movements related to the reading process, for each scenario, the two descriptions were presented auditorily through headphones (similarly to the procedure adapted for EEG, Bossi et al., 2020). Moreover, to allow a reliable baseline correction, we created a luminance-related baseline version of each scenario using MATLAB function Randblock (https://it.mathworks.com/matlabcentral/fileexchange/17981-randblock).

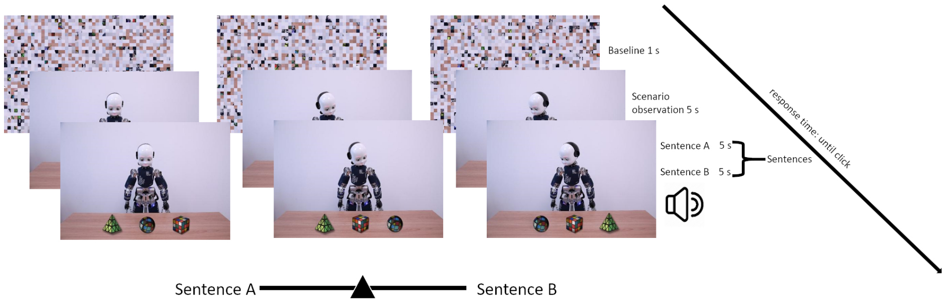


Figure 6.2. Experimental time line.

This function allowed us to create a scrambled version of each item scenario with randomized blocks of pixel positions. The scrambled items were used as specific baselines for each corresponding scenario. This process was necessary to control the different luminance levels of each item.

6.3.3 Pupil Data Preprocessing

All data were preprocessed (and analyzed) using R (version 3.4.0, available at http://www.rproject.org) and an open-source MATLAB (The Mathworks, Natick, MA, United States) toolbox provided by Kret and Sjak-Shie (Kret and Sjak-Shie, 2019). To clean and preprocess the data, we followed the pipeline proposed by Kret & Sjak-Shie: 1) first, we converted the eyetracker data to the standard format used by Kret & Sjak-Shie's MATLAB toolbox. Since we were interested in exploring how pupil dilation could predict participants' choice in the IST, we decided to take the duration of each sentence as our time window of interest. Thus, data were segmented and pre-processed separately for the selected time windows. By applying this procedure, we reduced the probability that the pupil dilation signal would be biased by the preprocessing procedure (Mathôt, 2018; Procházka et al., 2010). In this dataset, we included information relevant to the pupil diameter, start/end time stamps of each segment, and validity of the data point, in separate columns. 2) We filtered dilation speed outliers, trend-deviation outliers, and samples that were temporally isolated, applying the parameters described by Kret and Sjak-Shie (2019). In greater detail, in order to mitigate possible gaps due to non-uniform sampling, dilation speed data were normalized following the formula below:

$$d'^{[i]} = \max\left(\frac{|d\,[i] - d\,[i-1]|}{|t\,[i] - t\,[i-1]|}, \frac{|d\,[i+1] - d\,[i]|}{|t\,[i+1] - t\,[i]|}\right), \tag{6.1}$$

where d' $^{[i]}$ indicates the dilation speed at each sample, d[i] indicates the pupil size series, and t[i] indicates the corresponding time stamp. Dilation speed outliers were then identified using the median absolute deviation (MAD, Leys et al., 2013). MAD is a robust metric of dispersion, resilient to outliers. Samples within 50 ms of gaps were rejected; contiguous missing data sections larger than 75 ms were identified as gaps. The MAD metric was applied to identify absolute trend-line outliers. 3) We interpolated and smoothened the signal using a zero-phase low-pass filter with a cutoff of 4Hz (Jackson and Sirois, 2009). After having applied the pipeline described above, data were baseline-corrected by subtracting the mean pupil size during the baseline phase from the mean pupil size in our time of interest (ToI), and dividing by the mean pupil size during the baseline (Preuschoff et al., 2011).

$$\frac{M_{\text{pupil size in ToI}} - M_{\text{baseline pupil size}}}{M_{\text{baseline pupil size}}}, \tag{6.2}$$

This process allows a clean comparison of the resulting percentage of pupillary change relative to the baseline.

6.3.4 Sample Split and Dichotomization of the IST Response

In line with Bossi et al. (2020), in order to investigate individual biases, participants were grouped by their average individual InStance Score (ISS, the overall score across both robot behaviour conditions): mentalistically biased people (>0.5 SD over the mean score, N = 12, average ISS for this group: 62.25, SD: 7.64) and mechanistically biased people (<−0.5 SD below the mean score, N = 9, average ISS for this group: 28.23, SD: 5.66). People who were not clearly over or under the cutoff value (−0.5 < score < 0.5 SD, N = 13, average ISS for this group: 44.90, SD: 4) were considered as the "unbiased" group. Moreover, to be able to investigate participants' stance in the IST (mentalistic vs. mechanistic), we considered the type of selected sentence (by considering as mechanistic a score <50 and mentalistic a score >50) as the attributed explanation to the item (from here on, defined as "Attribution"), leading to a binomial distribution. Although this practice could lead to a considerable loss of information, it allowed for a higher control of the interindividual variability present in the raw IST scores that could bias the overall mean score.

6.3.5 Data Analysis: Pipeline Applied for (Generalized) Linear Mixed-Effects Models

Data analysis was conducted on the mean pupil size (baseline-corrected) for the time windows of interest (Sentence A and Sentence B time periods) using linear (or generalized linear where needed) mixed-effects models (Bates et al., 2015). When it comes to linear mixed-effects models (LMMs) or generalized linear mixed-effects models (GLMMs), it is important to specify the pipeline that was followed to create the models. (i) First, we included all the fixed effects that allowed the model to converge. (ii) We included random effects that presented a low correlation value ($|r| < 0.80$) with other random effects, to avoid overfitting. In all our models, Participant was included as a random effect. (iii) The significance level of the effects for the LMM was estimated using the Satterthwaite approximation for degrees of freedom, while for the GLMM, we performed a comparison with the corresponding null model (likelihood ratio tests, LRTs). Since time series analyses were not planned, autocorrelation of factors was not modeled. Detailed parameters for each model are reported in the Appendix D.

6.4 Results

In line with Marchesi et al. (2019), the score in the InStance Test was calculated ranging on a scale from 0 (extreme mechanistic value) to 100 (extreme mentalistic value). In order to obtain the average InStance Score (ISS) per participant, the scores across single scenarios were averaged. Before performing any preprocessing, the overall average score at the InStance Test after observing the mechanistic behaviour was 43.80, with SD: 17.69, and the overall average score after observing the humanlike behaviour was 43.44, with SD: 18.03 [t(65.97) = –0.08, p = 0.934]; thus, the type of robot behaviour that participants observed did not modulate the ISS. The overall sample average score at the InStance Test was 43.62, SD: 17.26.

As in the study by Bossi et al. (2020), given that our focus was the individual bias at the IST, in the present section, we will report the results from the mechanistically and mentalistically biased participants, leading to an overall total sample of N = 21 participants. Results on the very same models involving unbiased participants as well are reported in the Appendix D (overall N = 34 participants).

6.4.1 InStance Test Individual Attribution and Pupil Size

The first model (GLMM) aimed at investigating the relationship between pupil size and participants' attribution at the IST. Our fixed effects were as follows: 1) the mean pupil size, 2) robot behaviour previously observed, and 3) participants' general bias at the IST, while we considered the selected attribution as the dependent variable. Because of this, the distribution of the GLMM is binomial.

The main effect of RobotBehaviour emerged as statistically significant (b = -0.537, model comparison: χ^2 (1) = 24.286, p = <0.001). Results showed that participants chose more often an attribution congruent with the behaviour previously observed on the robot (more mechanistic attribution after watching machine-like behaviour and vice versa) (Figure 7.3).

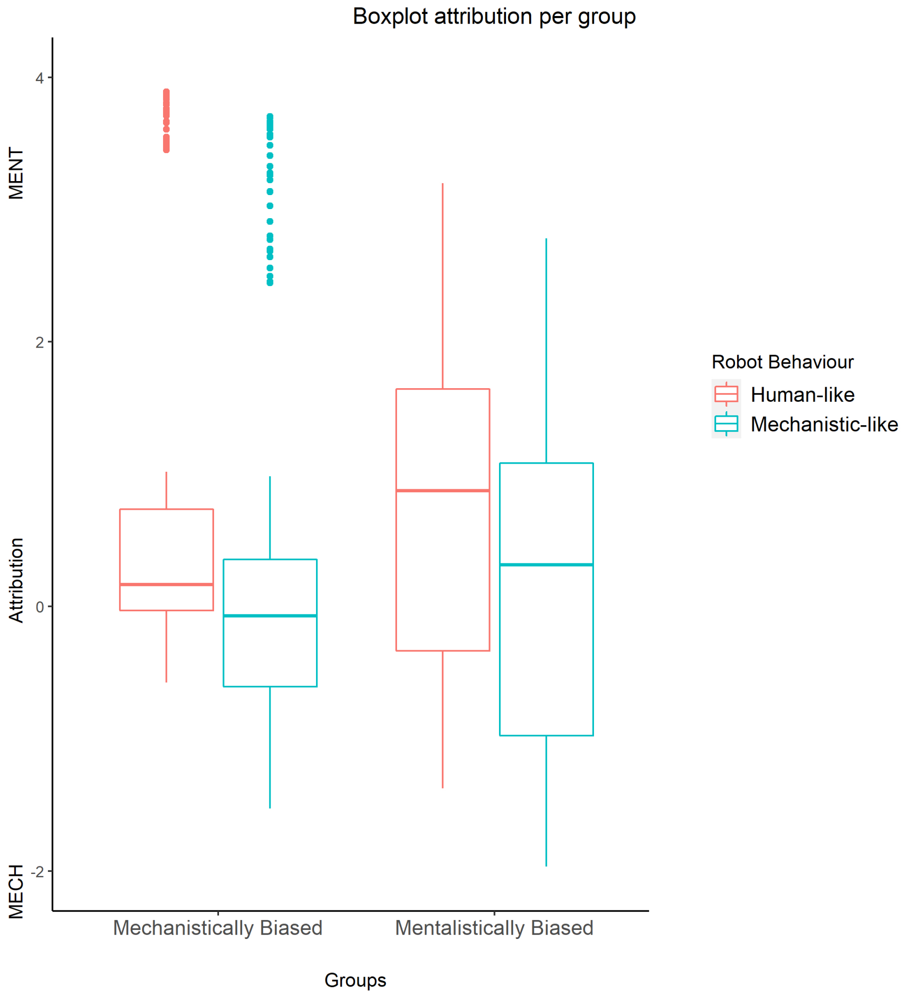


Figure 6.3. GLMM: boxplot showing the statistically significant effect of RobotBehaviour * Bias on attribution, with extreme values as predicted by the model.

The interaction effect between RobotBehaviour * mean pupil size was statistically significant as well (b = -9.291, model comparison: χ^2 (1) = 9.355, p = 0.002). Although the three-way interaction between RobotBehaviour*mean pupil size * individual bias was significant only when taking into account the Unbiased group (see Appendix D), our main a priori hypotheses aimed at exploring differences due to participants' individual bias in the IST. Therefore, we performed a planned comparison GLMM for each bias group (Ruxton and Beauchamp, 2008; Tucker, 1990; Kuehne, 1993;) to test the interaction between RobotBehaviour * mean pupil size: mechanistic group (model comparison: $\chi^2(1)$ = 7.701 p = 0.005); mentalistic group (model comparison: χ (1) = 3.001, p = 0.083). These results show that mechanistically biased participants showed a greater pupil dilation for attributions congruent with the robot behaviour (b = -9.28, z = -2.757, p = 0.005, Figure 7.4) when attributing a mechanistic description after the observation of the robot behaving in a machine-like way and when attributing a mentalistic score after the observation of the robot behaving in a humanlike way. On the other hand, mentalistically biased participants showed a tendency, although statistically not significant, toward greater pupil sizes for mentalistic attributions, relative to mechanistic attributions, regardless of the robot behaviour (b = -4.45, z = -1.73, p = 0.083, Figure 7.4).

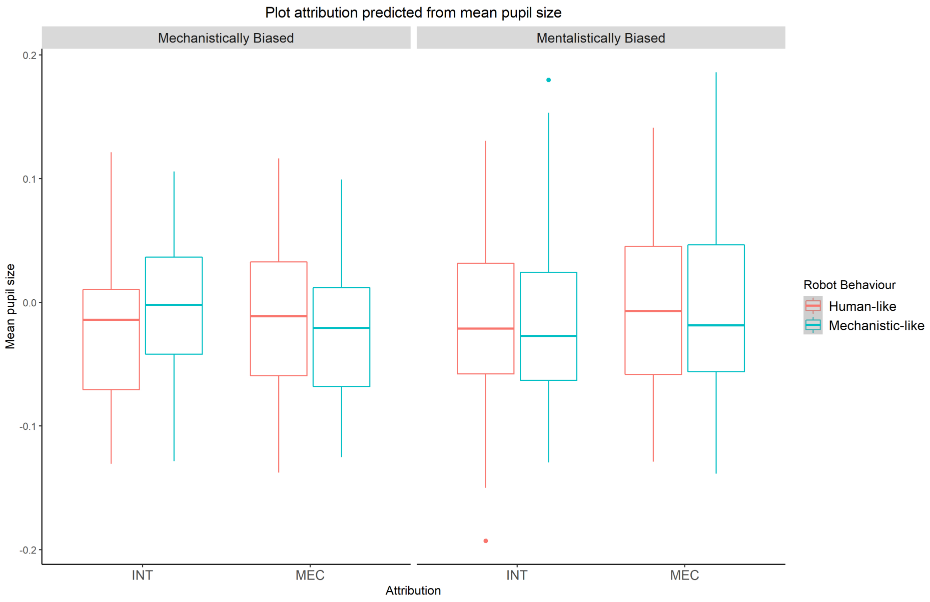


Figure 6.4. GLMM on the mechanistic group (N = 9) and the mentalistic group (N = 12). The mechanistic bias group shows the interaction effect between attribution and mean pupil size. No statistically significant effect on attribution and pupil size in the mentalistic bias group.

6.4.2 Behavioural Data Analysis

In order to investigate the relationship between behavioural data and participants' response times, we tested the quadratic effect of the z-transformed IST score (included as the fixed factor) on log-transformed response times (our dependent variable), as we expected them to be smaller in the extremes of the score distribution of the IST. Results showed a statistically significant quadratic effect of the IST score [b = -0.146, t $(1, 419.99)$ = -9.737, p = <0.001] (Figure 7.5). These results show that participants were overall faster when scoring on the extremes of the IST scale.

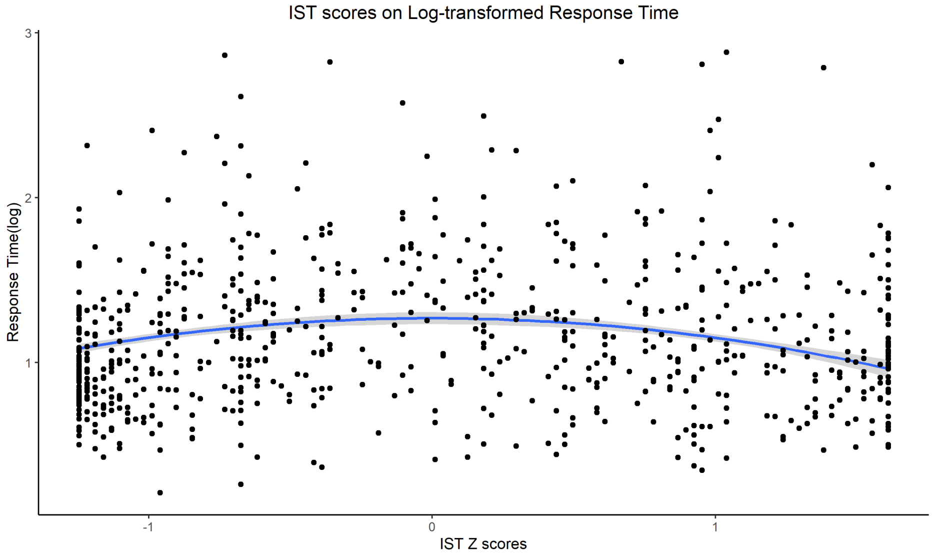


Figure 6.5. LMM: statistically significant quadratic effect of the IST-z score on log-transformed response time showing faster RTs for extreme scores.

6.5 Discussion

In the present study, we investigated whether adopting the intentional/design stance could be predicted by changes in pupil dilation and how both effects are modulated by participants' individual bias in adopting the intentional stance and by a behaviour of a robot observed prior to the test. To address these aims, we conducted an experiment in which participants first observed the embodied humanoid robot iCub, programmed to behave as if it was playing solitaire on a laptop positioned in front of it. From time to time, the robot was programmed to turn its head toward a second monitor on its left periphery, where a sequence of videos was being played. The behaviours exhibited by the robot were manipulated in a within-subjects design: in one condition, the robot exhibited a humanlike behaviour, and in the second condition, the robot exhibited a machine-like behaviour. After each session with the robot, participants' pupil data were recorded while they completed the InStance Test. Participants were then divided into two groups, based on the bias showed by their IST score: a mentalistically biased group and a mechanistically biased group.

We found that both mechanistically and mentalistically biased participants leaned more toward mentalistic attributions in the IST after observing the robot's humanlike behaviour, as compared to the mechanistic behaviour. This shows that participants had some sensitivity to the subtle differences in the robot behaviour, thereby attributing more "humanness" to the humanlike behaviour, independently of their initial bias (Ghiglino, De Tommaso et al., 2020).

We also explored the relationship between the individual bias and the changes in pupil dilation as a function of the behaviours displayed by the robot. We found that the two groups showed different patterns. On the one hand, for mechanistically biased people, pupil dilation was greater when they chose descriptions of the robot behaviour in terms that were "congruent" with the previously observed robot behaviour: a mentalistic attribution after the humanlike behaviour and a mechanistic attribution after the machine-like behaviour. We argue that this is due to the engagement of additional cognitive resources, caused by the cognitive effort in integrating the representation of the observed behaviour into the judgment (Kool et al., 2010; Kool and Botvinick, 2014). In other words, these participants might have had enough sensitivity to detect the "human-likeness" or "machine-likeness" in the behaviour of the robot. We argue that the integration of this piece of evidence into the judgment in the IST might have required additional cognitive resources.

On the other hand, mentalistically biased participants showed a tendency for greater pupil dilation when choosing the mentalistic description, independent of the observed robot behaviour. Perhaps this group of participants showed engagement of additional cognitive resources when they were choosing descriptions that were in line with their initial bias (Christie and Schrater, 2015). Adherence to the "mentalistic" descriptions, independent of observed behaviour, indicates, on the one hand, lower cognitive flexibility than the mechanistically oriented participants and, on the other hand, might be related to the general individual characteristic to structure and make the external world reasonable. This tendency to structure the external environment and engage in cognitive effortful tasks is defined as "need for cognition" (Cacioppo and Petty, 1982; Cohen et al., 1955; Epley et al., 2007). Mentalistically biased participants might have a lower need for cognition, and therefore pay less attention to all the subtle behavioural cues exhibited by the agent and stick to their original bias. Therefore, we may argue that this group is less prone to changing the stance adopted to interpret an agent's behaviour.

One last (and interesting) finding of our study was that RTs were faster on the extremes of the IST score distribution. This suggests that perhaps once participants made a clear decision toward mentalistic or mechanistic description, it was easier and more straightforward for them to indicate the extreme poles of the slider. On the other hand, when they were not convinced about which alternative to choose, they indicated this through keeping the cursor close to the middle and longer (more hesitant) responses.

Overall, it seems plausible that the general mechanistic bias leads to allocating a higher amount of attentional resources toward observation of the robot (Ghiglino, Willemse, De Tommaso et al., 2020), resulting in paying more attention to the details of the observed behaviour (in line also with Ghiglino, De Tommaso et al., 2020; Marchesi et al., 2020). This, in turn, might influence the subsequent evaluation of robot behaviour descriptions. On the other hand, a mentalistic bias might lead participants to stick to their spontaneous first impression (Spatola, 2019) and a lower need for cognition (Cacioppo and Petty, 1982; Cohen et al., 1955; Epley et al., 2007). Commonly, individual differences and expectations shape the first impression about a humanoid robot (Bossi et al., 2020; Horstmann and Krämer, 2019; Marchesi, Spatola et al., 2021; Ray et al., 2008). Perez-Osorio et al. (Perez-Osorio et al., 2019) showed that people with higher expectations about robots tend to explain the robot behaviour with reference to mental states. This might indicate that our participants with a mentalistic bias were predominantly influenced by their expectations about the abilities of the robot and, therefore, paid less attention to the mechanistic behaviours of the robot. To conclude, we interpret the results in light of the influence of individual differences in the allocation of cognitive resources that might differ between people who are prone to adopting the intentional stance toward humanoid robots and people who, by default, adopt the design stance (Bossi et al., 2020; Marchesi, Spatola et al., 2021).

6.6 Limitations of the Current Study and Future Work

In the present study, we opted for a within-subjects design to reduce the influence of interindividual differences related to prior knowledge/experience with the iCub robot. Nevertheless, we cannot rule out the fact that our approach was indeed too conservative, leading to a null effect of the robot behaviour manipulation on the raw IST scores due to a carry-over effect. Future research should consider adapting similar paradigms to a between-subjects design, since this option will allow for controlling possible carry-over effects.

6.7 Concluding Remarks

In conclusion, our present findings indicate that there might be individual differences with respect to people's sensitivity to subtle hints regarding human-likeness of the robot and the likelihood of integrating the representation of the observed behaviour into the judgment about the robot's intentionality. Whether these individual differences are the result of personal traits, attitudes specific to robots, or a particular state at a given moment of measurement remains to be answered in future research. However, it is important to keep such biases in mind (and their interplay with engagement of cognitive resources) when evaluating the quality of human–robot interaction. The evidence for different biases in interpreting the behaviour of a humanoid robot might translate into the design of socially attuned humanoid robots capable of understanding the needs of the users, targeting their biases to facilitate the integration of artificial agents into our social environment.

Chapter 7

Publication 5: Belief in sharing the same phenomenological experience increases the likelihood of adopting the intentional stance towards a humanoid robot

This chapter presents the submitted version of the following manuscript:

Marchesi, S., De Tommaso, D., Pérez-Osorio, J. and Wykowska, A., (in press). Belief in sharing the same, phenomenological experience increases the likelihood of adopting the intentional stance towards a humanoid robot. Materials and datasets are available at https://osf.io/xnm5c/

Author Contributions SM, JPO, and AW designed the experiments. SM and DD created the robot behaviours. DD programmed and implemented the behaviours on the robot. SM performed data collection and run the analyses, SM and AW discussed the data. SM and AW wrote the manuscript. All authors contributed to reviewing the manuscript and approved it.

7.1 Abstract

Humans interpret and predict others' behaviours by ascribing them intentions or beliefs, or in other words, by adopting the intentional stance. Since artificial agents are increasingly populating our daily environments, the question arises whether (and under which conditions) humans would apply the "human-model" to understand the behaviours of these new social agents. Thus, in a series of three experiments we tested whether embedding humans in a social interaction with a humanoid robot either displaying a human-like or machine-like behaviour, would modulate their initial bias towards adopting the intentional stance. Results showed that indeed, humans are more prone to adopt the intentional stance after having interacted with a more socially available and human-like robot, while no modulation of the adoption of the intentional stance emerged towards a mechanistic robot. We conclude that short experiences with humanoid robots, presumably inducing a "like-me" impression and social bonding, increase the likelihood of adopting the intentional stance.

7.2 Introduction

Being intrinsically social, humans need to develop the ability to interpret and understand the behaviours of others occupying the same environment (Baron-Cohen et al., 1999). Meltzoff suggests (Meltzoff, 2007) that the way we learn to understand others is through learning about ourselves and subsequently perceiving (and explaining) others "like me". The author proposes that understanding the similarities between the self and the other is the foundation of social cognition. This basic knowledge and ability provide toddlers (and, later in life, adults) a framework to interpret others' behaviours. The most efficient strategy to predict and interpret humans' behaviour (others' and one's own) is to refer to underlying inner mental states, such as desires, intentions and beliefs (Fletcher et al., 1995; Frith and Frith, 2012; Gallotti and Frith, 2013). Interestingly, referring to others' mental states to explain behaviour might not be limited to only humans. Evidence showed that attribution of mental states to others occurs also with respect to non-human entities (Apperly and Butterfill, 2009; Butterfill and Apperly, 2013; Happé and Frith, 1995; Heider and Simmel, 1944). Given that artificial agents, such as humanoid robots, are increasingly populating our daily lives in various contexts (Samani et al., 2013), it remains to be answered whether we deploy similar socio-cognitive mechanisms to interpret their behaviour as we do towards other humans (Hortensius and Cross, 2018; Wykowska, 2021; Wykowska et al., 2016. When it comes to unfamiliar agents, Wiese and colleagues suggest that we might interpret their behaviours as if they were intentional because this is the default way of making sense of the social world (Wiese et al., 2017). Therefore, when facing novel agents, we apply the schema and knowledge we are most familiar with: the "human model" (Perez-Osorio and Wykowska, 2020; Wiese et al., 2017). This reasoning is in line with the "like me" account of Melzoff, and recent literature shows that this account can be applied to human-robot interaction (Riddoch and Cross, 2021). In this context, empirical studies have investigated whether humans would indeed interpret the behaviour of artificial agents by ascribing to them mental states (Abu-Akel et al., 2020; H. L. Gallagher et al., 2002; Marchesi et al., 2019; for a review, see Perez-Osorio and Wykowska, 2020). In other words, literature investigated whether humans would adopt the intentional stance (Dennett, 1971, 1983; Dennett, 1989) towards artificial agents.

7.2.1 The intentional stance framework

Dennett's theoretical framework accounts for different strategies that humans might adopt when facing the need to interpret another entity's behaviour. These strategies, or "stances", explain and predict the behaviour with reference to different levels of abstraction: 1 – with reference to the physical domain of the agent, such as the reaction of a molecule when heated or the trajectory of a ball (physical stance); 2 – with reference to how the system was built to function, for example, one expects the car to stop if they push the brake (design stance); 3 – with reference to mental states and beliefs of the agents, i.e., expecting that our friend would enjoy an ice cream (intentional stance). According to Dennett (1989), while the first two levels apply to all systems, the third stance has stricter assumptions on the type of agents for whom the stance works efficiently: ‘Here is how it works: first, you decide to treat the object whose behaviour is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally, you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is, what you predict the agent will do’ (Dennett, 1989, p. 17). Therefore, when adopting the intentional stance, we assume that the behaviour we are predicting is the most rational one that the agent can exert in that context, given their beliefs, desires, and constraints. Dennett highlights that any system can be treated as a rational (and intentional) agent. However, only for truly intentional agents (“true believers”), the intentional stance is the most efficient strategy. For other agents or systems, it makes more sense to switch to a different, more efficient, stance (i.e., the design or the physical stance).

7.2.2 Operationalization of the intentional stance in human-robot interaction

In the context of investigating the adoption of intentional stance towards artificial agents, special interest has been given to humanoid robots, since they represent entities that are somewhat “in-between”: on the one hand, as man-made artefacts, they should elicit the adoption of the design stance; on the other hand, given their shape, physical features and perhaps behaviour, they might evoke the human (intentional) model, as discussed by Wiese and colleagues (Wiese et al., 2017). Thus, humans might anthropomorphize humanoid robots by ascribing typically human characteristics (Airenti, 2018; Epley, Waytz, Akalis et al., 2008; Epley et al., 2007; Złotowski et al., 2014). Recently, several authors empirically investigated the adoption of the intentional stance towards robots (Marchesi et al., 2019; Marchesi, Spatola et al., 2021; Thellman et al., 2017). For instance, Thellman and colleagues (Thellman et al., 2017) exposed participants to images of humans and humanoid robots. Participants’ task was to rate the perceived level of intentionality of the depicted agent. Participants reported similar levels of perceived intentionality between the two agents’ behaviours. Marchesi and colleagues (Marchesi et al., 2019) addressed the challenge of operationalizing a philosophical concept by creating a new tool, the InStance Test (IST), to assess people’s individual tendency to attribute intentionality to a humanoid robot. The IST includes 34 pictorial scenarios (each containing three pictures) depicting the iCub humanoid robot (Metta et al., 2010). Each scenario is associated with two descriptions: one always explains the robot behaviour with reference to a mechanistic vocabulary (mechanistic description), the other always describes the robot behaviour with reference to a mental state (mentalistic description). In other words, one sentence is related to the adoption of the design stance, the other one instantiates the adoption of the intentional stance. In the original study (Marchesi et al., 2019), participants were asked to move a cursor along a slider, towards the description that best represents their interpretation of the observed scenario. Results showed that participants adopted the intentional stance to some extent towards iCub. This suggests that social cognition and the adoption of the intentional stance may be the default and spontaneous way of making sense of others (Abu-Akel et al., 2020; Meyer, 2019; Raichle, 2015; Schilbach et al., 2008; Spreng and Andrews-Hanna, 2015; Waytz, Cacioppo et al., 2010). Moreover, recent studies reported that the spontaneous adoption of intentional stance towards robotic agents might be elicited by the individual tendency to anthropomorphize non-human agents (Marchesi, Spatola et al., 2021; Spatola, Monceau et al., 2020). Interestingly, the results of Marchesi et al.'s study (2019) showed also that individuals differed in their bias in adopting either one or the other stance towards a humanoid robot. Bossi and colleagues (Bossi et al., 2020) later found that it is possible to predict this individual bias in adopting the intentional or the design stance from neural oscillatory patterns during the resting state (i.e., before any task is given to participants). Furthermore, recently Ghiglino and colleagues showed that subtle differences in the robot behaviour might influence the individual tendency to adopt the intentional stance (Ghiglino, De Tommaso et al., 2020) and that including human-like behaviours in the robot can facilitate communication in human-robot interaction in interactive scenarios (Ghiglino et al., 2021).

7.2.3 Aim of study

The present study aimed at examining whether interaction with the humanoid robot iCub in a naturalistic context modulates the tendency to adopt intentional stance towards the robot. More specifically, we addressed the question of whether creating a "like me" context through human-like behaviour and social bonding would increase the likelihood of adopting the intentional stance, while generating a "different-from-me" mechanistic behaviour would have the opposite effect. To this aim, we conducted a series of three experiments: in Experiment 1 participants experienced a social context of watching a movie together with the iCub robot. In line with Meltzoff's account (Meltzoff, 2007), we created a context that should affect adoption of the intentional stance through the "like-me" impression of the robot displaying human-like contingent emotional reactions to the events in the movies. In addition, the context should create social bonding with iCub through the phenomenological experience of sharing a familiar social situation. We hypothesized that this manipulation should have activated the "human" model, leading to the adoption of the intentional stance towards the iCub. We measured whether the experimental manipulation would affect the degree to which intentional stance was adopted by administering half of the items of the IST before the interaction with the robot and the other half, after the interaction. In Experiment 2, we aimed at replicating the results of Experiment 1, and we tested the validity of the IST keeping the social interaction with the robot identical to Experiment 1 and changing the way IST was split into pre-and post-interaction items. In Experiment 3 the social context of watching the video remained the same as in Experiment 1 and 2. However, the "like-me" behaviour was no longer present, as the robot was made to behave in a mechanistic, robotic manner. We hypothesize that this should reduce (or eliminate) the "like-me" impression and social bonding. The robot's behaviours were programmed to display very repetitive and mechanical movements.

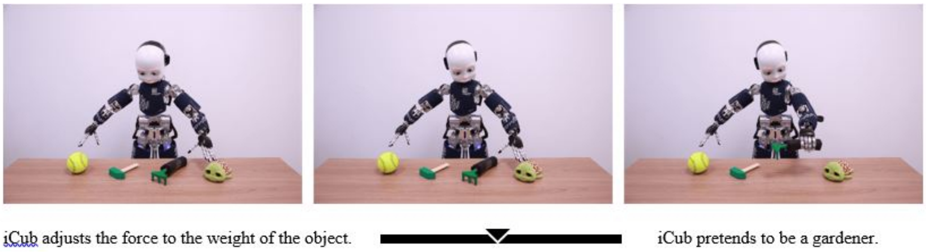
7.3 Robot platform and experimental measures

7.3.1 Robot platform and behaviours

The iCub is a humanoid robotic platform with 53 degrees-of-freedom (DoF) (Metta et al., 2010). Its design allows the investigation of human social cognition mechanisms by generating a context of interaction of high ecological validity. iCub can reliably perform humanlike movements and thereby can be used as a “proxy” of social interaction with another human. In Experiment 1 and 2, we designed three different behaviours of the robot, which were reactions (sadness, awe, and happiness) of the robot to the displayed videos. To implement movements that would be perceived as humanlike as possible, the behaviours followed the principles of animation (Sultana et al., 2013), and were implemented via the middleware YARP (Metta et al., 2006) using the position controller following a minimum jerk profile for head, torso, and arms joints movements. The gaze behaviour was implemented using the 6-DoF iKinGazeCtrl (Roncone et al., 2016) which uses inverse kinematics to produce eye and neck movements. Behaviours were programmed to occur in specific timeframes, corresponding to the apex event of each video. Moreover, to maximize the human-likeness during the verbal interaction at the beginning and at the end of the robot session, the verbal emotional reactions and sentences were pre-recorded from an actor and digitally edited to match the childish appearance of the iCub using Audacity® Cross-Platform Sound Editor. The greetings sentences at the beginning and at the end of the experiment were played by the experimenter via a Wizard-of-Oz manipulation (WoOz) (Kelley, 1983). The WoOz manipulation consists of an experimenter completely (or partially) controlling remotely a robot’s actions during an interaction (movements, speech, gestures, etc) (for a review see Riek, 2012). This method allows researchers to elicit more natural interaction between the robot and the participant, in the absence of AI solutions that would allow the robot to behave in a similar manner autonomously. In addition, since the robot would directly address the participants during the Wizard-of-Oz interaction, cameras from the robot’s eyes were actively recognizing participants’ faces, to create mutual gaze between the iCub and participants. Mutual gaze in human-robot interaction has been shown to be a pivotal mechanism that influences human social cognition (Kompatsiari et al., 2021; Kompatsiari, Ciardo et al., 2018b). Facial expressions on the robot were programmed to display the three different emotions (sadness, awe, and happiness) via the YARP emotion interface module. In Experiment 3, we designed the behaviour of the robot in reaction to the videos in such a way that it would perform always the same repetitive moments of the torso, head, and neck. Cameras were deactivated and, thus, there was no mutual gaze between the robot and participants. The Wizard-of-Oz manipulation was replaced with pre-programmed robotic actions, such as the calibration of joints. The verbal interaction was replaced with a verbal description of the robot’s calibration sequences, created and played via text-to-speech. All the emotional sounds reproduced during Experiments 1 and 2 during the videos were replaced with a “beep” sound. In all three experiments, all sound and recordings were played via two speakers positioned behind the robot, creating the impression that the source of the sound is the robot itself. Videos of the behaviours and verbal scripts are available at https://osf.io/xnm5c/.

7.3.2 Experimental procedure and measures common across all three experiments

The experimental structure consisted of three main parts that were identical across all three experiments: Part 1 – IST pre-interaction: participants would complete the first half of the IST, (Marchesi et al., 2019) to assess their initial individual tendency to adopt the intentional stance towards robots. In Experiment 1, the IST split was conducted in accordance with Marchesi and colleagues (Marchesi, Bossi et al., 2021), by assigning items to Group A or B in a way to obtain two groups with comparable means and standard deviation of the InStance score (based on data from Marchesi et al (2019)). In Experiment 2, the IST split was in accordance with the psychometric structure emerged from the original IST dataset (Marchesi et al., 2019), following the method proposed by Spatola, Marchesi, and Wykowska (under review). Spatola et al. describe a two-factor structure of the IST, one involving mostly the "Alone robot" construct, and the second a "Social" construct where the robot is depicted in the presence of another human. Thus, we performed a factorial analysis on the dataset reported by Marchesi et al. (2019) and split the 34 items of the IST, balancing the emerged factors in the two halves. In all experiments, the presentation of the two groups of items (Group A and B) was counterbalanced across participants between Pre- and Post-interaction with the robot. Regarding the IST task per se (pre-interaction), participants observed scenarios depicting the iCub robot, and they had to drag a slider towards the description of the scenario that they found fitting best to what is displayed in the pictures (Fig. 8.1). After completion of the IST, they would fill out the questionnaire to assess their negative attitudes towards robots (Negative Attitudes towards Robots Scale, NARS, Nomura, 2014; Nomura et al., 2011). Moreover, in Experiment 2 and 3 we assessed also participants' personality phenotype (Big Five Inventory, BFI, Goldberg, 1993).


iCub adjusts the force to the weight of the object.

iCub pretends to be a gardener.

Figure 7.1. Example of item from the IST (Marchesi et al., 2019)

Part 2 - Interaction Session: participants were then instructed to sit beside the robot (1.30 m distance) in a separate room, and they were told that the task would consist of watching three documentary videos with the robot. Each video was edited to last 1.21 minutes, for a total duration of 4.3 minutes. In Experiment 1 and 2, before and after the videos, the robot interacted with the participants via a Wizard-of-Oz manipulation. In more detail, the robot would greet participants, introduce itself, ask participants' names and invite them to watch some videos together. At the end of the videos, the robot would say goodbye to the participants and invite them to proceed to fill out some questionnaires. In Experiment 3, participants were not exposed to any type of social interaction with the robot. The robot only issued verbal utterances about the calibration process it is undergoing (script of the interactions are available under the following link: https://osf.io/xnm5c/) Part 3 – IST post-interaction: after the interaction session with the robot, participants were asked to complete three questionnaires. First, they completed the second half of the IST, to assess whether the robot session modulated their initial tendency to adopt the intentional stance. Subsequently, to assess participants' attitudes towards the robot after the interaction session they completed the Robotic Social Attitudes Scale (RoSAS) (Carpinella et al., 2017), and a set of 7 questions from Waytz and colleagues (Ruijten et al., 2019; Waytz, Cacioppo et al., 2010) to assess their individual tendency to attribute a mind, morality, and reasoning to the robot. In addition, in Experiment 2 and 3, participants completed the Godspeed questionnaire (Bartneck et al., 2009) to assess their level of anthropomorphism of the robot. All participants received monetary compensation of €30. All tests and questionnaires presented in all 3 experiments were administered in Italian and through Psychopy (v2020.1.3) (Peirce et al., 2019) or Opensesame (v3.2.5) (Mathôt et al., 2012). All analyses were conducted with JASP 0.14.0.1(JASP Team, 2020).

7.4 Experiment 1

7.4.1 Participants

Forty participants took part in the study. The study was approved by the local Ethical Committee (Comitato Etico Regione Liguria) and was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Each participant provided written informed consent before taking part in the experiment. All participants were naïve to the purpose of this experiment. Data from 1 participant were excluded from the analyses due to technical problems that occurred during data collection. The final sample was N = 39 (Mage = 25, SDage = 4.75, range = 19 – 42, 28 females).

7.4.2 Analyses

To test whether belief in sharing the same phenomenological experience with a humanoid robot would enhance the adoption of the intentional stance, we first re-coded participants' choices in the IST so that they would range from 0 = totally mechanistic to 100 = totally mentalistic. Subsequently, we conducted a paired sample t-test between the mean score at IST Pre- and Post-interaction with the iCub.

7.4.3 Results

Results showed a significant difference between the mean IST Pre-interaction score and the mean IST Post-interaction score $[t(38) = -3.44, p = .001, C.I.95\% = (-11.51; -2.98)]$ (Table 8.1).

						95% **CI for Mean Difference**			95% **CI for Cohen's d**	
IST_Pre	**IST_Post**	**t**	**p**	**Mean Difference**	**SE Difference**	**Lower**	**Upper**	**Cohen's d**	**Lower**	**Upper**
M_{IST_Pre} 42.12 SD_{IST_Pre} 18.33	M_{IST_Post} 49.37 SD_{IST_Post} 20.12	$-3.44(38)$	$.001$	-7.25	2.10	-11.50	-2.98	-0.55	-0.88	-0.21

Table 7.1. Results from paired sample t-test between IST_Pre and IST_Post interaction - Experiment 1.

After sharing a familiar context with the robot reacting in a human-like emotional manner contingent to the events in the videos, participants chose more often the mentalistic description, leading to an overall mean IST Post-interaction score higher (M IST_Post = 49.37, SD IST_Post = 20.12) than the mean IST_Pre-interaction score (M IST_Pre = 42.12, SD IST_Pre = 18.33). In addition to the t-test, we performed a correlation analysis between the mean IST_Pre-interaction and Post-interaction score and a battery of questionnaires administered before (NARS) and after (RoSAS, and Waytz) the robot experience. Regarding the questionnaire administered before the experience, results revealed no significant correlation between the mean IST score (either Pre or Post) with the NARS. One positive correlation emerged between the mean IST_Post and the Warmth subscale of the RoSAS $[r = 0.38, p = .015, C.I.(0.08; 0.62)]$ and two positive correlations between the Waytz score and the mean IST_Pre $[r = 0.34, p = .032, C.I.(0.03; 0.59)]$ and the Waytz score and the mean IST_Post $[r = 0.50, p = .001, C.I.(0.22; 0.70)]$.

7.4.4 Discussion Experiment 1

The main aim of Experiment 1 was to test whether the likelihood of adopting the intentional stance would be increased by creating a familiar social context that presumably elicits bonding and where the robot induces a "like-me" impression. Results showed that after the social interaction with the robot, participants indeed scored higher in the IST, meaning that they chose more often the mentalistic description of IST items in the post-interaction IST, relative to the pre-interaction IST. In addition, the more mentalistically the robot was described before the interaction, the more it was also described as warm (Warmth subscale of the RoSAS questionnaire). Finally, higher mentalistic attribution pre- and post-social interaction were correlated with higher attribution of mental abilities, morality, and reasoning to the robot (Waytz questionnaire).

7.5 Experiment 2

To test the reliability of the effect observed in Experiment 1, we conducted a follow-up experiment where we kept the robot interaction session identical to Experiment 1, but we changed the way the IST items were split into pre-and post-interaction halves (see Par. 8.3.2 above).

7.5.1 Participants

Forty-one participants took part in the study and received monetary compensation of €30. The study was approved by the local Ethical Committee (Comitato Etico Regione Liguria) and was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Each participant provided written informed consent before taking part in the experiment. All participants were naïve to the purpose of this experiment. Data from 1 participant were excluded from the analyses due to technical problems that occurred during data collection. The final sample was N = 40 (Mage = 29.12, SDage = 8.87, range = 18 – 60, 23 females).

7.5.2 Analyses

To test whether the experience of a shared social context with a humanoid robot, presumably eliciting bonding and "like-me" impression would enhance the adoption of the intentional stance, we first re-coded participants' choice in the IST so that it would range from 0 = totally mechanistic to 100 = totally mentalistic. Subsequently, we conducted a paired sample t-test between the mean score at IST_Pre- and Post-interaction with the iCub. Results showed a significant difference between the mean IST_Pre-interaction score and the mean IST Post-interaction score [t (39) = -5.31, p = $<$ .001, C.I. 95% = (-17.07; -7.65)]. Results confirm findings from Experiment 1. Indeed, participants chose more often the mentalistic description after the interaction with the robot, leading to an overall mean IST_Post-interaction score higher (M IST_Post = 54.08, SD IST_Post = 16.65) than the mean IST Pre-interaction score (M IST_Pre = 41.71, SD IST_Pre = 14.27) (Table 8.2).

						95% **CI for Mean Difference**			95% **CI for Cohen's d**	
IST_Pre	**IST_Post**	**t(df)**	**p**	**Mean Difference**	**SE Difference**	**Lower**	**Upper**	**Cohen's d**	**Lower**	**Upper**
M_{IST_Pre} 41.71 SD_{IST_Pre} 14.27	M_{IST_Post} 54.08 SD_{IST_Post} 16.65	$-5.31(39)$	$< .001$	-12.36	2.33	-17.07	-7.65	-0.84	-1.19	-0.47

Table 7.2. Results from paired sample t-test between IST_Pre and IST_Post interaction - Experiment 2.

In addition to the t-test, we performed a correlation analysis between the mean IST_Pre-interaction and Post-interaction scores and a battery of questionnaires administered before (BFI and NARS) and after the experience with the robot (RoSAS, Godspeed, and Waytz).

Regarding the questionnaires administered before the interaction, results revealed no significant correlation between the mean IST score (either Pre or Post) with the BFI and the NARS. Two significant positive correlations emerged between the mean IST_Post and three subscales of the Godspeed: the Likeability subscale [r = 0.31, p = .045, C.I. (0.008; 0.573)], the Animacy subscale [r = 0.32, p = .04, C.I. (0.17; 0.57)] and the Anthropomorphism subscale [r = 0.41, p = .008, C.I. (0.12; 0.64)]. Moreover, one positive correlation emerged between the mean IST_Pre and the Warmth subscale of the RoSAS [r = 0.41, p = .009, C.I. (0.11; 0.64)] and two positive correlations between the Waytz score and the mean IST_Pre [r = 0.38, p = .015, C.I. (0.08;0.62)] and the Waytz score and the mean IST_Post [r = 0.46, p = .003, C.I. (0.17; 0.67)].

7.5.3 Discussion Experiment 2

The main aim of Experiment 2 was to replicate and confirm results from Experiment 1 which showed an increased likelihood of adopting the intentional stance after an interaction with the iCub robot in a familiar social context, which presumably elicits social bonding and a "like-me" impression. To this aim, we split IST items into pre-and post-interaction halves considering the psychometric structure of the IST (Spatola, Marchesi & Wykowska, under review). Results confirmed findings of Experiment 1 since participants indeed scored higher (more "mentalistically") in the IST after the interaction with the robot, relative to the score before the interaction. Moreover, correlation analysis revealed that the more likely participants adopted the intentional stance after the social interaction, the more the robot was described as likable (Likeability subscale of the Godspeed questionnaire), active, and responsive (Animacy subscale), and anthropomorphic (Anthropomorphism subscale of the Godspeed questionnaire). In addition, the more mentalistically the robot was described before the interaction, the more it was also described as warm (Warmth subscale of the RoSAS questionnaire). Finally, higher mentalistic attribution pre- and post-social interaction were correlated with higher attribution of mental abilities, morality, and reasoning to the robot (Waytz questionnaire). Overall, our results confirmed our hypothesis that creating a familiar social context of sharing an experience and social bonding, together with human-like behaviour that might be interpreted as "like-me" increases adoption of the intentional stance towards a humanoid robot.

7.6 Experiment 3

7.6.1 Aim of Experiment 3

Since Experiment 1 and 2 results showed that it is possible to increase the likelihood of the adoption of the intentional stance through a social context, shared experience, and human-like behaviours (emotionally contingent on the events in the video), we needed to test whether the effect was indeed due to our experimental manipulation or rather due to simple exposure to the robot. To confirm that the observed effect was due to our manipulation, we conducted Experiment 3 in which the robot displayed repetitive and mechanistic behaviours in the same social context. We reasoned that behaviours that are not human-like and not emotionally contingent on the events occurring in the videos should disrupt the social bonding and the “like-me” impression. This in turn should not increase the likelihood of adopting the intentional stance after the interaction.

7.6.2 Participants

Forty-one participants took part in the study and received a monetary compensation of €30. The study was approved by the local Ethical Committee (Comitato Etico Regione Liguria) and was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Each participant provided written informed consent before taking part in the experiment. All participants were naïve to the purpose of this experiment. Data from 1 participant were excluded from analyses due to a poor understanding of the Italian language. The final sample was N = 40 (Mage = 34.27, SDage = 12.29, range = 18 – 54, 30 females).

7.6.3 Experimental procedure and measures

The experimental procedure and the battery of questionnaires pre- and post-interaction were the same as in Experiment 1, except for the behaviours of the robot during the Robot Session (see Par 8.3.2 above). The IST was split into pre-and post-interaction items in the same way as in Experiment 2.

7.6.4 Analyses

To test our hypothesis, we conducted a paired sample t-test between the mean score at IST Pre- and Post- interaction with the iCub. Results showed no significant difference between the mean IST_pre-interaction (M IST_pre =43.40, SD IST_pre = 14.61) score and the mean IST_post-interaction score (M IST_post = 44.97, SD IST_post = 16.30) $[t(39) = -0.57, p = .569, C.I.95\% = (-7.08; 3.95)]$, cf. Table 8.3. In addition, we performed the same correlation analyses between the mean IST_pre and post-interaction scores with the same battery of questionnaires as in Experiment 1 and Experiment 2. No correlation between the mean IST_pre or post-interaction and the BFI or the NARS emerged as significant. One positive correlation between the mean IST_post score and the Intelligence subscale of the Godspeed was significant [r = 0.47, p = .002, C.I. (0.19; 0.68)]. Two positive correlations emerged between the mean IST_post score and the Competence [r = 0.31, p = .049, C.I. (0.002; 0.569)] and the Warmth [r = 0.50, p = $<$ .001, C.I. (0.23; 0.70)] subscales of the RoSAS. As in Experiment 1 and 2, the Waytz questionnaire scores positively correlated with both the mean IST_pre [r = 0.42, p = .007, C.I. (0.12; 0.64)] and mean IST_post [r = 0.67, p = $<$ .001, C.I. (0.46; 0.81)] scores.

						95% **CI for Mean Difference**			95% **CI for Cohen's d**	
IST_Pre	**IST_Post**	**t(df)**	**p**	**Mean Difference**	**SE Difference**	**Lower**	**Upper**	**Cohen's d**	**Lower**	**Upper**
M_{IST_Pre} 43.40 SD_{IST_Pre} 14.61	M_{IST_Post} 44.97 SD_{IST_Post} 16.30	$-0.57(39)$	0.569	-1.56	2.72	-7.08	3.95	-0.091	-0.40	-0.22

Table 7.3. Results from paired sample t-test between IST-Pre and IST-Post interaction - Experiment 3.

7.6.5 Discussion Experiment 3

The main aim of Experiment 3 was to test whether the effects observed in Experiment 1 and Experiment 2 were indeed due to our manipulation, rather than the mere exposure to the robot. To address this aim, we exposed participants to an interaction with a robot displaying repetitive and pre-programmed behaviours, not emotionally contingent on the events of the videos. Results of Experiment 3 showed that our participants did not increase their initial tendency of adopting the intentional stance after the interaction with the mechanistically behaving robot. This suggests that the effects of Experiment 1 and Experiment 2 were indeed due to our intended manipulation rather than mere exposure to the robot. Thus, we can conclude that creating a familiar context of shared experience with a robot that creates an impression of being "like-me" increases the likelihood of adopting the intentional stance. Mere exposure to the robot is not sufficient, as results of Experiment 3 have shown. In addition to the effect of our primary interest, results of Experiment 3 showed, analogously to Experiment 1, that the more participants were choosing the mentalistic option in the IST, the more they described the robot as warm (Warmth subscale of the RoSAS questionnaire). Moreover, we found the same pattern emerging in describing the robot as competent (Competence subscale of the RoSAS questionnaire). Confirming previous findings, the more participants were adopting the intentional stance toward the iCub pre-and post-interaction, the more they attributed a mind, reasoning, and morality to it (Waytz questionnaire).

7.7 Comparison between experiments

To confirm the impact of human-like robot behaviours on the likelihood of adopting the intentional stance, and to control for age and gender, we decided to compare the results between experiments. Specifically, we conducted an analysis comparing Experiment 2 and Experiment 3 where the IST-pre- and post-interaction were administered in the same way (same way of splitting IST items into pre-and post-interaction sets) while the interaction itself differed with respect to human-likeness of robot behaviours. We first calculated the Δ-IST score as the difference between the IST-post and IST-pre for each participant. Subsequently, we performed an ANCOVA considering the Δ-IST score as our dependent variable, Experiment as a fixed factor, and age and gender as covariates. The Δ-IST score allows us to compare the magnitude of the modulation of the adoption of the intentional stance related to robot exposure. No main effect of gender or age emerged as significant. Furthermore, confirming previous results, the main effect of Experiment emerged as significant [F(1, 1696.86) = 6.64, p = 0.012, η^2 = 0.07]. Post-hoc comparisons with Bonferroni correction revealed a significant difference in the Δ-IST score between Experiment 2 and Experiment 3 [t = 2.57, p = 0.012, C.I. (2.18; 17.05)]

7.8 General discussion

The present study aimed at examining whether people might increase their likelihood of adopting the intentional stance towards a humanoid robot in a familiar context of a shared experience of watching a movie together, in which the robot displays human-like behaviours, emotionally contingent on the events in the video, and thus presumably creating a "like-me" impression and social bonding. To address this aim, we invited participants to watch three videos alongside the iCub robot. During the video-watching session, the robot would either exhibit an emotional and human-like reaction contingent to the narration of the videos (Experiment 1 and Experiment 2) or a very repetitive and machine-like behaviour, emotionally not contingent (no emotional reactions at all) on the events of the videos (Experiment 3). Moreover, before the video session, the robot would either greet and verbally interact with participants (Experiment 1 and 2) in a human-like manner (through a Wizard-of-Oz technique) or display a mechanistic calibration behaviour (Experiment 3). Our results showed that the behaviours displayed by the robot in Experiment 1 and 2 led participants to score higher (i.e., choose more mentalistic descriptions of robot behaviour) in a test probing the degree of adoption of the intentional stance (the intentional stance test (IST), Marchesi et al., 2019) after the interaction, relative to their scores before the interaction. Therefore, we can conclude that the short experience of sharing a social context with the robot presumably led participants to perceive the humanoid robot as "like-them", increasing the likelihood of adopting the intentional stance towards the robot. Conversely, short, repetitive, and machine-like behaviours (Experiment 3) did not affect the initial degree of adopting the intentional stance towards the robot, confirming that the differential effect observed in Experiment 1 was due to the experimental manipulation and not to mere exposure to the robot. Our results are in line with recent literature on the adoption of the intentional stance and mind attribution towards robot behaviours (Abubshait et al., 2021; Abubshait and Wykowska, 2020; Ciardo et al., 2021; Marchesi et al., 2020). Specifically, Ciardo and colleagues (in press) report that, when a robot behaviour is perceived as more mechanistic in a joint task, participants decrease their likelihood of adopting the intentional stance towards it. Along similar lines, Marchesi et al. (Marchesi et al., 2020) tested whether observing a robot exerting certain behaviours more frequently and other behaviours rarely would modulate the adoption of the intentional stance. The authors found that infrequent, unexpected behaviours increased the likelihood of adopting the intentional stance. Finally, Abubshait et al. (2021) found a pattern of results in a very similar direction as those presented here. In their experiment, participants performed a joint task with the iCub robot. In one condition, they believed they scored jointly with iCub while in another condition, they scored individually. The results showed that the “social framing” of the task, namely the belief that participants score as a team with iCub, increased the likelihood of adopting the intentional stance towards iCub. Hence, similarly as in the present study, the social “bonding” with the robot seemed to increase the likelihood of adopting the intentional stance. In addition to our main effects of interest, the results of this study showed that individual attitudes towards robots correlated with mind attribution towards the robot, in line with previous literature (Ghiglino, De Tommaso et al., 2020; Horstmann and Krämer, 2019; Perez-Osorio et al., 2019; Spatola, Kühnlenz et al., 2020). Taken together, we argue that our results support the idea that people might be more likely to adopt the intentional stance toward artificial agents when the agents create the impression of being “like-me” and when the context generates social bonding and shared experience. This is in line with such phenomena as shared intentionality (Dewey et al., 2014; Gilbert, 2009; Pacherie, 2014) and other effects occurring during shared social contexts (Boothby et al., 2014).

7.9 Concluding remarks

Our study demonstrates that adoption of the intentional stance is influenced by the phenomenology of shared experience with the robot, which is presumably induced by the behaviours displayed by the robot and the context of interaction. This interplay between the context and the robot behaviours should be examined further in various contexts of human-robot interaction.

Part IV

The influence of culture on the attribution of intentions to robot

Chapter 8

Publication 6: The mediating role of anthropomorphism in the adoption of the intentional stance towards humanoid robots

This chapter presents the submitted version of the following submitted manuscript:

Marchesi, S., Spatola, N., and Wykowska, A., (submitted, currently under revision). The mediating role of anthropomorphism in the adoption of the intentional stance towards humanoid robots. Materials and datasets are available at https://osf.io/3mzpj/

SM and AW designed the task. SM performed data collection, SM and NS performed data analysis, SM, NS, and AW discussed the data. SM, NS and AW wrote and edited the manuscript. All authors contributed to reviewing the manuscript and approved it.

8.1 Abstract

Evidence from cognitive psychology showed that cultural differences influence human social cognition, leading to a difference activation of social cognitive mechanisms. A growing corpus of literature in Human-Robot Interaction is investigating how culture shapes cognitive processes like anthropomorphism or mind attribution when humans face artificial agents, such as robots. The present paper aims at disentangling the relationship between cultural values, anthropomorphism and intentionality attribution to robots, in the context of the intentional stance theory. We administered a battery of tests to 600 participants from various nations worldwide and modelled our data with a path model. Results showed a consistent direct influence of collectivism on anthropomorphism, but not on the adoption of the intentional stance. Therefore, we further explored this result with a mediation analysis that revealed anthropomorphism as a true mediator between collectivism and the adoption of the intentional stance. We conclude that our finding extend previous literature by showing that the adoption of the intentional stance towards humanoid robots depends on anthropomorphic attribution in the context of cultural values.

8.2 Introduction

There is an increasing trend to introduce robots in human' environments in daily contexts such as school, elder care (for a review see Wykowska, 2020). As humans become exposed to robot presence, it is crucial to examine how robots are represented in the human mind, and how robot actions are perceived and interpreted. (for a review, see Wykowska, 2021). Previous studies have shown that people differ in the degree of anthropomorphism (Chin et al., 2004; Cullen et al., 2013) and in the likelihood of attributing intentionality to robot actions (Bossi et al., 2020; Hegel et al., 2008; Marchesi et al., 2019; Thellman and Ziemke, 2020). Apart from inter-individual differences such as personality traits (Airenti, 2018; A. D. Kaplan et al., 2019; Spatola and Wykowska, 2021; Waytz, Cacioppo et al., 2010), or level of education (Ghiglino and Wykowska, 2020), also cultural values might be a factor influencing the degree to which robots are perceived as human-like. In the present study, we investigated how the cultural values of individualism and collectivism may influence the anthropomorphic representation of robots (i.e., the attribution of human-characteristics to a non-human system) and consequently, the likelihood of adopting the intentional stance towards robots (Dennett, 1989). The cultural values such as individualism or collectivism are defined as the core norms and principles shared by a community to organize social life (Triandis, 1993)

8.2.1 The importance of cultural values

Cultural values influence our general attitudes towards others and social exclusion/inclusion behaviours. For example, Varnum and colleagues (Varnum et al., 2010) report that people from Eastern countries adopt a more holistic approach and are more prone to interdependency (i.e., the tendency to value more the relational harmony rather than the individual (Kitayama et al., 2010)) compared to people from Western countries. Indeed, several authors reported differences in collectivistic and individualistic values between Eastern and Western countries, leading to the conclusion that collectivism and individualism are "Cultural Syndromes" (for a review see Triandis, 1993). According to Triandis (1993), a Cultural Syndrome is defined when constructs (such as collectivism/individualism) are: 1 – organized around a theme, a value; 2 – the differences within-culture are smaller than across cultures and 3 - these constructs are clustered in line with geographic regions (not necessary countries). Therefore, individualism and collectivism can coexist in the same country, but they are applied and emphasized differently upon the situation. Cultural Syndromes are internalized by individuals, thus each person has an individual tendency to adhere to collectivism or individualism and this individual tendency influences our choices in everyday life, especially when we socially interact with other humans and shape the way we present ourselves to others (Bandura, 2002; Markus and Kitayama, 1991; Vygotsky, 1980)

8.2.2 Anthropomorphism

A second pivotal concept in our investigation is anthropomorphism in the definition given by

Epley and colleagues (2007). The authors describe anthropomorphism as the attribution of human physical or mental characteristics to non-human agents (of any kind and form) and can be elicited by three factors: 1- the availability of physical characteristics of the agent that can activate knowledge and heuristics related to humans; 2- the possibility of fulfilling the human need for connection and sociality; 3- individual traits (Epley et al., 2007). While the first two factors depend on the agent and how much it activates the "human-model" (Wiese et al., 2017), the third rely on each individual and researchers proved that different personality phenotypes have different tendencies to anthropomorphize (Spatola and Wykowska, 2021).

8.2.3 The intentional stance

The third concept core to the present work is the concept of intentional stance. Dennett (1989, 2009) defines the intentional stance as the strategy humans spontaneously adopt when interpreting the behaviours of other agents with reference to mental states (she desires, he wishes, etc.). Interestingly, according to Dennett, this process takes place towards all agents that fulfil the epistemological requirement of being treated as rational. When this latter requirement is not met, humans can switch to the adoption of a different stance that allows a better explanation of the agent behaviour, such as the design stance, defined as interpreting the agent's behaviour with reference to how it is designed to behave. Several authors showed that indeed, the adoption of intentional stance can take place towards biological and non-biological agents (Happé and Frith, 1995; Heider and Simmel, 1944; Zwickel, 2009). Therefore, we can conclude that attributing mental states to non-human agents can be considered part of a broader construct such as anthropomorphism (Airenti, 2018; Dacey, 2017; Waytz, Epley et al., 2010; Złotowski et al., 2014).

8.2.4 Cultural differences, anthropomorphism and the adoption of intentional stance in Human-Robot Interaction

Recent literature in HRI investigated the influences of cultural differences when humans face artificial agents (for a review, see (Lim et al., 2020). The cultural background seems to influence HRI regarding several features of social interaction: likeability, trust, and engagement (Rau et al., 2010), proxemics (Remland et al., 1991), recognition of facial expressions (R. E. Jack et al., 2009), and preferred style of communication (Papadopoulos and Koulouglioti, 2018). Findings from social psychology show a higher in-group bias for cultures with high collectivist values (Brewer and Chen, 2007; Chen et al., 2002; Yamagishi et al., 1998). Kuchenbrandt and colleagues reported that in-group bias influence people's evaluation of robots (Kuchenbrandt et al., 2013). That is because robots are not members of the human (in-) group, and thus, collectivist cultures could be less prone to anthropomorphize them. In addition, Kovačić found that individualistic cultures present a more positive attitude toward technology adoption (compared to collectivist cultures). Kovačić argue that individuals perceive technology as potentially useful tool to help them perform better (Kovačić, 2005). As positive attitudes tend to result in a more positive evaluation, the consequence would be the attribution of positive (human-like) characteristics such as warmth or sociability (Dupree and Fiske, 2017). Nevertheless, little is known about how the cultural background and specifically, the individual tendency towards collectivism/individualism, shapes adoption of intentional stance towards a humanoid robot via the level of evoked anthropomorphism. In this context, one could argue that the individual differences in collectivism should influence how individuals interpret and explain the behaviour of a humanoid robot (i.e., adopting either the intentional or the design stance). That is because non-biological agents such as a social humanoid robot can fulfil the first two factors described by Epley and colleagues (2007) (1- the availability of physical characteristics of the agent that can activate knowledge and heuristics related to humans; 2- the possibility of fulfilling the human need for connection and sociality) and, therefore, be anthropomorphized. In this case, once the "human model" representation (Wiese et al., 2017) is activated, it could trigger the adoption of the intentional stance (Chaminade et al., 2012; H. L. Gallagher et al., 2002; Özdem et al., 2017; Spunt et al., 2015). A higher anthropomorphic attribution could lead to perceive the robot as closer to the human in-group and to a higher likelihood of adopting the intentional stance towards it (Eyssel et al., 2012).

8.2.5 Assessment of the intentional stance in Human-Robot Interaction and the relationship with anthropomorphism

Recently, researchers have investigated whether the intentional stance can be adopted towards humanoid robots (Abu-Akel et al., 2020; Marchesi et al., 2019; Thellman et al., 2017 see also Perez-Osorio and Wykowska, 2020 for a review). Marchesi et al. (2019) created the InStance Test (IST), a novel tool to assess the adoption of the intentional stance. The IST consists of a set of scenarios, depicting the iCub humanoid robot (Metta et al., 2010; Natale et al., 2019) involved in several activities (Fig.1). Below each scenario, two sentences are displayed as possible descriptions of the scenario. One description addresses the scenario with reference to a mental state (i.e., intentional stance) and the other with reference to a mechanistic explanation (i.e., design stance). Marchesi and colleagues asked participants to choose which sentence fits best as a description of the scenario by moving a button along a slider. The authors report that indeed, there are individual differences in the adoption of the intentional stance, which called for a deeper investigation of the relation between the individual likelihood to adopt the intentional stance towards a robot and the individual tendency to anthropomorphize non-human systems. Indeed, Marchesi et al. (Marchesi, Bossi et al., 2021) report that the individual differences in anthropomorphism are associated with different levels of individual tendency to adopt the intentional stance towards robots (a higher tendency to anthropomorphize is associated with a higher tendency to adopt the intentional stance to explain the behaviour of a humanoid robot)

8.2.6 Aim of the present research

State-of-the art research suggests that humans do adopt the intentional stance towards humanoid robots under some circumstances. Moreover, individual traits and other factors might differentiate individuals regarding their likelihood of adopting the intentional stance. Other biases and heuristics such as our tendency to anthropomorphize and cultural differences such as collectivistic or individualistic values might influence the individual differences. Thus, differences in the level of individualism/collectivism can lead to a higher/lower tendency to anthropomorphize artificial agents. Hence, the present paper aims to disentangle the relationship between culture, the various dimensions of anthropomorphism, and the adoption of the intentional stance towards humanoid robots. More in detail, we hypothetise that individual differences in cultural values should influence the adoption of the intentional stance via the anthropomorphic attributions towards the robot.

To address the aims of our study, we asked participants to complete a battery of questionnaires to model the relationship between the above-mentioned constructs: Cultural Values Scale (CVSCALE, Hofstede, 2011; Yoo et al., 2011) to measure individual adherence to cultural values, the Human-Robot Interaction scale (HRIES, (Spatola, Kühnlenz et al., 2020) to measure anthropomorphism, and the InStance Test (IST, Marchesi et al., 2019) to measure the degree of adopting the intentional stance. Our hypotheses were two-folded: first, individual differences in the collectivism subscale of the CVSCALE should influence anthropomorphism (Epley, Akalis et al., 2008; Epley, Waytz, Akalis et al., 2008; Eyssel et al., 2012, and second, building on results of Marchesi et al. (2021), anthropomorphism should positively mediate the causal relationship between the individual tendency towards collectivism and the adoption of the intentional stance. That is, the higher the individual level in the collectivism subscale, the higher the anthropomorphic attribution towards the robot and thus, the higher the adoption of the intentional stance towards the same robot.

8.3 Methods and measures

8.3.1 Participants

Six hundred-and-one participants from Singapore/Malaysia; United-States of America (USA), United-Kingdom (UK); Germany, Spain and Italy were recruited from Prolific.co. We selected these countries to access different level of individualism-collectivism values and therefore to reach a representative sample (see table 9.1 for demographics of each country) in return of a payment of £3.3. The study was approved by the local Ethical Committee (Comitato Etico Regione Liguria) and was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Each participant provided written informed consent before taking part in the experiment by clicking on the "accept" button at the beginning of the survey. All participants were naïve to the purpose of this experiment. (Inclusion criteria are reported in Appendix E).

Table 8.1. Demographic description of the sample

	Sample Demographics					
Country	n	Males	Females	Non-Binary	Prefer Not to Answer	Mage
USA	100	56	42	2		36.5
UK	100	31	70			38.4
Italy	100	54	42	4		25.8
Spain	100	71	27	1	1	27.7
Germany	100	68	29	2	1	31.0
Singapore/Malaysia	101	32	68	1		27.2

Table 8.2. Descriptive Statistics Collectivism

	CV_COLL					
	Ger	ITA	Singa_Malay	Spain	UK	USA
Valid	95	96	95	94	98	94
Mean	4.248	4.115	4.394	4.723	4.296	4.314
Median	4.330	4.167	4.500	4.670	4.330	4.330
Std. Deviation	1.006	1.064	1.024	0.934	1.120	1.281
Minimum	1.330	1.167	1.500	2.670	1.170	1.000
Maximum	6.670	6.500	7.000	6.830	6.670	6.670

8.3.2 Task procedure and measures

Participants completed a series of questionnaires tailored to assess different variables that, according to our hypothesis, may influence the adoption of the intentional stance toward a humanoid robot. The order of presentation was as follows: Cultural Values Scale, InStance Test, and the Human-Robot Interaction Scale. All questionnaires were administered in English. To guarantee that the respondents would have sufficient knowledge of the English language to understand the task, and the items of the questionnaires, we administered the LexTALE test (Lemhöfer and Broersma, 2012), which provides a measure of the individual level of English proficiency.

Cultural values measure.

Hofstede (2011) developed a framework to describe cultural values with a five-dimensions model that encompasses power distance, collectivism, uncertainty avoidance, masculinity vs. femininity, and long-term orientation vs. short-term orientation. Power distance refers to how many individuals accept and expect that power is distributed unevenly. Collectivism refers to the extent to which people rely on themselves (individualism) or their group (collectivism). Uncertainty avoidance refers to how much tolerance individuals have for ambiguity. Masculinity vs. femininity refers to how the inequalities between genders are perceived in society and how individuals internalize these inequalities. Finally, long-term orientation refers to the temporal orientation of a society, meaning to what extent individuals feel pressured to plan their life with a long-term mindset. For the purposes of our study, we adopted the collectivism subscale of the Cultural Values Scale (CVSCALE, Yoo et al., 2011), which has proven to be reliable in measuring Hofstede's dimensions of collectivism/individualism at the individual level. The five dimensions are power distance (6 items, e.g., people in higher positions should make most decisions without consulting people in lower positions; ω = .816, CI95% [.793, .839]), uncertainty avoidance (5 items, e.g., It is important to closely follow instructions and procedures; ω = .862, CI95% [.845, .880]), collectivism (6 items, e.g., Individuals should sacrifice their self-interest for the group; ω = .860, CI95% [.843, .877]), long-term orientation (6 items, e.g., Long-term planning is important; ω = .749, CI95% [.719, .780]), and masculinity (4 items, e.g., It is more important for men to have a professional career than it is for women; ω = .831, CI95% [.808, .853]). In addition, as Hofstede posited a dimension named "indulgency", which addressed the extent to which a society values leisure time, moral discipline, happiness and well-being (Enkh-Amgalan, 2016). Khans and Cox reported that indulgent cultures tend to create new technology as a way to improve life, to value the expression of emotions and freedom of speech. To explore the indulgence dimension, we developed 5 items (e.g., Freedom of speech is important; ω = .715, CI95% [.679, .751]) (Khan and Cox, 2017; Prim et al., 2017). For each item, participants had to indicate the extent to which they agree or disagree with the statement, from 1 "Disagree strongly" to 7 "Agree strongly".

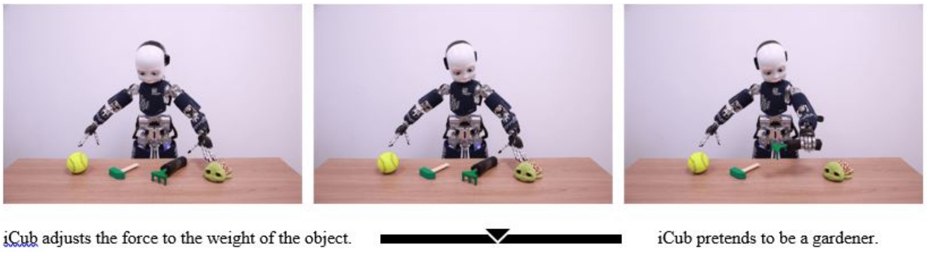
Anthropomorphism measure.

Spatola and colleagues (2020), created the Human-Robot Interaction Evaluation Scale (HRIES) to assess the individual tendency to anthropomorphize social robot. The authors report four subscales:

Sociability (i.e., to what extent the robot is perceived as social, ω = .840, C.I.95%[.811, .865]), Agency (i.e., to what extent the robot it perceived to have agency in its environment, ω = .767, C.I.95%[.727, .803]), Animacy (i.e., to what extent the robot is perceived as having smooth movements, ω = .807, C.I.95%[.778, .833]), and Disturbing (i.e., to what extent the robot is perceived as uncanny, ω = .883, C.I.95%[.863, .900]). Participants were asked to evaluate how closely a series of adjectives would describe the robot (i.e., "Warm", "Alive", "Real"). The items were evaluated on a 7-point Likert scale (1: "Not at all", 7: "Totally").

Intentional stance measure.

The InStance Test was developed by Marchesi et al. (2019) to assess the individual likelihood of adopting of the intentional stance towards a humanoid robot. Overall, it consists of 34 scenarios (ω = .880, C.I.95% [.867, .894]), each one made of three static images depicting the iCub robot (Metta et al., 2010). Below, two sentences (one mentalistic, thus representing the adoption of the intentional stance and the other mechanistic, thus representing the adoption of the design stance) are presented as possible description of each scenario (Fig. 9.1). Participants' task is to choose the sentence they consider a good fit for the depicted scenario by moving a button on a slider. The obtained score ranges from 0 (totally mechanistic, i.e., design stance) to 100 (totally mentalistic, i.e., intentional stance). For the complete list of items, refer to Marchesi et al., 2019 Appendix A.



iCub adjusts the force to the weight of the object.

iCub pretends to be a gardener.

Figure 8.1. Exemplification of the IST items with exemplification of Sentence A and Sentence B (Marchesi et al., 2019).

8.4 Analyses and results

8.4.1 Filtering procedure

Participants with missing values and that resulted over or below 2.5 SD on the variable of interests were considered as outliers, leading to a total sample of N = 572 participants (Singapore/Malaysia n = 95; USA n = 94; UK n = 98; Germany n = 95; Spain n = 94; Italy n = 96). All the analyses were performed with JASP v.0.14.1 (2020), using the SEM package.

8.4.2 Comparison across Countries on IST and HRIES subscales

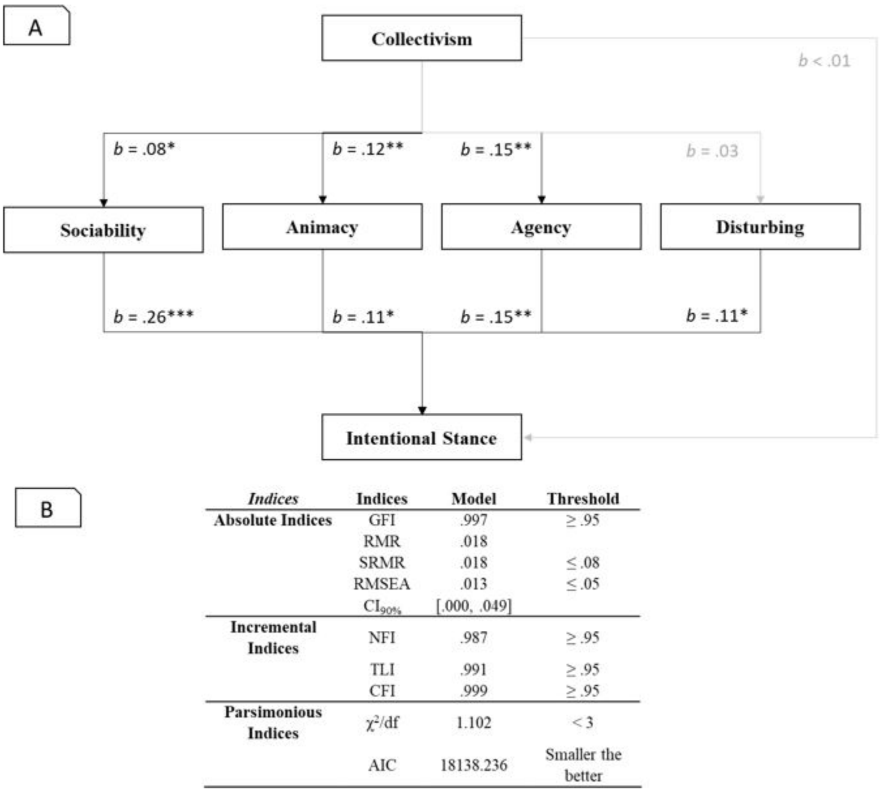
To check for differences in the adoption of the intenitonal stance and in the tendency to anthropomorphize across Countries, we performed non-parametrical Kruskal-Wallis test on the z-scored corrected IST and HRIES subscales scores. No significant effect emerged from the analysis (all p > .06).

8.4.3 Structural Equation Model

First, we explored the structural relationship between our variables, we conducted a path model analysis. We defined our model introducing all the subscales from the CVSCALE (to control for covariation) as first ordered variables, all the subscales from the HRIES as second ordered variables and the IST as outcome variable. We controlled for covariance between same level variables. In addition, to account for possible differences in the knowledge of the English language, we added the LexTALE test (Lemhöfer and Broersma, 2012) as a covariate (see Appendix E for model parameters). All 95% Confidence Interval were calculated with a bias corrected percentile bootstrap analysis with 10,000 samples (Zhao et al., 2010), with Maximum Likelihood estimator, that has been proven to provide more precise model fit indexes and less biased parameters, compared to other estimators such as Generalized Least Squares and Weighted Least Squares (Olsson et al., 2000). All variables were corrected using z-scores. The model fit indices and Information Criteria are reported respectively in Figure 9.2.

8.4.4 Structural Equation Model Results

Figure 9.2 represents all the significant path emerged from the path model, revealing a consistent direct influence of the Collectivism on the considered anthropomorphism constructs (specifically on the Sociability, Agency and Animacy subscales) but not on the IST. Given our initial hypothesis, we further explored this relationship (collectivism →anthropomorphism →intentional stance) with a mediation model, considering the above-mentioned variables respectively as predictor, mediator, and outcome. All paths' parameters are reported in the Appendix E.



Indices	Indices	Model	Threshold
Absolute Indices	GFI	.997	≥ .95
	RMR	.018	
	SRMR	.018	≤ .08
	RMSEA	.013	≤ .05
	$CI_{90\%}$	[.000, .049]	
Incremental Indices	NFI	.987	≥ .95
	TLI	.991	≥ .95
	CFI	.999	≥ .95
Parsimonious Indices	χ^2/df	1.102	< 3
	AIC	18138.236	Smaller the better

Figure 8.2. Panel A: Model path. We reported the significant (in black) and non-significant (in grey) paths with the associated b statistics. Panel B: Fit indices of our model.

8.4.5 Indirect effects

As a second step, building upon the significant paths in the path model analysis, we tested the mediating effect of anthropomorphism between the individual tendency to be collectivistic towards the adoption of intentional stance. As for the previous analysis, all datapoints were corrected using z-scores. The strategy to calculate the 95% Confidence Interval was kept constant (a bias corrected percentile bootstrap analysis with 10, 000 samples), as the Maximum Likelihood estimator.

The indirect effects of Collectivism on IST through the HRIES were significant when mediated by the Sociability subscale [b = 0.025, bootSE = 0.011, z = 2.253, p = 0.024, bootC.I. 95% (0.006; 0.054)], the Animacy subscale [b = 0.016, bootSE = 0.008, z = 1.996, p = 0.046, bootC.I. 95% (0.003; 0.039)] and the Agency subscale [b = 0.023, bootSE = 0.010, z = 2.424, p = 0.015, bootC.I. 95% (0.007; 0.047)]. No indirect effect of Collectivism on IST mediated by the Disturbing subscale (bootC.I. included 0, see Table 9.2).

Table 8.3. Indirect effects

									95% C.I.	
					Estimate	Std. Error	z-value	p	Lower	Upper
CVS_COLL	→	HRIES_SOC	→	IST	0.025	0.011	2.253	0.024	0.006	0.054
CVS_COLL	→	HRIES_AN	→	IST	0.016	0.008	1.996	0.046	0.003	0.039
CVS_COLL	→	HRIES_AG	→	IST	0.023	0.010	2.424	0.015	0.007	.047
CVS_COLL	→	HRIES_DIST	→	IST	0.001	0.005	0.280	0.780	-0.009	0.015

Note. Delta method standard errors, bias-corrected percentile bootstrap confidence intervals, ML estimator. NB: Not all bootstrap samples were successful: CI based on 9998 samples.

Following Zhao and colleagues classification (Zhao et al., 2010), anthropomorphism seems to play the role of a true mediator between the individual tendency to be collectivistic and the adoption of the intentional stance towards a humanoid robot. This seems particularly true for the mediation role of the perceived level of Sociability, Animacy and Agency of the robot, while the Disturbing seemed not to mediate the effect. Nonetheless, the total indirect effect of Collectivism on the IST is significant [b = 0.065, bootSE = 0.018, z = 3.623, p = $<$ .001, bootC.I. 95% (0.029; 0.104)], indicating the total mediating role of attributed anthropomorphism on the relationship between the individual tendency towards collectivism and the individual likelihood of adopting the intentional stance towards a humanoid robot (Table 9.2). The higher the level of collectivism, the higher the anthropomorphic attribution, the higher the likelihood of adopting the intentional stance towards a humanoid robot.

8.4.6 Exploratory analyses

In addition, to confirm that our model was indeed the best to describe the relation between the considered constructs, we compared post-hoc, our mediation analysis with a second model that considered the adoption of the intentional stance as mediator (collectivism →intentional stance →anthropomorphism). Results showed that there is no mediation effect of the intentional stance on the considered anthropomorphism constructs (model parameters are reported in Appendix E). This result showed that the influence of collectivism seemed to be on anthropomorphism and, only through anthropomorphism, we can observe an influence on the adoption of the intentional stance.

8.5 Discussion

In the present paper, we aimed at disentangling the relationship between cultural values, the tendency to anthropomorphize, and the tendency to adopt the intentional stance towards a humanoid robot. To address this aim, we first built a structural model to test whether our hypothesized relationship between the variables (and relative sub-constructs) would hold: differences in cultural values should influence anthropomorphism and the adoption of the intentional stance. Results showed that cultural values, in particular the level of individualism/collectivism, directly influence anthropomorphism, but not adoption of the intentional stance. The higher the collectivism, the higher the anthropomorphism. Our second step was to investigate whether anthropomorphism would play a mediating role between collectivism and the adoption of the intentional stance. The mediation results showed that when considering the role of anthropomorphism as a mediator between culture (as a predictor) and the adoption of the intentional stance towards robots (as the outcome), the relationship was significant. In particular, our results showed that the Agency, Sociability, and Animacy subscales of the HRIES were positive mediators between collectivism and adoption of the intentional stance. Interestingly, the "disturbing" subscale of HRIES (i.e., negative appraisal) seems to rely on a parallel process, as it is not influenced by collectivism. However, it still has a positive effect on Intentional Stance. The higher the score on the disturbing subscale, the higher the IST. A tentative interpretation of this result could rely on the distinctiveness threat hypothesis discussed by Ferrari et al. (2016). The authors argue that individuals could feel threatened by the increased (imaginative) similarity of robots' appearance and capacities to humans (F. Ferrari et al., 2016). In this context, the observer does not adopt a design stance but, on the contrary, overestimates the abilities of the robot, attributing to the artificial agent humanlike characteristics such as intentions. In other words, the distinctiveness threat results in an overestimation of the human-like properties of the robot (as a negative expectation).

More generally, the model showed a strong influence of collectivism on anthropomorphism and – through anthropomorphism – on the adoption of intentional stance towards a humanoid robot. The relationship in the reverse direction (intentional stance → anthropomorphism) was not observed. In other words, cultural values, in particular collectivism, influence our perception of robots in terms of anthropomorphic attributions. This result is in line with previous literature that shows how collectivistic cultures tend to anthropomorphize, like more and trust more, anthropomorphically-shaped or zoomorphic robots (for a review see Lim et al., 2020). The present results extend the previous findings by demonstrating that adoption of intentional stance towards robots depends on anthropomorphism in a context of cultural values. Therefore, interpretation of the robot's actions would depend on prior attribution of anthropomorphic characteristics. This attribution would be influenced by the cultural values adopted by the observer.

Future literature should attempt to create a vaster and more comprehensive theoretical model of how these variables and their facets interact with each other when humans predict the behaviour of a robot. Moreover, more theoretical and empirical investigations should explore the presence of new variables that may play a role in the adoption of intentional stance towards humanoid robots

Part V

Project results

Chapter 9

Conclusions

9.1 Overview of results

The Ph.D. project presented in this thesis aimed at unravelling the factors that might contribute to evoking in humans the adoption of the intentional stance towards humanoid robots. The adoption of the intentional stance is a default strategy adopted by humans to predict and interpret the behaviours of other (social) agents they might face (Dennett, 1971; Dennett, 1989). To address the key aims of this project, we conducted a series of studies described in this thesis within Parts 2-4.

In **Part 2**, I describe a novel tool that we developed (namely the InStance Test, IST) to assess the adoption of the intentional stance. The IST consists of 34 pictorial scenarios depicting the humanoid iCub robot involved in several activities, either alone or with other human characters. Each scenario is associated with two sentences: one is describing the scenario making references to mental states (therefore, representing the adoption of the intentional stance), the other sentence is describing the scenario by means of a mechanistic explanation (and therefore, representing the adoption of the design stance). We showed that, indeed, it is possible for humans to adopt the intentional stance towards a humanoid robot. Interestingly, we also observed that participants differ in their individual likelihood of adopting the intentional stance towards a robot (Publication 1). This was later explored in Publication 2, where we examined the behavioural correlates of individual differences in the likelihood to adopt the intentional/design stance.

We then proceeded with the investigation of the behavioural indices of the adoption of the intentional stance towards two different agents (a humanoid robot compared to a human). We adapted the IST to measure participants' response times (RTs) when judging each scenario associated with one sentence at a time. This modification allowed us to have an implicit measure (RTs) associated with the explicit one (participants' choice). To be able to compare RTs across various agents, we created a version of the IST with a human agent as the main character instead of the iCub robot. Results demonstrated a dissociation between implicit and explicit measures: although the intentional stance might be considered as a strategy to predict and explain the behaviour of a humanoid robot (i.e., implicit level, A. I. Jack, Dawson, Begany et al., 2013; Schilbach et al., 2008) when facing a humanoid robot, humans might rationalize their attributions by consciously choosing to adopt the design stance. This means that although participants might by default adopt the intentional stance towards a robot, they might switch to the design stance upon deliberate reflection. Furthermore, to investigate the role of the individual tendency to anthropomorphize non-human agents, we clustered our participants according to their anthropomorphism score and found that those who were faster in adopting the intentional stance towards the robot, were the ones with a higher anthropomorphism score (Publication 2).

Part 3 aimed at exploring whether it was possible to manipulate humans' adoption of the intentional stance towards humanoid robots in an interactive scenario. To this end, we designed a series of experiments that required our participants to face the embodied robot iCub and to be involved in a task with it. In Publication 3 we asked our participants to observe the iCub robot involved in a decision-making task while exerting two different types of behaviours: a more hesitant and a more decisive behaviour. The frequency of hesitation was manipulated between groups (80% for Group 1, 20% for Group 2). Participants were asked to infer the strategy used by the robot during the task. Before and after the task with the robot, our participants filled a battery of questionnaires and tests to assess the adoption of the intentional stance before and after the task, their individual differences in personality, expectations and attitudes towards robots. Results showed that our participants in Group 2 had a higher likelihood of adopting the intentional stance after having observed the robot, compared to participants in Group 1. Moreover, we reported correlations between the IST and the other questionnaires that help us to highlight the link between the adoption of the intentional stance and humans' attitudes and expectations about robots.

Subsequently, we further explored the physiological responses to humanlike vs. machinelike behaviours of the robot (Publication 4). In this experiment, we asked our participants to observe the iCub robot behave either in a more human-like or machine-like way. All participants observed both conditions in two different sessions. After each session, they were asked to complete the IST while their pupillary response was recorded. Results showed that participants' pupillary response was predictive of their individual bias towards intentional stance, measured by the IST. Moreover, we report that both participants with a higher likelihood of adopting the intentional stance and the ones with a higher likelihood of adopting the design stance, were opting more often for the mentalistic attributions in the IST after observing the robot's humanlike behaviour, as compared to the machinelike behaviour. This shows that participants had some sensitivity to the subtle differences in the robot behaviour, thereby attributing more "humanness" to the humanlike behaviour, independently of their initial tendency. Furthermore, participants with a higher tendency towards the intentional stance showed a greater pupil dilation when choosing the mentalistic option, regardless of the behaviour of the robot (indicating cognitive effort to adhere to their tendency), whereas the mechanistic participants showed a greater pupil dilation when opting for the description that was "incongruent" with the behaviour they observed during the robot session (i.e., mentalistic option after machinelike behaviour). We conclude that participants with a tendency to adopt the intentional stance might pay less attention to behavioural cues exerted by the robot, sticking to their initial tendency, while participants with a tendency for the design stance are more prone to notice different subtle cues in the robot's behaviour, leading them to change the stance they initially adopted (and therefore, causing a higher cognitive effort).

Finally, as presented in Part 3, we designed a series of three experiments to investigate whether the adoption of the intentional stance can be modulated by different robot behaviours in the same experiential context (Publication 5). In Experiment 1 and 2 we introduced our participants into a room and asked them to sit beside the robot. Via a Wizard of Oz (WOZ) manipulation, the robot would welcome participants and verbally interact with them. After the verbal interaction, the robot would invite participants to watch some videos together. During this phase, the robot would emotionally react in a contingent way to the events shown in the videos. Finally, the robot concluded the interaction by saying goodbye to participants via a WOZ manipulation. In these two experiments, we used two different ways to split the IST questionnaire into pre- and post sessions. Results showed that the degree to which intentional stance was adopted increased after the interaction with a humanlike behaving robot. This was independent of the way how the pre- and post- items of the IST were distributed. In Experiment 3, we used the version of pre- post- IST split presented in Experiment 2. We manipulated the behaviours shown by the robot to be very mechanical. The robot displayed a series of calibration movements. There was no verbal interaction and no social consideration of the participants by the robot. Moreover, during the videos, the robot would show always the same set of movements. Its reactions would not be emotionally contingent on the events in the videos, and the robot would only display mechanical beeping sounds. This was done in order to make the conditions relatively equivalent with respect to the amount of auditory stimulation. Results showed this kind of robot behaviour did not modulate the IST score (pre- vs. post interaction).

Part 4 aimed to investigate whether the adoption of the intentional stance would be modulated by cultural differences. The original plan involved data collection in collaboration with A*STAR (Singapore) and my physical presence at A*STAR. Unfortunately, due to the COVID-19 pandemic, it was not possible to collect the Singaporean sample during my stay at A*STAR facilities in the period 07/02/2020-07/05/2020. As a contingency measure, an online study was carried out to examine the impact of cultural differences on the likelihood of adoption of the intentional stance towards robots.

To explore the relationship between cultural differences, anthropomorphism, and the adoption of the intentional stance, we collected a large sample of participants from different countries and administered a battery of tests and questionnaires. We then modelled our data with a path model (SEM) and observed a consistent and direct effect of collectivism on anthropomorphism, but no direct effect on the adoption of the intentional stance. Thus, to further explore this result, we performed a mediation analysis where anthropomorphism would play the role of mediator between collectivism and the adoption of the intentional stance. Indeed, results showed anthropomorphism as a true mediator in the relationship between collectivism and the adoption of the intentional stance towards a humanoid robot (Publication 6).

9.2 Scientific contribution and impact

The present thesis aimed at investigating whether humans would predict and interpret the behaviour of humanoid robots by ascribing them intentions, beliefs, and desires. In other words, whether humans would adopt the intentional stance towards these new (social) agents.

To pursue this aim, we asked three research questions, stated in Chapter 2. I will now provide answers to each research question.

RQ 1 - *As intentional stance is a philosophical concept, how can we operationalize it and design empirical tests of the intentional stance? Is it possible to identify neural and physiological markers of adoption of the intentional stance?*

To answer the first part of RQ 1, we developed the InStance Test (IST), a tool that aims at assessing the adoption of the intentional stance towards humanoid robots. We administered the IST to a pool of participants and results showed that indeed humans can adopt the intentional stance toward robots to some extent. Following this result, we adapted the IST to explore the behavioural (response times) indices and neurophysiological (EEG) correlates of the adoption of the intentional stance. We showed a dissociation between the explicit choice of participants and their behavioural response in the IST. Moreover, we found that it is possible to predict the adoption of the intentional (or design) stance from EEG activity at rest. Thus, we were able to identify the behavioural and neurophysiological underpinnings of the adoption of the intentional stance as measured by the IST.

RQ 2 – *Can we modulate the likelihood of adopting the intentional stance towards a humanoid robot by manipulating the robot's behaviour?*

Once we developed and tested the IST as a measure of the adoption of the intentional stance, we tested it in more interactive scenarios where participants were facing the embodied robot iCub. We designed several experiments, showing that indeed, the degree of human-likeness embedded in the robot's behaviour can influence the individual likelihood to adopt the intentional stance and that physiological responses (measured by pupil dilation) reflect the cognitive effort in switching from one stance to another.

RQ 3 - *Is the adoption of the intentional stance towards a humanoid robot modulated by cultural differences and the individual tendency to anthropomorphize robots?*

Given the consequences of the Covid-19 pandemic on this research question, we explored it from a different angle. We created a path model of the relationship between culture, anthropomorphism, and the adoption of the intentional stance. We showed that anthropomorphism is a mediator between the tendency to be collectivistic and the individual likelihood of adopting the intentional stance. The higher the collectivistic stance, the higher the anthropomorphic attribution, the higher the likelihood of adopting the intentional stance towards a humanoid robot. These results show that, in order to be able to understand what leads humans to attribute mental states to humanoid robots, we need to take into account not only their individual differences in personality and attitudes towards robots and technology but their cultural background as well, since it can indirectly influence our perception of robots.

The general goal of the present Ph.D. thesis was to present the investigation of the possible application of a philosophical and psychological concept, namely the adoption of the intentional stance, Dennett, 1989), to the field of human-robot interaction. This makes the present thesis an interdisciplinary journey towards an integration of the research fields that can result in a cross-contamination of methodologies, theories, and impact. In the next paragraph, I will discuss how the application of theories and methodologies from social and cognitive psychology helped in casting a light on how humans' process the adoption of the intentional stance towards an artificial agent. I will then discuss the implication of these results both for the theory of the intentional stance applied to HRI and for possible future applications.

9.2.1 How to investigate humans' social cognition in human-robot interaction: methods and impact of results on the field

To pursue the general aim of exploring whether humans would adopt the intentional stance towards robots, I employed the theoretical, methodological and practical skills acquired during my training as a cognitive scientist. These methods include the creation of empirical designs that apply different measures to test the hypotheses, and the use of different statistical analyses and tools to adapt to the specific experimental context and data. Moreover, the interdisciplinary approach of the present thesis allows the impact to spread across disciplines. In the following subsections, I will discuss the methods used and the impact of the results. I will also present some research that was inspired by some works previously presented. I will finally discuss future directions that can arise from the most recent findings related to this thesis.

Impact of the IST as a tool to assess the adoption of the intentional stance in HRI

The first method was the psychometric evaluation of the IST. With the help of psychometric tools such as the Principal Component Analysis (PCA) and Item Analysis, we were able to investigate the structure of the IST. The results reported in Chapter 3 led to the use of the IST in several experimental contexts, both involving screen-based and the embodied robot. For example, Spatola and colleagues (Spatola et al., 2021c) further evaluated the psychometric properties of the IST. The authors identified a two-factor structure, named "Social" and "Non-Social", that identified a subset of 12 items that are specifically informative about the (non-) socialness of the depicted interaction. This further psychometric evaluation was employed by O'Reilly and colleagues (O'Reilly et al., 2021). The authors explored how autistic traits modulate the IST scores in the Social and Non-Social factors. The IST has been employed as an evaluation test by several studies investigating the adoption of the intentional stance with different paradigms (in addition to the ones reported in the present thesis in Chapters 5, 6, and 7). For example, recently Abubshait and Wykowska (Abubshait and Wykowska, 2020) reported a decreased adoption of the intentional stance after long exposure to repetitive robot behaviour; Ciardo and colleagues (Ciardo et al., 2021) reported a decreased adoption of the intentional stance when participants are exposed to mechanical erring movements. Willemse and colleagues (Willemse et al., 2022) found that individual differences in adopting the intentional stance moderate motor behaviour in a screen based task mimicking a gaze following paradigm. Parenti and colleagues used the IST to assess the adoption of the intentional stance towards a virtual avatar of the iCub robot, compared to the virtual avatar of a human kid (Parenti et al., 2021). In a series of three experiments, the authors highlight that the adoption of the intentional stance towards a virtual agent can be modulated by participants' expectations.

In summary, the development of the IST employed psychometrics methods to operationalise a philosophical concept to create a tool that is flexible and that can serve HRI empirical research.

The IST as tool to investigate human social cognition

The second set of methods used within this thesis are techniques from classical cognitive psychology and neuroscience. Specifically, Chapter 4 showed how it is possible to adapt the IST to a paradigm that investigates the cognitive processes related to the decision-making process in the adoption of the intentional stance. Although, compared to the classical cognitive psychology paradigms, there are some limitations (i.e., the long response times), Chapter 4 presents the first attempt to find a behavioural (and thus, quantitative) measure of the cognitive processes underlying the adoption of the intentional stance in HRI. Recent literature, reported an adaptation of the IST to an EEG recording setting and showed that it is possible to identify neural oscillation that predicts the stance adoption at the IST (Bossi et al., 2020). This result confirms that the adoption of the intentional is a default mechanism, closely related to the Default Mode Network (DMN, Spunt et al., 2015). In addition, the work published by Bossi and colleagues shows that the IST is a versatile tool that can serve researchers both in social neuro-cognition and HRI.

Chapter 6 represents the first integration of the use of the IST to assess the adoption of the intentional stance towards the embodied robot and the recording of a physiological measure (i.e., pupil dilation). Results from such experimental adaptation of the IST may influence not only social cognition and HRI, but human factors as well. Following the exploration of the cognitive processes and costs of the intentional stance, Spatola et al., (Spatola et al., 2021a) explored in a series of 4 experiments how loading cognitively participants at an early stage of the scenario observation, results in a higher mentalistic attribution. Future investigations on the cognitive flexibility required to switch from one stance and the other could help to cast a light on the relation between the intentional stance and the DMN. A deeper understanding of this relationship could allow drawing more definitive conclusions on the interplay between lower and higher level cognitive processes active in adopting the intentional stance, both towards other humans and artificial agents.

Is the robot acting as I would? Integrating the phenomenological approach in the discussion about the adoption of the intentional stance and human-likeness

As humans, we are first and foremost experts in predicting other humans' behaviours. I already discussed in Chapter 2 several theories that investigate how and why this phenomenon happens. Nevertheless, if "others" include non-humans animals, the discussion becomes more tricky. That is because we cannot be sure that animals, especially the non-mammals, have at least similar capabilities to mental states (Penn and Povinelli, 2007; Premack, 2007). The further from humans we go, the more difficult it is to argue that a system has mental states. And robots are quite far from humans. But what about a humanoid robot? This is where the intentional stance becomes a very flexible strategy to adopt when interpreting the behaviour of another agent: as previously explained, the adoption of the intentional stance does not require the observed system to *actually* have mental states. Adopting of the intentional stance happens within the observer's mind, meaning that it is the observer that assumes that the interpreted system is rational, and therefore, must act accordingly and follow its mental states. This "beholder" point of view allows us to consider any system among the candidates towards which humans could adopt the intentional stance, humanoid robots included. I also discussed how the humanoid shape may facilitate the activation of the "human model" (Epley et al., 2007; Kahn and Shen, 2017; Wiese et al., 2017), and how this model could induce the adoption of the intentional stance towards them. Thus, in the quest of understanding what types of behaviour may induce or facilitate the adoption of the intentional stance, it is plausible that the more human-like the behaviour, the higher the probability of adopting the intentional stance towards the robot.

Throughout the thesis, I presented three different approaches to the implementation of behaviours on the robot. In Chapter 5 the robot was showing two different behaviours (a decisive behaviour vs. a hesitant behaviour) and we learned that it is not only the quality of the behaviour (i.e., hesitant or not) to increase the adoption of the intentional stance, but the frequency with which this behaviour was displayed. In other words, what can induce the adoption of the intentional stance towards a humanoid robot is also the inherent resemblance to a crucial aspect of human behaviour: variability. Following the idea of implementing a behaviour that was as human-like and variable as possible, in Chapter 6 we implemented the behaviour recorded from directly from a human. This allowed the robot to show variable and unpredictable, subtle patterns of movement. Although recent research shows that humans are not able to explicitly recognise a subtle human behaviour implemented on a robot, their implicit responses seem to be affected even by such subtle variations (Ghiglino, De Tommaso et al., 2020; Ghiglino et al., 2018; Ghiglino et al., 2021).

Results from Chapter 6 are in line with the literature, showing that some participants were affected by the behaviours, resulting in different pupil dilation when switching from one stance to the other. So far, the behaviours implemented on the robot were mostly focused on being "as human-like as possible". Although this approach was showing interesting results, it was only focused on making the robot closer to the human behaviour in the attempt to evoke the human model in the observer. However, this strategy was excluding the human observer from the creation of the meaningful context to adopt the intentional stance, almost resulting in considering the human as a passive agent to be activated by the robot's behaviours. Thus, within Chapter 7 we took a different approach, that would include the human participant in the creation of the contextual experience of adopting the intentional stance. Although I don't consider myself as a phenomenologist, I do consider ourselves embedded in environments that we experience and make sense of in an active (and sometime enactive) way. With the help of methods from social psychology, Chapter 7 showed that when participants are embedded in a meaningful and familiar context (i.e., watching videos with another agent), where the human-like behaviours are created to be perceived as matching a given context (plausible reactions to emotion-eliciting events in the videos), humans do adopt the intentional stance more easily towards the robot. This increase in the likelihood of adoption of the intentional stance could result from the perception that the robot is acting "like-me" (Meltzoff, 2007). That is because when we are sharing a social context with agents whose behaviours are perceived as "how I would behave", we might activate what Fuchs and De Jaegher (Fuchs and De Jaegher, 2009) defined as "mutual affective resonance". This mutual attunement between agents, could result in an attunement of their affective and kinematics behaviours, ultimately resulting in a "mutual incorporation" of the other in our perception of the experience (Fuchs, 2017). Moreover, as mentioned in Chapter 2, Higgins argued in favour of the minimal relational self as one of the ontological primitive of selfhood (Higgins, 2020). Therefore, we could speculate that the shared social and affective context that the participants experienced in Chapter 7, led them to build the expectation that the robot could be an interactive-embodied agent able to perceive the context similar to how they were experiencing it themselves. Once the "like-me" perception was elicited, or in other words, once the human model was activated, participants spontaneously adopted the intentional stance towards the iCub.

9.2.2 Summary of research impact: from the theory to the application

In the previous paragraph, I described how methods from social and cognitive psychology and social neuroscience can help the researchers to investigate and explore how humans attribute mental state to (social) humanoid robots. The results emerging from such an interdisciplinary approach should impact both HRI and social cognition studies. Hence, the way we (as researchers) think about the humanoid robot itself should be twofold: as a powerful and reliable tool to evoke and study human cognition (for a review see Wykowska, 2020), or as proper social companion and assistive agents in healthcare and educational context (for reviews see Prescott and Robillard, 2021; Ramsey et al., 2021; Wykowska, 2021).

The present thesis offers to the communities of HRI and cognitive science a tool that can assess the degree to which intentional stance is adopted towards a humanoid robot (Chapter 3). This would allow predicting which stance will be adopted by an individual (i.e., Chapter 4). Moreover, the understanding of which robotic behavioural cues induce various stances (intentional, design in humans) (Chapter 5, 6, and 7) could be extremely useful to design robots that elicit either stances, depending on the purpose of the robot itself. This is relevant especially in the context of human-factor, where it would be crucial to consider the cognitive cost of adopting either stance and to switch from one stance to the other. For example, in a real life interaction with a humanoid robot, one could combine the methods and results from Chapters 4, 6 and 7 to understand whether that context and that type of human-robot interaction could be negatively (or positively) affected by the adoption of the intentional stance under stressful conditions for the human.

As we saw throughout with the thesis, the adoption of the intentional stance is an individual tendency, and as such, it varies across us. While for me, in a stressful situation it might be cognitively costly switching to the design stance, for another the opposite could be true. On top of the individual differences, results from experiments such as the one presented in Chapter 8 can help to consider also cultural differences in the design of robots, always with the aim to create robots that can easily interact and socially attune with us. First steps are already taken in the development of cognitive architectures that would allow robots to be "truly social", based on the incorporation of evidence from social cognition in HRI (for a review see Kotseruba and Tsotsos, 2018). For example, recently, Prescott and Camilleri (Prescott and Camilleri, 2019) present the challenges of implementing a robot sense of self by developing equivalent sub-systems within an integrated biomimetic cognitive architecture for a humanoid robot, such as the iCub robot. Vinanzi and colleagues (Vinanzi et al., 2021) propose the implementation of an intention reading artificial model that could allow the robot to correctly predict the goals of their human partners. Such implementation should allow an easier social attunement within the interaction and lead to a higher likelihood of positive outcomes from the interaction. Lombardi et al. (Lombardi et al., 2022) developed an attentive architecture for the iCub robot. Specifically, Lombardi et al. endowed the robot with the ability to reliably detect whether the human is engaging in mutual gaze with it. This cognitive architecture could be implemented during a joint task between the humans and the robot, and use the mutual gaze both as a parameter of the human's attentional resources and as a trigger for the robot to implement the subsequent behaviour. Robotic architectures such as the one briefly presented above, in combination with the results reported in the present thesis could make it possible to imagine real life "IST scenarios" in the near future, where humans and humanoid robots are sharing the same environments, and in which the robots are perceived as truly social partners, able to promptly interact with us (Henschel et al., 2021)

9.3 Limitations and future directions

In the present thesis, I described an approach in which we used methods from cognitive neuroscience and cognitive psychology to combine explicit and implicit measurements to investigate the adoption of the intentional stance. However, some limitations regarding the IST and the experimental approaches need to be acknowledged.

First, I would like to explore some limitation related to the InStance Test. The IST has been tested and validated only with the iCub robot as the main character. Although the adoption of the intentional stance is considered a strategy that should apply to any type of agent, we proved that the level of anthropomorphism and human-likeness might have an influence on the likelihood of adopting it. Therefore, future works should explore this point by developing, for example, a virtual version of the IST or by changing the main robot character. Such modification would allow generalising results from this thesis. In addition, it would allow researchers in HRI that work with different platforms to access the tool and apply it to their research. Moreover, the IST presents participants with two possible explanations (a mentalistic and a mechanistic one), providing already a set of mentalistic (and mechanistic) semantic representations. This does not allow us to definitely conclude that participants, provided with only the scenarios and simply asked to describe the behaviour of the robot, would have naturally adopted the intentional stance. Nevertheless, evidence from literature tells us that brain oscillations in the resting state phase (thus, before being exposed to the IST) are predictive of the stance (intentional or designed) adopted when completing the IST (Bossi et al., 2020). This result tells us that the tendency to adopt either stance is a default mechanism that is active before being exposed to any artificial agent. This being said, future research should investigate the narrative evoked by the IST scenarios dissociated by their options. This evaluation would allow a comparison between the vocabularies used by participants, with the one used in the two options, leading to a possible future semantic analysis of the language that emerged in the observation of the scenarios. On a more technical side, a second limitation of the IST that needs to be discussed is the slider scale without ticks. Although the choice of the ticks was taken to avoid biasing participants to overthink their stance adoption, it is true that evidence shows that the lack of ticks can result in a biased distribution (Matejka et al., 2016). Although some recent studies investigated the psychometric composition of the IST (Spatola et al., 2021b, 2021c), no study to investigate how the appearance of the scale would influence the distribution of the IST scores has been conducted yet. Future studies should address this issue to further explore the density distribution of the IST scores. Finally, the IST items consist of static images of the robots involved in different actions (either alone or with one or two humans). The choice of opting for static scenarios instead of videos was made for two main reasons:

1. we opted to follow the structure implemented by the most influential empirical tests for Theory of Mind and mentalization developed in literature (Baron-Cohen, 1997; Baron-Cohen et al., 2001; Völlm et al., 2006.

2. evidence shows that several factors influence the attribution of a mind to an artificial agent, one of which is the kinematics of the robot itself (Cross et al., 2016; Gray and Wegner, 2012; Saygin et al., 2010; Saygin et al., 2012). Specifically, these results show that a violation of the expectations in the behaviour of the robot (i.e., it resembles a human, but it does not have a biological motion) can lead to the "uncanny valley effect" (Mori et al., 2012; Urgen et al., 2018) and, therefore, to a different degree of adoption of the intentional stance. That is, robotic motion has to be considered "yet another factor" to be experimentally controlled and manipulated.

Now that the static version of the IST has been validated and implemented (although more psychometric evaluations are still needed), future studies should expand the investigation, creating, for example, a virtual version of the IST, where the robot movements can be compared to a human virtual agent's movements. Moreover, a video version of the IST with the embodied robot as the main character could allow for a comparison with the virtual and the static version of the IST, helping to cast a light on the importance of embodiment to mind perception and the adoption of the intentional stance towards humanoid robots. Finally, the IST depicts the iCub robot as the main character. Thus, researchers in HRI who use other robotic platforms finds it difficult to use the IST as a measure, as the depicted robot in the test and the one used during the interaction might have different shapes. A solution would be to test if the IST reveals to be a consistent measure with other depicted robots, and/or test the generalisability of the results by using the IST as it is to evaluate other robots. This latter solution would require to control also for a number of different factors that we know having an influence on the adoption of the intentional stance, such as anthropomorphism and partiicpants' cultural background.

To further explore the limitations to the present thesis, it is important to discuss the nature of the presented studies, which were either screen-based or run in a laboratory setup. Indeed, according to the second person-neuroscience framework (Schilbach et al., 2013), we should consider that in a more natural and ecological scenario outside the laboratory, the socio-cognitive processes involved might differ from the ones observed and reported in this thesis. Future studies should be designed taking into account this factor and further explore how, for example, being involved in a joint task would affect the adoption of the intentional stance towards an artificial agent. First steps in this direction have been taken by Ciardo and colleagues (Ciardo et al., 2021) and Abubshait and colleagues (Abubshait et al., 2021), who report two studies where they observed modulations of the adoption of the intentional stance in socially framed joint tasks.

Finally, a limitation that has to be mentioned is the unfortunate lack of data from the study implemented in Singapore, at A*STAR. The Covid-19 pandemic emergency and subsequent restrictions did not allow me to complete the data collection within the expected timeline. Nonetheless, we managed to implement a contingency plan and to organize a second stay in Singapore to finish the data collection, unfortunately already after the planned completion of this Ph.D. project. The contingency plan consisted of an online study where 600 participants from six different countries across the world were recruited. The aim of the study was to build a structural model of how the cultural value of collectivism influence the adoption of the intentional stance via the level of anthropomorphisation of the robot. Although the number of participants was sufficient to build the model, the experiment is inherently limited for two main reasons:

1. **the reductionist approach** that was adopted as the theoretical framework. On the one hand, Hofstede's account allows considering the individual variability in the cultural values explored with his framework. That is, considering how cultural values vary within each individual. On the other hand, this approach does not consider the variability across

the lifetime in the same values or in the case of cultural mixed individuals (for example, individuals who have parents from different cultures, or that are raised in a country different from the one of origin). Future studies should approach the problem with a less reductionist approach, considering the cultural context as a whole and not only made of single individuals.

2. **the inclusion criteria** applied to data collection. As reported in Appendix E, the inclusion criteria lead the data collection to include individuals from specific countries, thus future studies should replicate the results to check their consistency and, in case, compare any differences. For example, as we had to include participants who were fluent in English, one could hypothetise that the same experiment with a different translation of the items could modulate the results. This is because of the different morphological and semantics conveyed by different languages. A professional translation could be performed to help address this issue.

I plan to be able to expand further the answer to RQ3 with the next data collection in Singapore. In particular, preliminary results show that in the same experimental setup reported in Chapter 7 (adapting the WoZ interaction by translating the sentences into English), Singaporean participants do increase their intentional stance adoption in the Human-like condition compared to the Machine-like condition, but the increase is lower compared to the Italian sample. This result could be interpreted in light of the cultural differences in the activation of the “like-me” processes. One could speculate that Singaporean participants may have different expectations about the same shared social context due to a different minimal relational self-development (Marchesi et al., in prep).

I hope that the work presented here will serve as a basis for future researchers who wish to explore the overlap between human social cognition and human-robot interaction. I also hope that the integration between methods from cognitive psychology and the implementation of more ecological scenarios often used in HRI will represent a new and exciting challenge for new methodological and theoretical exchanges between research fields, leading to a greater corpus of multidisciplinary research.

9.4 Concluding Remarks

Within the present thesis, we discussed the importance of empirically examining how humans’ social cognition mechanisms are activated when we face a new kind of (possibly) social agents: humanoid robots. In a future where these entities will populate our social environments, it is crucial to understand the factors that will allow an easy communication, the creation of a sense of bonding and trust between humans and social robots. Robots that are socially attuned with humans may assist us in many aspects of our lives, from healthcare to home assistance. Therefore, it is crucial to design robots that would take into consideration how humans perceive, predict and react to their behaviours.

In conclusion, investigating not only the robot behaviour as the potential factor influencing adoption of the intentional stance, but also the individual and cultural differences should eventually provide specific guidelines to facilitate the integration of artificial agents into our social environments. I expect that the findings would translate into the design of socially attuned humanoid robots, capable of understanding the specific needs of each user, including their different personalities and cultures.

References

Aaltonen, I., Arvola, A., Heikkilä, P. & Lammi, H. (2017). Hello pepper, may i tickle you? children's and adults' responses to an entertainment robot at a shopping mall. *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 53–54.

Abu-Akel, A. M., Apperly, I. A., Wood, S. J. & Hansen, P. C. (2020). Re-imaging the intentional stance. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1925), 1–9. https://doi.org/10.1098/rspb.2020.0244

Abubshait, A., Perez-Osorio, J., De Tommaso, D. & Wykowska, A. (2021). Collaboratively framed interactions increase the adoption of intentional stance towards robots. *2021 30th IEEE International Conference on Robot Human Interactive Communication (RO-MAN)*, 886–891. https://doi.org/10.1109/RO-MAN50785.2021.9515515

Abubshait, A. & Wykowska, A. (2020). Repetitive Robot Behavior Impacts Perception of Intentionality and Gaze-Related Attentional Orienting. *Frontiers in Robotics and AI*, *7*(November), 1–11. https://doi.org/10.3389/frobt.2020.565825

Agassi, J. (1973). Anthropomorphism in science.

Airenti, G. (2018). The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. *Frontiers in Psychology*, *9*(NOV), 1–13. https://doi.org/10.3389/fpsyg.2018.02136

Airenti, G., Cruciani, M. & Plebe, A. (2019). The cognitive underpinnings of anthropomorphism. *Frontiers in psychology*, *10*, 1539.

Andrews, K. & Huss, B. (2014). Anthropomorphism, anthropectomy, and the null hypothesis. *Biology & Philosophy*, *29*(5), 711–729.

Apperly, I. A. (2008). Beyond simulation–theory and theory–theory: Why social cognitive neuroscience should use its own concepts to study "theory of mind". *Cognition*, *107*(1), 266–283. https://doi.org/https://doi.org/10.1016/j.cognition.2007.07.019

Apperly, I. A. & Butterfill, S. A. (2009). Do Humans Have Two Systems to Track Beliefs and Belief-Like States? *Psychological Review*, *116*(4), 953–970. https://doi.org/10.1037/a0016923

Baack, S. A., Brown, T. S. & Brown, J. T. (1991). Attitudes toward computers: Views of older adults compared with those of young adults. *Journal of Research on Computing in Education*, *23*(3), 422–433.

Bacher, J., Wenzig, K. & Vogler, M. (2004). *Spss twostep cluster - a first evaluation* (Vol. 2004-2). Universität Erlangen-Nürnberg, Wirtschafts- und Sozialwissenschaftliche Fakultät, Sozialwissenschaftliches Institut Lehrstuhl für Soziologie.

Bandura, A. (2002). Social cognitive theory in cultural context. *Applied Psychology*, *51*(2), 269–290. https://doi.org/10.1111/1464-0597.00092

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. MIT press.

Baron-Cohen, S., Leslie, A. M. & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Baron-Cohen, S., Ring, H. A., Wheelwright, S., Bullmore, E. T., Brammer, M. J., Simmons, A. & Williams, S. C. (1999). Social intelligence in the normal and autistic brain: An fMRI study. *European Journal of Neuroscience*, *11*(6), 1891–1898. https://doi.org/10.1046/j.1460-9568.1999.00621.x

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. & Plumb, I. (2001). The "reading the mind in the eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241–251.

Barrouillet, P. (2015). Theories of cognitive development: From piaget to today.

Bartneck, C., Kulić, D., Croft, E. & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Bartneck, C. & Reichenbach, J. (2005). Subtle emotional expressions of synthetic characters. *International journal of human-computer studies*, *62*(2), 179–192.

Bates, D., Kliegl, R., Vasishth, S. & Baayen, H. (2015). Parsimonious Mixed Models. (2000). http://arxiv.org/abs/1506.04967

Becchio, C., Koul, A., Ansuini, C., Bertone, C. & Cavallo, A. (2018). Seeing mental states: An experimental strategy for measuring the observability of other minds. *Physics of Life Reviews*, *24*, 67–80. https://doi.org/https://doi.org/10.1016/j.plrev.2017.10.002

Birks, M., Bodak, M., Barlas, J., Harwood, J. & Pether, M. (2016). Robotic Seals as Therapeutic Tools in an Aged Care Facility: A Qualitative Study. *Journal of Aging Research*, *2016*. https://doi.org/10.1155/2016/8569602

Bohl, V. & van den Bos, W. (2012). Towards an integrative account of social cognition: Marrying theory of mind and interactionism to study the interplay of Type 1 and Type 2 processes. *Frontiers in Human Neuroscience*, *6*(SEPTEMBER), 1–15. https://doi.org/10.3389/fnhum.2012.00274

Bolis, D. & Schilbach, L. (2020). ‘I Interact Therefore I Am’: The Self as a Historical Product of Dialectical Attunement. *Topoi*, *39*(3), 521–534. https://doi.org/10.1007/s11245-018-9574-0

Boothby, E. J., Clark, M. S. & Bargh, J. A. (2014). Shared Experiences Are Amplified. *Psychological Science*, *25*(12), 2209–2216. https://doi.org/10.1177/0956797614551162

Bossi, F., Willemse, C., Cavazza, J., Marchesi, S., Murino, V. & Wykowska, A. (2020). The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots. *Science Robotics*, *5*(46). https://doi.org/10.1126/SCIROBOTICS.ABB6652

Breazeal, C. & Scassellati, B. (1999). How to build robots that make friends and influence people. *Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No. 99CH36289)*, *2*, 858–863.

Brentano, H. & Heidegger, S. (1874). What is phenomenology? *Psychology from an empirical standpoint. London: Routledge & Kegan Paul.*

Brewer, M. B. & Chen, Y.-R. (2007). Where (who) are collectives in collectivism? toward conceptual clarification of individualism and collectivism. *Psychological review*, *114*(1), 133.

Butterfill, S. A. & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind and Language*, *28*(5), 606–637. https://doi.org/10.1111/mila.12036

Butterfill, S. A. & Sinigaglia, C. (2014). Intention and motor representation in purposive action. *Philosophy and Phenomenological Research*, *88*(1), 119–145.

Cacioppo, J. T. & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Carpinella, C. M., Wyman, A. B., Perez, M. A. & Stroessner, S. J. (2017). The Robotic Social Attributes Scale (RoSAS): Development and Validation. *ACM/IEEE International Conference on Human-Robot Interaction*, *Part F1271*(March), 254–262. https://doi.org/10.1145/2909824.3020208

Chalmers, D. (2007). The hard problem of consciousness. *The Blackwell companion to consciousness*, 225–235.

Chaminade, T., Hodgins, J. & Kawato, M. (2007). Anthropomorphism influences perception of computer-animated characters’ actions. *Social Cognitive and Affective Neuroscience*, *2*(3), 206–216. https://doi.org/10.1093/scan/nsm017

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutcher, E., Cheng, G. & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*, *6*(MAY 2012), 1–9. https://doi.org/10.3389/fnhum.2012.00103

Chaminade, T., Zecca, M., Blakemore, S.-J., Takanishi, A., Frith, C. D., Micera, S., Dario, P., Rizzolatti, G., Gallese, V. & Umiltà, M. A. (2010). Brain response to a humanoid robot in areas implicated in the perception of human emotional gestures. *PLoS one*, *5*(7), e11577.

Chen, Y.-R., Brockner, J. & Chen, X.-P. (2002). Individual–collective primacy and ingroup favoritism: Enhancement and protection effects. *Journal of Experimental Social Psychology*, *38*(5), 482–491.

Chin, M. G., Sims, V. K., Clark, B. & Lopez, G. R. (2004). Measuring individual differences in anthropomorphism toward machines and animals. *Proceedings of the human factors and ergonomics society annual meeting*, *48*(11), 1252–1255.

Churchland, P. M. (1981). Eliminative materialism and propositional attitudes. *the Journal of Philosophy*, *78*(2), 67–90.

Ciardo, F., De Tommaso, D. & Wykowska, A. (2021). Effects of erring behavior in a human-robot joint musical task on adopting intentional stance toward the icub robot. *2021 30th IEEE International Conference on Robot Human Interactive Communication (RO-MAN)*, 698–703. https://doi.org/10.1109/RO-MAN50785.2021.9515434

Ciardo, F., De Tommaso, D. & Wykowska, A. (2022). Joint action with artificial agents: Human-likeness in behaviour and morphology affects sensorimotor signaling and social inclusion. *Computers in Human Behavior*, *132*, 107237. https://doi.org/https://doi.org/10.1016/j.chb.2022.107237

Ciaunica, A. & Fotopoulou, A. (2017). The touched self: Psychological and philosophical perspectives on proximal intersubjectivity and the self. MIT Press.

Cohen, A. R., Stotland, E. & Wolfe, D. M. (1955). An experimental investigation of need for cognition. *The Journal of Abnormal and Social Psychology*, *51*(2), 291.

Cross, E. S., Ramsey, R., Liepelt, R., Prinz, W. & Hamilton, A. F. d. C. (2016). The shaping of social perception by stimulus and knowledge cues to human animacy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686), 20150075.

Cross, E. S., Riddoch, K. A., Pratts, J., Titone, S., Chaudhury, B. & Hortensius, R. (2019). A neurocognitive investigation of the impact of socializing with a robot on empathy for pain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1771). https://doi.org/10.1098/rstb.2018.0034

Cullen, H., Kanai, R., Bahrami, B. & Rees, G. (2013). Individual differences in anthropomorphic attributions and human brain structure. *Social Cognitive and Affective Neuroscience*, *9*(9), 1276–1280. https://doi.org/10.1093/scan/nst109

Currie, G., Ravenscroft, I. et al. (2002). *Recreative minds: Imagination in philosophy and psychology*. Oxford University Press.

Dacey, M. (2017). Anthropomorphism as cognitive bias. *Philosophy of Science*, *84*(5), 1152–1164.

Davidson, D. (1999). The emergence of thought. *Erkenntnis*, *51*(1), 511–521.

De Gee, J. W., Knapen, T. & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(5), 1–8. https://doi.org/10.1073/pnas.1317557111

De Houwer, J., Teige-Mocigemba, S., Spruyt, A. & Moors, A. (2009). Implicit Measures: A Normative Analysis and Review. *Psychological Bulletin*, *135*(3), 347–368. https://doi.org/10.1037/a0014211

De Jaegher, H. & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the cognitive sciences*, *6*(4), 485–507.

De Jaegher, H., Di Paolo, E. & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in cognitive sciences*, *14*(10), 441–447.

Demoulin, S., Leyens, J.-P., Paladino, M.-P., Rodriguez-Torres, R., Rodriguez-Perez, A. & Dovidio, J. (2004). Dimensions of "uniquely" and "non-uniquely" human emotions. *Cognition and emotion*, *18*(1), 71–96.

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, *68*(4), 87–106.

Dennett, D. C. (1981). True believers: The intentional strategy and why it works.

Dennett, D. C. (1983). Intentional systems in cognitive ethology: The "Panglossian paradigm" defended. *Behavioral and Brain Sciences*, *6*(3), 343–355. https://doi.org/10.1017/S0140525X00016393

Dennett, D. C. (1990). The interpretation of texts, people and other artifacts. *Philosophy and phenomenological research*, *50*, 177–194.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Dewey, J. A., Pacherie, E. & Knoblich, G. (2014). The phenomenology of controlling a moving object with another person. https://doi.org/10.1016/j.cognition.2014.05.002

Dietrich, E. & Markman, A. B. (2003). Discrete thoughts: Why cognition must use discrete representations. *Mind & Language*, *18*(1), 95–119.

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, *30*(100), 412–431. https://doi.org/10.1016/0001-6918(69)90065-1

Dupree, C. H. & Fiske, S. T. (2017). Universal dimensions of social signals: Warmth and competence.

Durt, C., Fuchs, T. & Tewes, C. (2017). *Embodiment, enaction, and culture: Investigating the constitution of the shared world*. MIT Press.

Ebstein, R. P., Israel, S., Chew, S. H., Zhong, S. & Knafo, A. (2010). Genetics of human social behavior. *Neuron*, *65*(6), 831–844.

Edwards, C., Edwards, A., Spence, P. R. & Westerman, D. (2016). Initial interaction expectations with robots: Testing the human-to-human interaction script. *Communication Studies*, *67*(2), 227–238.

Enkh-Amgalan, R. (2016). The indulgence and restraint cultural dimension: A cross-cultural study of mongolia and the united states.

Epley, N., Akalis, S., Waytz, A. & Cacioppo, J. T. (2008). Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and ghreyhounds. *Psychological Science*, *19*(2), 114–120. https://doi.org/10.1111/j.1467-9280.2008.02056.x

Epley, N., Waytz, A., Akalis, S. & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, *26*(2), 143–155. https://doi.org/10.1521/soco.2008.26.2.143

Epley, N., Waytz, A. & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Eyssel, F., Kuchenbrandt, D., Bobinger, S., De Ruiter, L. & Hegel, F. (2012). 'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism. *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 125–126. https://doi.org/10.1145/2157689.2157717

Ferketich, S. (1991). Focus on psychometrics. aspects of item analysis. *Research in nursing & health*, *14*(2), 165–168.

Ferrari, F., Paladino, M. P. & Jetten, J. (2016). Blurring Human–Machine Distinctions: Anthropomorphic Appearance in Social Robots as a Threat to Human Distinctiveness. *International Journal of Social Robotics*, *8*(2), 287–302. https://doi.org/10.1007/s12369-016-0338-y

Ferrari, P. F. & Rizzolatti, G. (2014). Mirror neuron research: The past and the future.

Fink, J. (2012). Anthropomorphism and human likeness in the design of robots and human-robot interaction. *International Conference on Social Robotics*, 199–208.

Fisher, J. A. (1991). Disambiguating anthropomorphism: An interdisciplinary review. *Perspectives in Ethology*, *9*(9), 49–85.

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. & Frith, C. D. (1995). Other minds in the brain: a functional imaging study of ”theory of mind” in story comprehension. *Cognition*, *57*(2), 109–128. https://doi.org/10.1016/0010-0277(95)00692-R

Friese, M., Hofmann, W. & Wänke, M. (2008). When impulses take over: Moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behaviour. *British Journal of Social Psychology*, *47*(3), 397–419.

Frith, C. D. & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*, *63*, 287–313. https://doi.org/10.1146/annurev-psych-120710-100449

Fuchs, T. (2013). The phenomenology and development of social perspectives. *Phenomenology and the cognitive sciences*, *12*(4), 655–683.

Fuchs, T. (2017). Intercorporeality and interaffectivity. *Intercorporeality: Emerging socialities in interaction*, 3–23.

Fuchs, T. & De Jaegher, H. (2009). Enactive intersubjectivity: Participatory sense-making and mutual incorporation. *Phenomenology and the cognitive sciences*, *8*(4), 465–486.

Gallagher, H. L., Jack, A. I., Roepstorff, A. & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage*, *16*(3 I), 814–821. https://doi.org/10.1006/nimg.2002.1117

Gallagher, S. (2001). The practice of mind: Theory, simulation or primary interaction? *Journal of Consciousness Studies*, *8*(5-7), 83–108.

Gallagher, S. (2012). In defense of phenomenological approaches to social cognition: Interacting with the critics. *Review of Philosophy and Psychology*, *3*(2), 187–212.

Gallagher, S. & Zahavi, D. (2020). *The phenomenological mind*. Routledge.

Gallese, V. (2001). The’shared manifold’hypothesis. from mirror neurons to empathy. *Journal of consciousness studies*, *8*(5-6), 33–50.

Gallese, V. (2005). Embodied simulation: From neurons to phenomenal experience. *Phenomenology and the cognitive sciences*, *4*(1), 23–48.

Gallotti, M. & Frith, C. D. (2013). Social cognition in the we-mode. *Trends in Cognitive Sciences*, *17*(4), 160–165. https://doi.org/10.1016/j.tics.2013.02.002

Ghiglino, D., De Tommaso, D., Willemse, C., Marchesi, S. & Wykowska, A. (2020). Can I get your (robot) attention? Human sensitivity to subtle hints of human-likeness in a humanoid robot’s behavior. https://doi.org/10.31234/osf.io/kfy4g

Ghiglino, D., De Tommaso, D. & Wykowska, A. (2018). Attributing human-likeness to an avatar: The role of time and space in the perception of biological motion. *International Conference on Social Robotics*, 400–409.

Ghiglino, D., Willemse, C., De Tommaso, D., Bossi, F. & Wykowska, A. (2020). At first sight: Robots' subtle eye movement parameters affect human attentional engagement, spontaneous attunement and perceived human-likeness. *Paladyn, Journal of Behavioral Robotics*, *11*(1), 31–39.

Ghiglino, D., Willemse, C., De Tommaso, D. & Wykowska, A. (2021). Mind the Eyes: Artificial Agents' Eye Movements Modulate Attentional Engagement and Anthropomorphic Attribution. *Frontiers in Robotics and AI*, *8*(May), 1–13. https://doi.org/10.3389/frobt.2021.642796

Ghiglino, D. & Wykowska, A. (2020). When robots (pretend to) think. *Artificial intelligence* (pp. 49–74). mentis.

Gilbert, M. (2009). Shared intention and personal intentions. *Philosophical Studies*, *144*(1), 167–187. https://doi.org/10.1007/s11098-009-9372-z

Gliem, J. A. & Gliem, R. R. (2003). Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales.

Golan, O., Baron-Cohen, S., Hill, J. J. & Golan, Y. (2006). The "reading the mind in films" task: Complex emotion recognition in adults with and without autism spectrum conditions. *Social Neuroscience, 1*(2), 111–123.

Goldberg, L. R. (1993). 'The structure of phenotypic personality traits": Author's reactions to the six comments. *American Psychologist*, *48*(12), 1303–1304. https://doi.org/10.1037/0003-066x.48.12.1303

Goldman, A. (1993). The psychology of folk psychology. *Behavioral and Brain sciences*, *16*(1), 15–28.

Goldman, A. (2005). Imitation, mind reading, and simulation. *Perspectives on imitation: From neuroscience to social science*, *2*, 79–93.

Gonzàles, A., Ramirez, M. & Viadel, V. (2012). Attitudes of the elderly toward information and communication technologies. *Educ. Gerontol*, *38*(9), 585–594.

Gopnik, A. & Wellman, H. M. (1992). Why the child's theory of mind really is a theory.

Gould, S. J. (1996). Can we truly know sloth and rapacity? *Natural History*, *105*(4), 18–25.

Gray, K. & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130. https://doi.org/https://doi.org/10.1016/j.cognition.2012.06.007

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A. & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological review*, *109*(1), 3.

Grezes, J. & Decety, J. (2001). Functional anatomy of execution, mental simulation, observation, and verb generation of actions: A meta-analysis. *Human brain mapping*, *12*(1), 1–19.

Happé, F. & Frith, U. (1995). Theory of Mind in Autism. *Learning and Cognition in Autism*, 177–197. https://doi.org/10.1007/978-1-4899-1286-2_10

Haring, K. S., Silvera-Tawil, D., Matsumoto, Y., Velonaki, M. & Watanabe, K. (2014). Perception of an android robot in japan and australia: A cross-cultural comparison. *International conference on social robotics*, 166–175.

Haslam, N., Bain, P., Douge, L., Lee, M. & Bastian, B. (2005). More human than you: Attributing humanness to self and others. *Journal of personality and social psychology*, *89*(6), 937.

Hegel, F., Krach, S., Kircher, T., Wrede, B. & Sagerer, G. (2008). Understanding social robots: A user study on anthropomorphism. *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 574–579.

Heider, F. & Simmel, M. (1944). University of Illinois Press http://www.jstor.org/stable/1416950 . *The American Journal of Psychology*, *57*(2), 243–259.

Henschel, A., Laban, G. & Cross, E. S. (2021). What makes a robot social? a review of social robots from science fiction to a home or hospital near you. *Current Robotics Reports*, *2*(1), 9–19.

Hess, E. H. & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, 1190–1192.

Higgins, J. (2020). The ‘we’in ‘me’: An account of minimal relational selfhood. *Topoi*, *39*(3), 535–546.

Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture*, *2*(1), 1–26. https://doi.org/10.9707/2307-0919.1014

Horstmann, A. C. & Krämer, N. C. (2019). Great expectations? Relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment. *Frontiers in Psychology*, *10*(APR), 1–14. https://doi.org/10.3389/fpsyg.2019.00939

Hortensius, R. & Cross, E. S. (2018). From automata to animate beings: The scope and limits of attributing socialness to artificial agents. *Annals of the New York Academy of Sciences*, *1426*(1), 93–110. https://doi.org/10.1111/nyas.13727

Huebner, B. (2010). Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? *Phenomenology and the cognitive sciences*, *9*(1), 133–155.

Jack, A. I., Dawson, A. J., Begany, K. L., Leckie, R. L., Barry, K. P., Ciccia, A. H. & Snyder, A. Z. (2013). Fmri reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage*, *66*, 385–401.

Jack, A. I., Dawson, A. J. & Norr, M. E. (2013). Seeing human: Distinct and overlapping neural signatures associated with two forms of dehumanization. *NeuroImage*, *79*, 313–328.

Jack, A. I. & Robbins, P. (2012). The phenomenal stance revisited. *Review of Philosophy and Psychology*, *3*(3), 383–403.

Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G. & Caldara, R. (2009). Cultural Confusions Show that Facial Expressions Are Not Universal. *Current Biology*, *19*(18), 1543–1548. https://doi.org/10.1016/j.cub.2009.07.051

Jackson, I. & Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental Science*, *12*(4), 670–679. https://doi.org/10.1111/j.1467-7687.2008.00805.x

Jensen, A. R. (1987). Mental chronometry in the study of learning disabilities. *Mental Retardation & Learning Disability Bulletin*.

Johnson, S. C. (2000). The recognition of mentalistic agents in infancy. *Trends in Cognitive Sciences*, *4*(1), 22–28.

Jones, R. A. (2021). Projective anthropomorphism as a dialogue with ourselves. *International Journal of Social Robotics*, 1–7.

Kahn, P. H., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., Ruckert, J. H. & Gill, B. (2011). The new ontological category hypothesis in human-robot interaction. *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 159–160. https://doi.org/10.1145/1957656.1957710

Kahn, P. H. & Shen, S. (2017). Noc noc, who's there? a new ontological category (noc) for social robots. *New perspectives on human development*, 13–142.

Kaplan, A. D., Sanders, T. & Hancock, P. A. (2019). The relationship between extroversion and the tendency to anthropomorphize robots: A Bayesian analysis. *Frontiers Robotics AI*, *6*(JAN). https://doi.org/10.3389/frobt.2018.00135

Kaplan, F. (2004). Who is afraid of the humanoid? investigating cultural differences in the acceptance of robots. *International Journal of Humanoid Robotics*, *01*(03), 465–480. https://doi.org/10.1142/S0219843604000289

Kelley, J. F. (1983). An empirical methodology for writing user-friendly natural language computer applications. *Conference on Human Factors in Computing Systems - Proceedings*, (December), 193–196. https://doi.org/10.1145/800045.801609

Khan, R. & Cox, P. (2017). Country culture and national innovation. *Archives of Business Research*, *5*(2).

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

Kitayama, S., Karasawa, M., Curhan, K. B., Ryff, C. D. & Markus, H. R. (2010). Independence and interdependence predict health and wellbeing: Divergent patterns in the united states and japan. *Frontiers in psychology*, *1*, 163.

Knoblich, G. & Sebanz, N. (2006). The social nature of perception and action. *Current directions in psychological science*, *15*(3), 99–104.

Komatsu, T., Kurosawa, R. & Yamada, S. (2012). How does the difference between users' expectations and perceptions about a robotic agent affect their behavior? *International Journal of Social Robotics*, *4*(2), 109–116.

Kompatsiari, K., Pérez-Osorio, J., De Tommaso, D., Metta, G. & Wykowska, A. (2018). Neuroscientifically-grounded research for improved human-robot interaction. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3403–3408. https://doi.org/10.1109/IROS.2018.8594441

Kompatsiari, K., Bossi, F. & Wykowska, A. (2021). Eye contact during joint attention with a humanoid robot modulates oscillatory brain activity. *Social cognitive and affective neuroscience*, *16*(4), 383–392. https://doi.org/10.1093/scan/nsab001

Kompatsiari, K., Ciardo, F., Tikhanoff, V., Metta, G. & Wykowska, A. (2018a). On the role of eye contact in gaze cueing. *Scientific reports*, *8*(1), 1–10.

Kompatsiari, K., Ciardo, F., Tikhanoff, V., Metta, G. & Wykowska, A. (2018b). On the role of eye contact in gaze cueing. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-36136-2

Kool, W., McGuire, J. T., Rosen, Z. B. & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of experimental psychology: general*, *139*(4), 665.

Kotseruba, I. & Tsotsos, J. K. (2018). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*(1), 17–94. https://doi.org/10.1007/s10462-018-9646-y

Kovačić, Z. J. (2005). The impact of national culture on worldwide egovernment readiness. *Informing Science*, *8*.

Kozak, M. N., Marsh, A. A. & Wegner, D. M. (2006). What do i think you're doing? action identification and mind attribution. *Journal of personality and social psychology*, *90*(4), 543.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F. & Kircher, T. (2008). Can machines think? interaction and perspective taking with robots investigated via fmri. *PloS one*, *3*(7), e2597.

Kret, M. E. & Sjak-Shie, E. E. (2019). Preprocessing pupil size data: Guidelines and code. *Behavior Research Methods*, *51*(3), 1336–1342. https://doi.org/10.3758/s13428-018-1075-y

Kriegel, U. (2020). *The oxford handbook of the philosophy of consciousness*. Oxford University Press.

Kuchenbrandt, D., Eyssel, F., Bobinger, S. & Neufeld, M. (2013). When a robot's group membership matters. *International Journal of Social Robotics*, *5*(3), 409–417.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G. & Banaji, M. R. (2019). Relationship between the implicit association test and intergroup behavior: A meta-analysis. *American psychologist*, *74*(5), 569.

Larsen, R. S. & Waters, J. (2018). Neuromodulatory correlates of pupil dilation. *Frontiers in Neural Circuits*, *12*(March), 1–9. https://doi.org/10.3389/fncir.2018.00021

Lemhöfer, K. & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Leslie, A. M. (1994). Pretending and believing: issues in the theory of ToMM. *Cognition*, *50*(1-3), 211–238. https://doi.org/10.1016/0010-0277(94)90029-9

Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Lim, V., Rooksby, M. & Cross, E. S. (2020). Social Robots on a Global Stage: Establishing a Role for Culture During Human–Robot Interaction. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-020-00710-4

Lombardi, M., Maiettini, E., De Tommaso, D., Wykowska, A. & Natale, L. (2022). Toward an attentive robotic architecture: Learning-based mutual gaze estimation in human–robot interaction. *Frontiers in Robotics and AI*, *9*. https://doi.org/10.3389/frobt.2022.770165

Luce, R. D. et al. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press on Demand.

Malle, B. F. & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101–121. https://doi.org/https://doi.org/10.1006/jesp.1996.1314

Marchesi, S., Perez-Osorio, J., De Tommaso, D. & Wykowska, A. (2020). Don't overthink: Fast decision making combined with behavior variability perceived as more human-like. *29th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2020*, 54–59. https://doi.org/10.1109/RO-MAN47096.2020.9223522

Marchesi, S., Bossi, F., Ghiglino, D. & Tommaso, D. D. (2021). I Am Looking for Your Mind : Pupil Dilation Predicts Individual Differences in Sensitivity to Hints of Human-Likeness in Robot Behavior. *8*(June), 1–10. https://doi.org/10.3389/frobt.2021.653537

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E. & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology*, *10*(MAR). https://doi.org/10.3389/fpsyg.2019.00450

Marchesi, S., Roselli, C. & Wykowska, A. (2021). Cultural values, but not nationality, predict social inclusion of robots. *International Conference on Social Robotics*, 48–57.

Marchesi, S., Spatola, N., Wykowska, A. & Perez-Osorio, J. (2021). Human vs humanoid. A behavioral investigation of the individual tendency to adopt the intentional stance. *ACM/IEEE International Conference on Human-Robot Interaction*, 332–340. https://doi.org/10.1145/3434073.3444663

Markus, H. R. & Kitayama, S. (1991). 17. Markus &Kitayama (1991). *Psychological Review*, *98*(2), 224–253.

Martini, M. C., Gonzalez, C. A. & Wiese, E. (2016). Seeing minds in others–can agents with robotic appearance have human-like preferences? *PloS one*, *11*(1), e0146310.

Matejka, J., Glueck, M., Grossman, T. & Fitzmaurice, G. (2016). The effect of visual appearance on the performance of continuous sliders and visual analogue scales. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5421–5432. https://doi.org/10.1145/2858036.2858063

Mathôt, S. (2018). Pupillometry: Psychology, Physiology, and Function. *Journal of Cognition*, *1*(1), 1–23. https://doi.org/10.5334/joc.18

Mathôt, S., Schreij, D. & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

McLeod, S. (2007). Jean piaget's theory of cognitive development.

Meltzoff, A. N. (2007). 'Like me': A foundation for social cognition. *Developmental Science*, *10*(1), 126–134. https://doi.org/10.1111/j.1467-7687.2007.00574.x

Metta, G., Fitzpatrick, P. & Natale, L. (2006). YARP: Yet another robot platform. *International Journal of Advanced Robotic Systems*, *3*(1), 043–048. https://doi.org/10.5772/5761

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., von Hofsten, C., Rosander, K., Lopes, M., Santos-Victor, J., Bernardino, A. & Montesano, L. (2010). The iCub humanoid robot: An open-systems platform for research in cognitive development. *Neural Networks*, *23*(8-9), 1125–1134. https://doi.org/10.1016/j.neunet.2010.08.010

Meyer, M. L. (2019). Social by Default: Characterizing the Social Functions of the Resting Brain. *Current Directions in Psychological Science*, *28*(4), 380–386. https://doi.org/10.1177/0963721419857759

Michael, J. (2011). Interactionism and mindreading. *Review of Philosophy and Psychology*, *2*(3), 559–578.

Mitchell, R. W., Thompson, N. S. & Miles, H. L. (1997). *Anthropomorphism, anecdotes, and animals*. Suny Press.

Mori, M., MacDorman, K. F. & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics Automation Magazine*, *19*(2), 98–100. https://doi.org/10.1109/MRA.2012.2192811

Mubin, O., Ahmad, M. I., Kaur, S., Shi, W. & Khan, A. (2018). Social robots in public spaces: A meta-review. *International Conference on Social Robotics*, 213–220.

Naefgen, C., Dambacher, M. & Janczyk, M. (2018). Why free choices take longer than forced choices: evidence from response threshold manipulations. *Psychological Research*, *82*(6), 1039–1052. https://doi.org/10.1007/s00426-017-0887-1

Nagel, T. (1974). What is it like to be a bat. *Readings in philosophy of psychology*, *1*, 159–168.

Nass, C. & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, *56*(1), 81–103.

Natale, L., Bartolozzi, C., Nori, F., Sandini, G. & Metta, G. (2019). Icub. In A. Goswami & P. Vadakkepat (Eds.), *Humanoid robotics: A reference* (pp. 291–323). Springer Netherlands. https://doi.org/10.1007/978-94-007-6046-2_21

Natale, L., Bartolozzi, C., Pucci, D., Wykowska, A. & Metta, G. (2017). Icub: The not-yet-finished story of building a robot child. *Science Robotics*, *2*(13).

Nichols, S. & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds.* Clarendon Press/Oxford University Press.

Nomura, T. (2014). Comparison on negative attitude toward robots and related factors between Japan and the UK. *CABS 2014 - Proceedings of the 5th ACM International Conference on Collaboration Across Boundaries*, 87–90. https://doi.org/10.1145/2631488.2634059

Nomura, T., Suzuki, T., Kanda, T., Yamada, S. & Kato, K. (2011). Attitudes toward robots and factors influencing them. https://doi.org/10.1075/ais.2.06nom

Nomura, T., Syrdal, D. S. & Dautenhahn, K. (2015). Differences on social acceptance of humanoid robots between Japan and the UK. *AISB Convention 2015*.

Olsson, U. H., Foss, T., Troye, S. V. & Howell, R. D. (2000). The performance of ml, gls, and wls estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural equation modeling*, *7*(4), 557–595.

O'Reilly, Z., Ghiglino, D., Spatola, N. & Wykowska, A. (2021). Modulating the intentional stance: Humanoid robots, narrative and autistic traits. In H. Li, S. S. Ge, Y. Wu, A. Wykowska, H. He, X. Liu, D. Li & J. Perez-Osorio (Eds.), *Social robotics* (pp. 697–706). Springer International Publishing.

Özdem, C., Wiese, E., Wykowska, A., Müller, H., Brass, M. & Van Overwalle, F. (2017). Believing androids–fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents. *Social Neuroscience*, *12*(5), 582–593. https://doi.org/10.1080/17470919.2016.1207702

Pacherie, E. (2014). How does it feel to act together? *Phenomenology and the Cognitive Sciences*, *13*(1), 25–46. https://doi.org/10.1007/s11097-013-9329-8

Papadopoulos, I. & Koulouglioti, C. (2018). The Influence of Culture on Attitudes Towards Humanoid and Animal-like Robots: An Integrative Review. *Journal of Nursing Scholarship*, *50*(6), 653–665. https://doi.org/10.1111/jnu.12422

Parenti, L., Marchesi, S., Belkaid, M. & Wykowska, A. (2021). Exposure to robotic virtual agent affects adoption of intentional stance. *Proceedings of the 9th International Conference on Human-Agent Interaction*, 348–353. https://doi.org/10.1145/3472307.3484667

Pasquali, D., Aroyo, A. M., Gonzalez-Billandon, J., Rea, F., Sandini, G. & Sciutti, A. (2020). Your eyes never lie: A robot magician can tell if you are lying. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 392–394.

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. *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 293–302.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E. & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Penn, D. C. & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 731–744.

Perez-Osorio, J., Marchesi, S., Ghiglino, D., Ince, M. & Wykowska, A. (2019). *More Than You Expect: Priors Influence on the Adoption of Intentional Stance Toward Humanoid Robots* (Vol. 11876 LNAI). Springer International Publishing. https://doi.org/10.1007/978-3-030-35888-4_12

Perez-Osorio, J. & Wykowska, A. (2019). Adopting the intentional stance towards humanoid robots. *Wording robotics* (pp. 119–136). Springer.

Perez-Osorio, J. & Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, *33*(3), 369–395. https://doi.org/10.1080/09515089.2019.1688778

Perner, J. (1991). *Understanding the representational mind.* The MIT Press.

Perner, J. & Wimmer, H. (1985). "john thinks that mary thinks that…" attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, *39*(3), 437–471.

Posner, M. I. (1978). *Chronometric explorations of mind.* Lawrence Erlbaum.

Premack, D. (2007). Human and animal cognition: Continuity and discontinuity. *Proceedings of the National Academy of Sciences*, *104*(35), 13861–13867. https://doi.org/10.1073/pnas.0706147104

Premack, D. & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(4), 515–526.

Prescott, T. J. & Camilleri, D. (2019). The synthetic psychology of the self. In M. I. Aldinhas Ferreira, J. Silva Sequeira & R. Ventura (Eds.), *Cognitive architectures* (pp. 85–104). Springer International Publishing. https://doi.org/10.1007/978-3-319-97550-4_7

Prescott, T. J. & Robillard, J. M. (2021). Are friends electric? The benefits and risks of human-robot relationships. *iScience*, *24*(1), 101993. https://doi.org/10.1016/j.isci.2020.101993

Preuschoff, K., 't Hart, B. M. & Einhäuser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience*, *5*(SEP), 1–12. https://doi.org/10.3389/fnins.2011.00115

Prim, A. L., FILHO, L. S., Zamur, G. A. C. & Di Serio, L. C. (2017). The relationship between national culture dimensions and degree of innovation. *International Journal of Innovation Management*, *21*(01), 1730001.

Procházka, A., Mudrová, M., Vyšata, O., Háva, R. & Araujo, C. P. S. (2010). Multi-channel EEG signal segmentation and feature extraction. *INES 2010 - 14th International Conference on Intelligent Engineering Systems, Proceedings*, 317–320. https://doi.org/10.1109/INES.2010.5483824

Raichle, M. E. (2015). The Brain's Default Mode Network. *Annual Review of Neuroscience*, *38*(April), 433–447. https://doi.org/10.1146/annurev-neuro-071013-014030

Ramsey, R., Kaplan, D. M. & Cross, E. S. (2021). Watch and learn: The cognitive neuroscience of learning from others' actions. *Trends in Neurosciences*, *44*(6), 478–491.

Ratcliffe, M. (2006). Folk psychology'is not folk psychology. *Phenomenology and the Cognitive Sciences*, *5*(1), 31–52.

Rau, P. L., Li, Y. & Li, D. (2010). A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics*, *2*(2), 175–186. https://doi.org/10.1007/s12369-010-0056-9

Ray, C., Mondada, F. & Siegwart, R. (2008). What do people expect from robots? *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3816–3821.

Redcay, E. & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews Neuroscience*, *20*(8), 495–505. https://doi.org/10.1038/s41583-019-0179-4

Reddy, V. & Morris, P. (2004). Participants don't need theories: Knowing minds in engagement. *Theory & Psychology*, *14*(5), 647–665.

Remland, M. S., Jones, T. S. & Brinkman, H. (1991). Proxemic and haptic behavior in three European countries. *Journal of Nonverbal Behavior*, *15*(4), 215–232. https://doi.org/10.1007/BF00986923

Riddoch, K. A. & Cross, E. S. (2021). "Hit the Robot on the Head With This Mallet" – Making a Case for Including More Open Questions in HRI Research. *Frontiers in Robotics and AI*, *8*(February), 1–17. https://doi.org/10.3389/frobt.2021.603510

Riek, L. (2012). Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, *1*(1), 119–136. https://doi.org/10.5898/jhri.1.1.riek

Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.*, *27*, 169–192.

Rizzolatti, G., Fogassi, L. & Gallese, V. (2009). The mirror neuron system: A motor-based mechanism for action and intention understanding.

Robbins, P. (2006). The phenomenal stance. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *127*(1), 59–85.

Roesler, E., Manzey, D. & Onnasch, L. (2021). A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics*, *6*(58), eabj5425. https://doi.org/10.1126/scirobotics.abj5425

Roncone, A., Pattacini, U., Metta, G. & Natale, L. (2016). A cartesian 6-DoF gaze controller for humanoid robots. *Robotics: Science and Systems*, *12*. https://doi.org/10.15607/rss.2016.xii.022

Ruijten, P. A., Haans, A., Ham, J. & Midden, C. J. (2019). Perceived Human-Likeness of Social Robots: Testing the Rasch Model as a Method for Measuring Anthropomorphism. *International Journal of Social Robotics*, *11*(3), 477–494. https://doi.org/10.1007/s12369-019-00516-z

Ruxton, G. D. & Beauchamp, G. (2008). Time for some a priori thinking about post hoc testing. *Behavioral Ecology*, *19*(3), 690–693. https://doi.org/10.1093/beheco/arn020

Samani, H., Saadatian, E., Pang, N., Polydorou, D., Fernando, O. N. N., Nakatsu, R. & Koh, J. T. K. V. (2013). Cultural robotics: The culture of robotics and robotics in culture. *International Journal of Advanced Robotic Systems*, *10*, 1–10. https://doi.org/10.5772/57260

Saygin, A. P., Chaminade, T. & Ishiguro, H. (2010). The perception of humans and robots: Uncanny hills in parietal cortex. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*(32).

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J. & Frith, C. (2012). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social cognitive and affective neuroscience*, *7*(4), 413–422.

Schellen, E. & Wykowska, A. (2019). Intentional mindset toward robots-open questions and methodological challenges. *Frontiers Robotics AI*, *6*(JAN), 1–11. https://doi.org/10.3389/frobt.2018.00139

Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R. & Vogeley, K. (2008). Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the ”default system” of the brain. *Consciousness and Cognition*, *17*(2), 457–467. https://doi.org/10.1016/j.concog.2008.03.013

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T. & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(4), 393–414. https://doi.org/10.1017/S0140525X12000660

Schramm, L. T., Dufault, D. & Young, J. E. (2020). Warning: This robot is not what it seems! exploring expectation discrepancy resulting from robot design. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 439–441.

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J. & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, *147*(3), 293.

Sebanz, N., Bekkering, H. & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76. https://doi.org/10.1016/j.tics.2005.12.009

Serafin, M. & Surian, L. (2004). Il test degli occhi: Uno strumento per valutare la" teoria della mente". *Giornale italiano di psicologia*, *31*(4), 839–862.

Spatola, N. (2019). L'homme et le robot, de l'anthropomorphisme à l'humanisation. *Topics in Cognitive Psychology*, (119), 515–563.

Spatola, N., Kühnlenz, B. & Cheng, G. (2020). Perception and evaluation in human-robot interaction: The Human-Robot Interaction Evaluation Scale (HRIES) – a multicomponent approach of anthropomorphism. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-020-00667-4

Spatola, N., Marchesi, S. & Wykowska, A. (2021a). Cognitive load affects early processes involved in mentalizing robot behaviour. https://doi.org/10.31234/osf.io/54bhe

Spatola, N., Marchesi, S. & Wykowska, A. (2021b). The intentional stance test-2: How to measure the tendency to adopt intentional stance towards robots. *Frontiers in Robotics and AI*, *8*. https://doi.org/10.3389/frobt.2021.666586

Spatola, N., Marchesi, S. & Wykowska, A. (2021c). Robot humanization measure/task. https://doi.org/10.31234/osf.io/3gde2

Spatola, N., Monceau, S. & Ferrand, L. (2020). Cognitive Impact of Social Robots: How Anthropomorphism Boosts Performances. *IEEE Robotics and Automation Magazine*, *27*(3), 73–83. https://doi.org/10.1109/MRA.2019.2928823

Spatola, N. & Urbanska, K. (2020). God-like robots: the semantic overlap between representation of divine and artificial entities. *AI and Society*, *35*(2), 329–341. https://doi.org/10.1007/s00146-019-00902-1

Spatola, N. & Wykowska, A. (2021). The personality of anthropomorphism: How the need for cognition and the need for closure define attitudes and anthropomorphic attributions toward robots. *Computers in Human Behavior*, *122*, 106841. https://doi.org/10.1016/j.chb.2021.106841

Spaulding, S. (2010). Embodied cognition and mindreading. *Mind & Language*, *25*(1), 119–140.

Spaulding, S. (2015). Phenomenology of social cognition. *Erkenntnis*, *80*(5), 1069–1089.

Spears, R. & Haslam, S. A. (1997). Stereotyping and the burden of cognitive load.

Spreng, R. N. & Andrews-Hanna, J. R. (2015). *The Default Network and Social Cognition* (Vol. 3). Elsevier Inc. https://doi.org/10.1016/B978-0-12-397025-1.00173-1

Spreng, R. N., Stevens, W. D., Viviano, J. D. & Schacter, D. L. (2016). Attenuated anticorrelation between the default and dorsal attention networks with aging: Evidence from task and rest. *Neurobiology of aging*, *45*, 149–160.

Spunt, R. P., Meyer, M. L. & Lieberman, M. D. (2015). The default mode of human brain function primes the intentional stance. *Journal of cognitive neuroscience*, *27*(6), 1116–1124.

Stenzel, A., Chinellato, E., Bou, M. A. T., Del Pobil, Á. P., Lappe, M. & Liepelt, R. (2012). When humanoid robots become human-like interaction partners: Corepresentation of robotic actions. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(5), 1073.

Stich, S. & Nichols, S. (2003). Folk psychology. *The blackwell guide to philosophy of mind*, 235–55.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S. et al. (2016). Artificial intelligence and life in 2030: The one hundred year study on artificial intelligence.

Strohkorb Sebo, S., Traeger, M., Jung, M. & Scassellati, B. (2018). The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 178–186.

Sultana, N., Meissner, N. & Peng, F. L. Y. (2013). Exploring believable character animation based on principles of animation and acting principles. *Proceedings - 2013 International Conference on Informatics and Creative Multimedia, ICICM 2013*, 321–324. https://doi.org/10.1109/ICICM.2013.69

Tapus, A., Mataric, M. J. & Scassellati, B. (2007). Socially assistive robotics [grand challenges of robotics]. *IEEE robotics & automation magazine*, *14*(1), 35–42.

Thellman, S., Silvervarg, A. & Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology*, *8*(NOV), 1–14. https://doi.org/10.3389/fpsyg.2017.01962

Thellman, S. & Ziemke, T. (2020). Do You See what I See? Tracking the Perceptual Beliefs of Robots. *iScience*, *23*(10), 101625. https://doi.org/10.1016/j.isci.2020.101625

Tomasello, M., Carpenter, M., Call, J., Behne, T. & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(5), 675–691.

Tomasello, M. & Rakoczy, H. (2003). What makes human cognition unique? from individual to shared to collective intentionality. *Mind & language*, *18*(2), 121–147.

Triandis, H. C. (1993). Collectivism and individualism as cultural syndromes. *Cross-cultural research*, *27*(3-4), 155–180.

Tucker, M. L. (1990). A compendium of textbook views on planned versus post hoc tests. *Annual Meeting of the Southwest Educational Research Association*.

Urgen, B. A., Kutas, M. & Saygin, A. P. (2018). Uncanny valley as a window into predictive processing in the social brain. *Neuropsychologia*, *114*, 181–185. https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2018.04.027

Urquiza-Haas, E. G. & Kotrschal, K. (2015). The mind behind anthropomorphic thinking: Attribution of mental states to other species. *Animal Behaviour*, *109*, 167–176. https://doi.org/10.1016/j.anbehav.2015.08.011

Van Kempen, J., Loughnane, G. M., Newman, D. P., Kelly, S. P., Thiele, A., O'Connell, R. G. & Bellgrove, M. A. (2019). Behavioural and neural signatures of perceptual decision-making are modulated by pupil-linked arousal. *eLife*, *8*(50), 1–27. https://doi.org/10.7554/eLife.42541

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods*, *49*(2), 653–673. https://doi.org/10.3758/s13428-016-0721-5

Varnum, M. E., Grossmann, I., Kitayama, S. & Nisbett, R. E. (2010). The origin of cultural differences in cognition: The social orientation hypothesis. *Current Directions in Psychological Science*, *19*(1), 9–13. https://doi.org/10.1177/0963721409359301

Vinanzi, S., Cangelosi, A. & Goerick, C. (2021). The collaborative mind: Intention reading and trust in human-robot interaction. *iScience*, *24*(2), 102130. https://doi.org/https://doi.org/10.1016/j.isci.2021.102130

Völlm, B. A., Taylor, A. N., Richardson, P., Corcoran, R., Stirling, J., McKie, S., Deakin, J. F. & Elliott, R. (2006). Neuronal correlates of theory of mind and empathy: A functional magnetic resonance imaging study in a nonverbal task. *NeuroImage*, *29*(1), 90–98. https://doi.org/10.1016/j.neuroimage.2005.07.022

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard university press.

Waytz, A., Cacioppo, J. & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232. https://doi.org/10.1177/1745691610369336

Waytz, A., Epley, N. & Cacioppo, J. T. (2010). Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science*, *19*(1), 58–62. https://doi.org/10.1177/0963721409359302

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H. & Cacioppo, J. T. (2010). Making sense by making sentient: effectance motivation increases anthropomorphism. *Journal of personality and social psychology*, *99*(3), 410–435. https://doi.org/10.1037/a0020240

Wiese, E., Metta, G. & Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology*, *8*(OCT), 1–19. https://doi.org/10.3389/fpsyg.2017.01663

Wiese, E., Weis, P. P. & Lofaro, D. M. (2018). Embodied social robots trigger gaze following in real-time hri. *2018 15th International Conference on Ubiquitous Robots (UR)*, 477–482.

Wiese, E., Wykowska, A., Zwickel, J. & Müller, H. J. (2012). I see what you mean: How attentional selection is shaped by ascribing intentions to others.

Willemse, C., Abubshait, A. & Wykowska, A. (2022). Motor behaviour mimics the gaze response in establishing joint attention, but is moderated by individual differences in adopting the intentional stance towards a robot avatar. *Visual Cognition*, *30*(1-2), 42–53. https://doi.org/10.1080/13506285.2021.1994494

Willemse, C., Marchesi, S. & Wykowska, A. (2018). Robot Faces that Follow Gaze Facilitate Attentional Engagement and Increase Their Likeability. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2018.00070

Willemse, C. & Wykowska, A. (2019). In natural interaction with embodied robots, we prefer it when they follow our gaze: A gaze-contingent mobile eyetracking study. *Philosophical Transactions of the Royal Society B*, *374*(1771), 20180036.

Wimmer, H. & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Wykowska, A., Kajopoulos, J., Ramirez-Amaro, K. & Cheng, G. (2015). Autistic traits inversely correlate with implicit sensitivity to human-like behavior. *Interac. Stud.*, *16*, 219–248.

Wykowska, A. (2020). Social Robots to Test Flexibility of Human Social Cognition. *International Journal of Social Robotics*, *12*(6), 1203–1211. https://doi.org/10.1007/s12369-020-00674-5

Wykowska, A. (2021). Robots as Mirrors of the Human Mind. *Current Directions in Psychological Science*, *30*(1), 34–40. https://doi.org/10.1177/0963721420978609

Wykowska, A., Chaminade, T. & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1693). https://doi.org/10.1098/rstb.2015.0375

Wykowska, A., Kajopoulos, J., Obando-Leiton, M., Chauhan, S. S., Cabibihan, J.-J. & Cheng, G. (2015). Humans are well tuned to detecting agents among non-agents: Examining the sensitivity of human perception to behavioral characteristics of intentional systems. *International Journal of Social Robotics*, *7*(5), 767–781.

Wykowska, A., Wiese, E., Prosser, A. & Müller, H. J. (2014). Beliefs about the minds of others influence how we process sensory information. *PLoS ONE*, *9*(4). https://doi.org/10.1371/journal.pone.0094339

Yamagishi, T., Jin, N. & Miller, A. S. (1998). In-group bias and culture of collectivism. *Asian Journal of Social Psychology*, *1*(3), 315–328.

Yoo, B., Donthu, N. & Lenartowicz, T. (2011). Measuring hofstede's five dimensions of cultural values at the individual level: Development and validation of CVSCALE. *Journal of International Consumer Marketing*, *23*(3-4), 193–210. https://doi.org/10.1080/08961530.2011.578059

Zahavi, D. (2011). Empathy and direct social perception: A phenomenological proposal. *Review of Philosophy and Psychology*, *2*(3), 541–558.

Zahavi, D. & Gallagher, S. (2008). The (in) visibility of others: A reply to herschbach. *Philosophical Explorations*, *11*(3), 237–244.

Zhang, T., Ramakrishnan, R. & Livny, M. (1996). Birch: An efficient data clustering method for very large databases. *ACM sigmod record*, *25*(2), 103–114.

Zhao, X., Lynch, J. G. & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, *37*(2), 197–206. https://doi.org/10.1086/651257

Zickuhr, K. & Smith, A. (2012). Digital differences. *Pew Research Center's Internet &American Life Project*, 1–41. http://www.english.illinois.edu/-people-/faculty/debaron/482/482readings/PEW_Class.pdf

Złotowski, J., Strasser, E. & Bartneck, C. (2014). Dimensions of anthropomorphism: From humanness to humanlikeness. *ACM/IEEE International Conference on Human-Robot Interaction*, 66–73. https://doi.org/10.1145/2559636.2559679

Zwickel, J. (2009). Agency attribution and visuospatial perspective taking. *Psychonomic Bulletin and Review*, *16*(6), 1089–1093. https://doi.org/10.3758/PBR.16.6.1089

Appendices

Appendix A

Supplementary Materials Chapter 3

A.1 Human control experiment

A.2 Materials and methods

A.2.1 Sample

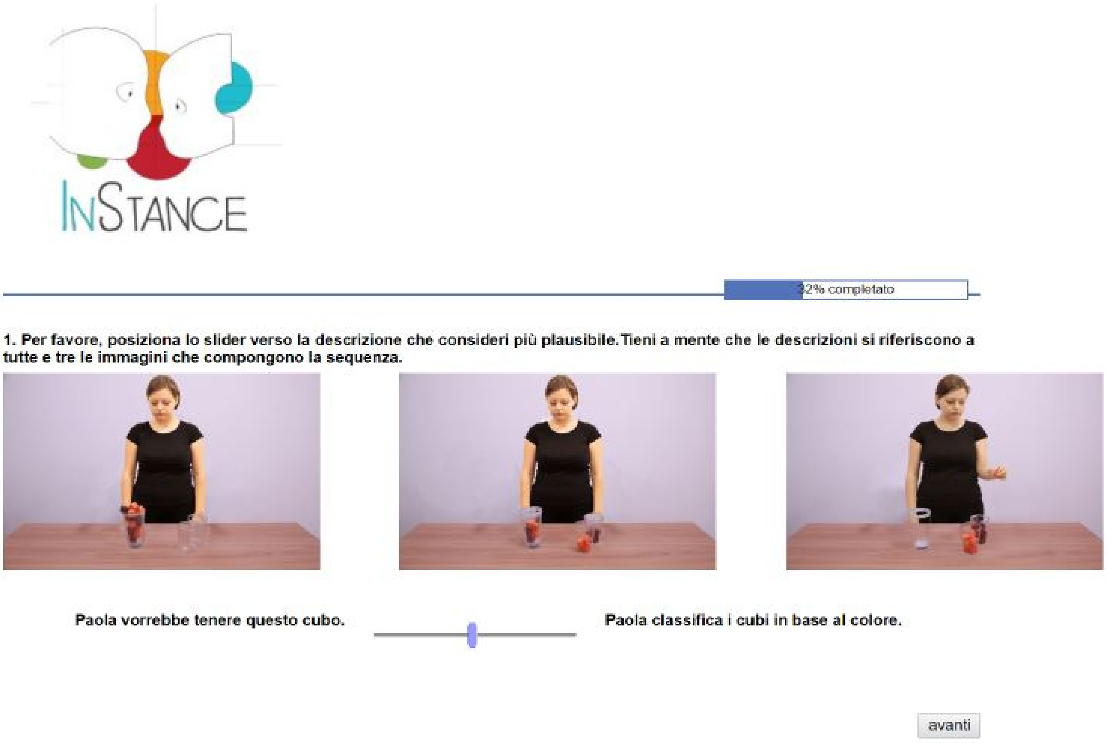
First, we collected data of one hundred and twenty Italian native speakers with different social and educational backgrounds (see Table 9.1 for demographical details) who completed our Human Control Test (HCT): HCT mapping not-counterbalanced, N= 120). Due to a mistake in counterbalancing the position (left vs. right) of mechanistic or mentalistic descriptions, we administered the HCT with the proper counterbalanced mapping to a second sample of hundred and three participants, from which we excluded one participant who already filled in the first HCT: HCT mapping counterbalanced, N=102. Data collection was conducted in accordance with the ethical standards laid down in the Code of Ethics of the World Medical Association (Declaration of Helsinki), procedures were approved by the regional ethics committee (Comitato Etico Regione Liguria).

	Demographic characteristics
Age(year), mean (SD) [min - max]	$29.81(10.39)[18-68]$
Female, n (%)	$153(68.92\%)$
Education (years), mean (SD) [min - max]	$16.20(2.97)[8,24]$

Table A.1. Demographic details of the sample (N= 222).

A.2.2 Human Control Test (HCT)

We selected 15 scenarios out of the 34 original ones that could be adapted to a human agent with respect to the mechanistic descriptions (we excluded all items that used very implausible mechanistic descriptions such as motor calibration) and digitally edited them (AdobePhotoshop CC 2018) depicting a human agent (Paola) instead of the iCub robot. Each item of the HCT was identical with the IST, except for the depicted agent in the scenarios, which for HCT was a human (see Figure A.1 for an example, all scenarios included in the HCT are included below). Each scenario was composed of three pictures (size 900 x 173.2 pixels). Out of the 15 scenarios, 9 involved one (or more) other human interacting with Paola; 1 scenario showed a human arm pointing to an object; 6 scenarios depicted only Paola. The types of action performed by Paola in the scenarios were: grasping, pointing, gazing, and head movements. As in the original IST, each item included two sentences, in addition to the scenario. One of the sentences was always explaining Paola's behaviour referring to the design stance (i.e., mechanistic explanation), whereas the other was always describing Paola's behaviour referring to mental states (i.e., mentalistic explanation). We kept the human agent's emotional expression constant within and across the scenarios, not to bias towards mentalistic explanations, see below for a complete list of items.



1. Per favore, posiziona lo slider verso la descrizione che consideri più plausibile.Tieni a mente che le descrizioni si riferiscono a tutte e tre le immagini che compongono la sequenza.

Paola vorrebbe tenere questo cubo.

Paola classifica i cubi in base al colore.

Figure A.1. Screenshot from the Human Control Test in Italian.

A.3 Data Analysis and Results

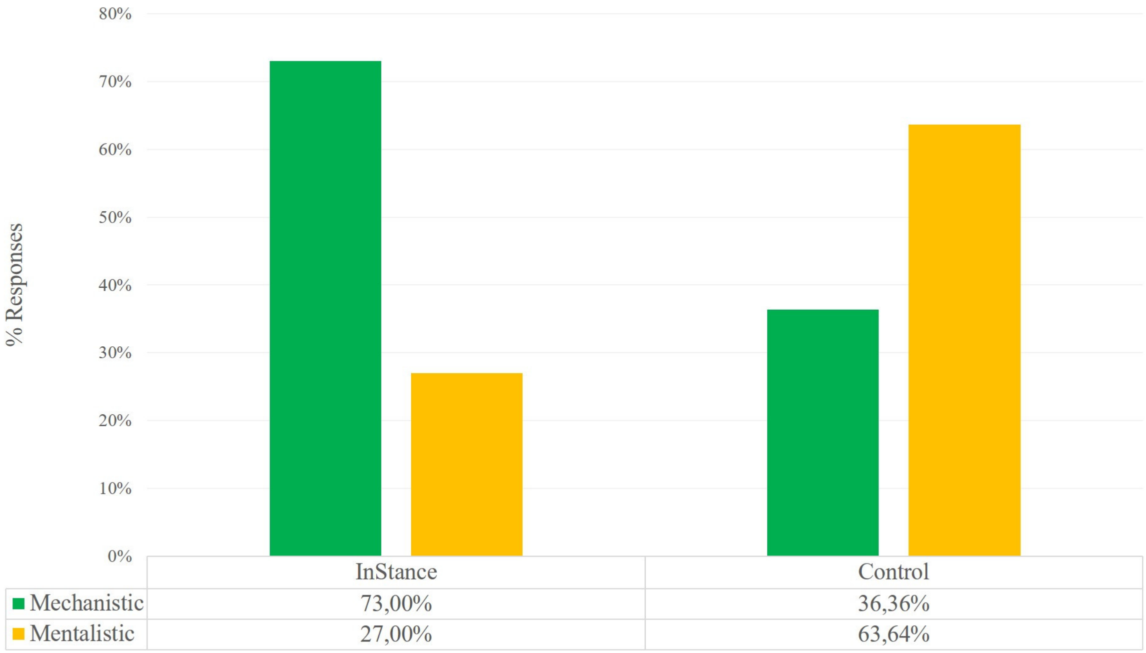
All statistical analyses were performed in R (version 3.4.0, freely available at http://www.rproject.org). Data analysis was conducted in three steps. First, we analysed responses in the HCT as we did for the InStance Test (IST). Second, we selected from our original IST only responses to the same 15 items included in the HCT, and we performed analyses on those. Finally, we compared responses in the two questionnaires (15 items) between groups.

A.3.1 Human control Test (HCT)

For each participant, we calculated the average score. As we did for the IST, we converted the bipolar scale into a 0-100 scale, where 0 corresponded to a completely mechanistic and 100 to a completely mentalistic explanation. The null value of the scale, i.e., the starting position of the slider that was equally distant from both limits, corresponded to 50. Scores under 50 meant the answer was 'mechanistic', scores above 50 meant they were 'mentalistic'. Firstly, we investigated the effect of the mapping by comparing the average score across groups (i.e., not counterbalanced mapping sample vs. counterbalanced mapping sample). Independent sample t-test showed that the average score did not differ across groups, t (220) < 1. Given that no differences were found between mappings, from the total sample of N= 222 participants, we randomly selected 106 respondents, to match the sample size of our original IST. The overall average score for the HCT was 54.62 (with 0 value indicating the most mechanistic score and 100 indicating the most mentalistic score). We tested the distribution of the average score for normality with the Shapiro–Wilk test. Results showed that the average scores were distributed normally, W = 0.99, p > .05. Moreover, in order to check that the average score for the HCT was not the result of random choice, we conducted one-sample t-tests against a critical value of 50 (i.e., the position at which the slider was equally distant from both statements). Results showed that the average score significantly differed from 50, t (105) = 3.67, p <.001). Then, as we did for the IST, we focused only on respondents who were not familiar with robots. From the original sample of N= 222, we selected randomly N= 89 respondents (to match the sample of the non-familiar group in IST), who reported no familiarity with robots. The average score was 54.82, and it was distributed normally, Shapiro–Wilk test: W = 0.97, p = .052. Then, as we did for the IST, we estimated the percentage of participants who attributed 'mechanistic' or 'mentalistic' descriptions according to their average score. Participants who scored below 50 (0 - 50 in our scale) were assigned to the Mechanistic group (N= 32), whereas participants with an average score above 50 (50 – 100) were classified as the Mentalistic group (N= 56). To check whether the percentage of respondents in the Mechanistic and Mentalistic group differed from chance level (i.e., expected frequency of 0.5), we performed a chi-square test. Results revealed that the frequency of participants who scored in HCT "Mechanistic" (36.4 %) and the frequency of participants who scored in HCT "Mentalistic" (63.4 %) were both different from the chance level, χ^2 (1. N= 88) = 51.26, p < .001. In order to compare if the average scores of the two groups (Mechanistic and Mentalistic) significantly differed from the null value of our scale (i.e. 50, which corresponded to the position at which the slider was equally distant from both statements), we ran one-sample t-tests against a critical value of 50 (i.e., the null value of our scale). Results showed that the average score significantly differed from the null value of 50 both for the Mechanistic (M= 42.95, SD = 7.48, t (31) = -5.33, p < .001) and the Mentalistic group (M = 61.69, SD = 9.60, t (55) = 9.11, p < .001).

A.3.2 InStance Test: selected 15 items

From the original IST, we selected responses to the same 15 items as those included in the HCT. The average InStance score (ISS) was 38.32, and it was distributed normally, Shapiro–Wilk test: W = 0.98, p > .05. In order to compare if the ISS significantly differed from a completely mechanistic bias, we conducted one-sample t-tests against a critical value of 0 (i.e. the value corresponding to a mechanistic bond). Results showed that the average ISS significantly differed from 0, t (105) = 24.52, p < .001). For respondents who were not familiar with robots (N = 89), the average score was 38.67, and it was distributed normally, Shapiro– Wilk test: W = 0.98, p > .05. Then, as we did for the full version of the IST, we estimated the percentage of participants who attributed 'mechanistic' or 'mentalistic' descriptions according to their average score. Participants who scored below 50 (0 - 50 in our scale) were assigned to the Mechanistic group (N= 65), whereas participants with an average score above 50 (50 – 100) were classified as the Mentalistic group (N= 24). To check whether the percentage of respondents in the Mechanistic and Mentalistic categories differed from chance level (i.e., expected frequency of 0.5), we conducted a chi-square test. Results revealed that the frequency of participants who scored in IST "Mechanistic" (73.0 %) and the frequency of participants who scored in IST "Mentalistic" (27.0 %) were both different from the chance level, χ^2(1. N =89) = 18.89, p < .001. In order to compare if the mean average scores of the two groups (Mechanistic and Mentalistic) significantly differed from the null value of our scale (i.e. 50, which corresponded to the position at which the slider was equally distant from both the two statements), we conducted one-sample t-tests against a critical value of 50 (i.e., the null value of our scale). Results showed that the average score significantly differed from the null value of 50 both for the Mechanistic (M = 31.19 SD = 11.15, t (64) = -13.61, p < .001) and the Mentalistic group (M = 58.96, SD = 6.08, t (23) = 7.22, p < .001).



	InStance	Control
🟩 Mechanistic	73,00%	36,36%
🟨 Mentalistic	27,00%	63,64%

Figure A.2. Percentage of Mechanistic (green bars) and Mentalistic (yellow bars) respondents to the reduced version of the IST on the left, and to the HCT on the right, for participants who were not familiar with robots (N=89).

A.3.3 Comparison between InStance and Human control Tests

Average scores for IST and HCT for 15 items are shown in Table A.2. In order to compare if the average score for the HCT statistically differed from the ISS, we conducted an independentsamples t-test. Results showed that the ISS for the shorter version of our InStance Test

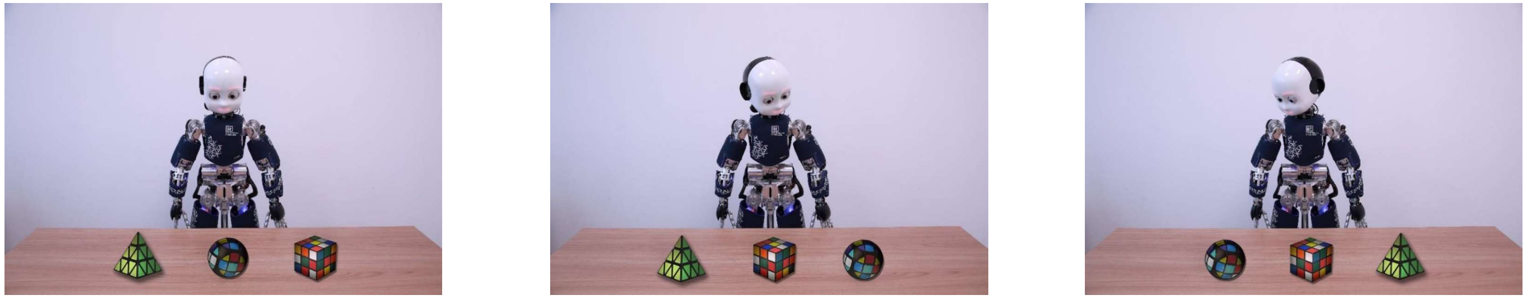
(M= 38.32, SD= 16.09) significantly from the average score of the HCT (M= 54.62, SD= 12.95), t(210) = 8.12, p < .001, d = 1.12. Similar results were found when considering only respondents who were not familiar with robots (N= 89), t(176) = 7.5, p < .001, d = 1.12.

Item IST	N° humans in the scenario	Mean	SD	Item HCT	N° humans in the scenario	Mean	SD
3	1	50.28	41.71	3	1	52.05	41.61
4	1	31.51	35.74	4	1	51.47	47.90
8	0	22.84	30.21	8	0	42	40.90
10	1	32.35	37.37	10	1	45.61	40.99
11	1	66.05	39.47	11	1	74.92	36.36
13	0	42.65	40.26	13	0	56.47	42.87
16	0	33.28	36.97	16	0	51.97	41.55
18	1	38.09	38.33	18	1	66.82	37.46
21	2	24.42	33.80	21	2	35.99	42.48
22	0	34.32	37.02	22	0	59.30	41.65
23	1	52.42	42.11	23	1	78.22	33.44
25	1	78.16	30.59	25	1	72.58	37.44
28	0	10.07	20.62	28	0	43.53	45.63
31	0	11.97	25.22	31	0	53.76	44.95
33	0	46.40	41.53	33	0	34.33	40.75

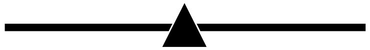
Table A.2. Average score and standard deviation for each item of the IST and HCT ($N = 106$).

The InStance Questionnaire
Supplementary material 2 ENGLISH

Item 0 (Example)



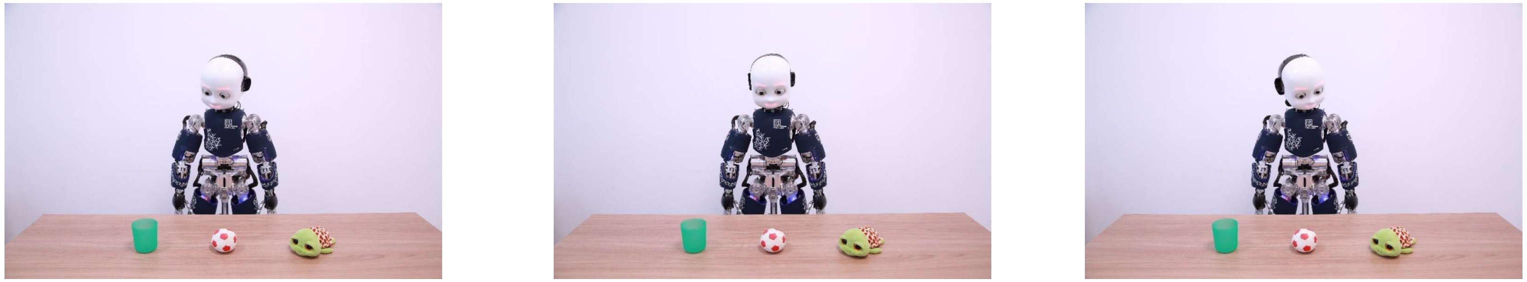
iCub categorizes
objects by their shapes.

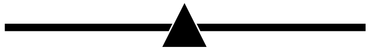


iCub likes round
objects.

M= 55.08; SD: 44.39

Item 1



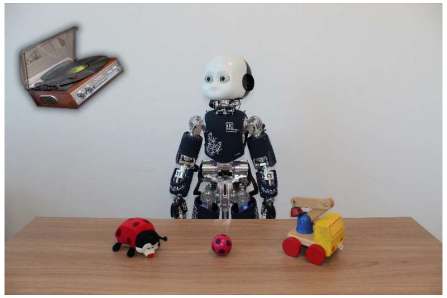


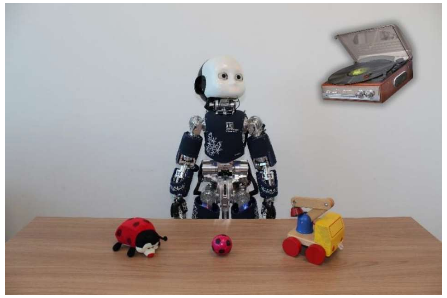
iCub is scanning
the environment.

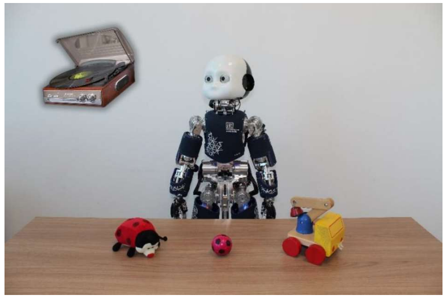
iCub is interested in
these objects.

M= 35.25; SD: 39.72

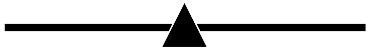
Item 2







iCub looks at
objects at eye level.

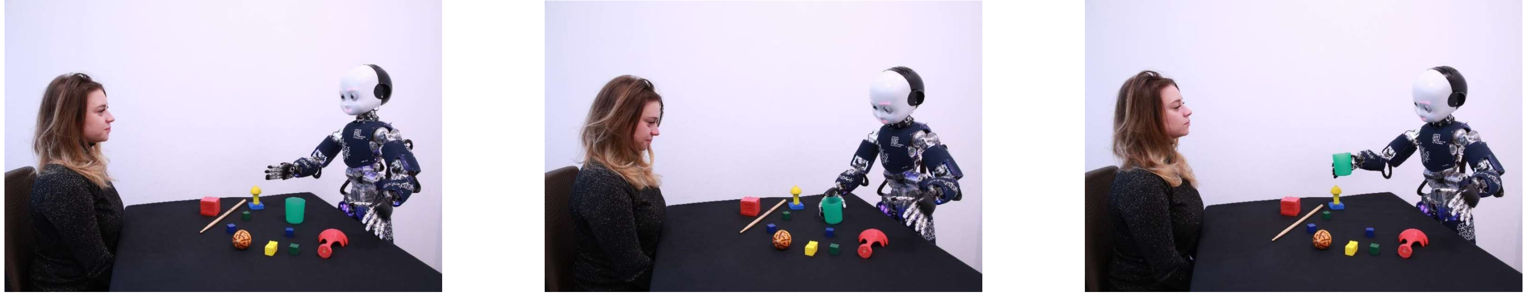


iCub is surprised to see
the object in the air.

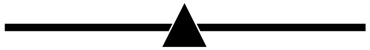
M= 46.37; SD: 41.91

4

Item 3



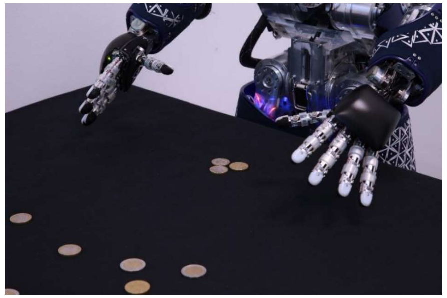
iCub grasps cylindrical
objects best.

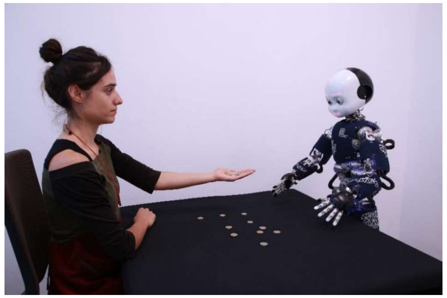


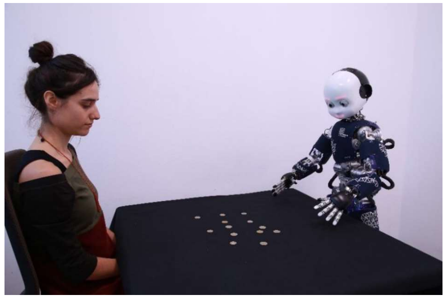
iCub believes that the girl
likes the cup.

M= 50.28; SD: 41.71

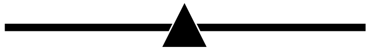
Item 4







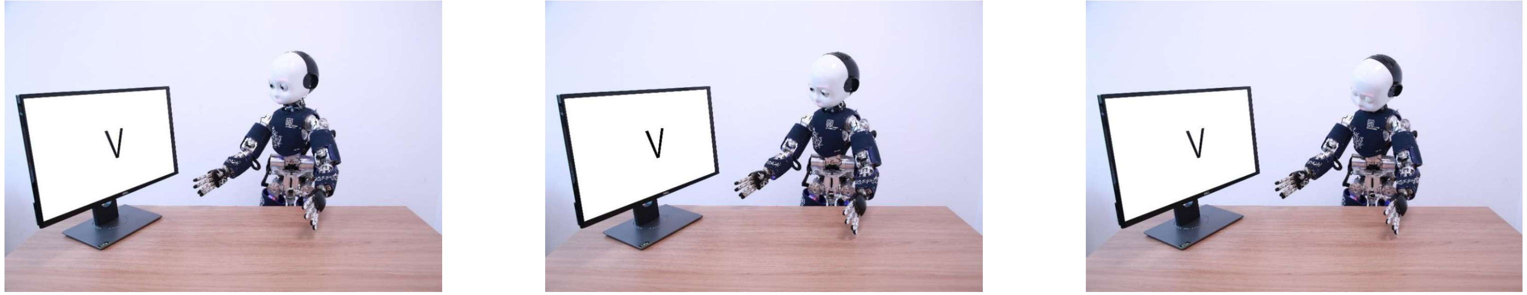
iCub is calculating the
number coins.



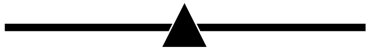
iCub decided to give
back the money.

M= 31.51; SD: 35.74

Item 5



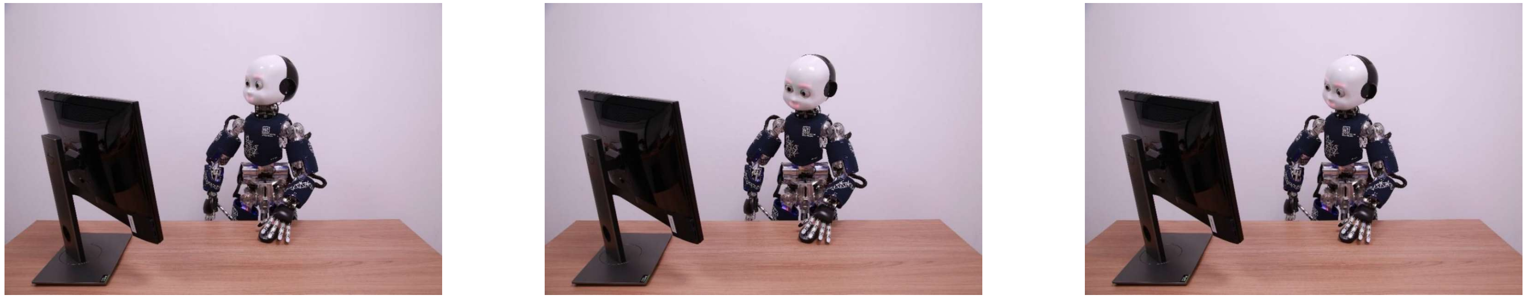
iCub is
turning off.



iCub is
bored.

M= 31.45; SD: 38.51

Item 6

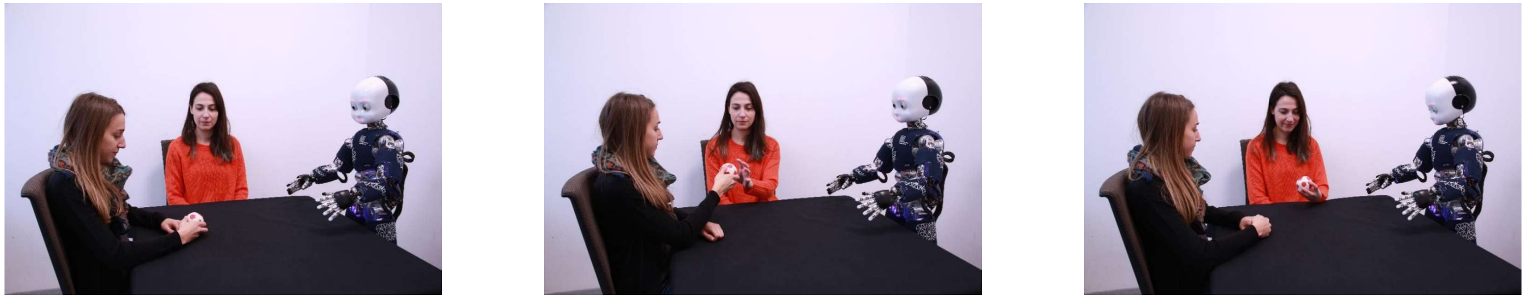


iCub is tracking the mouse cursor.

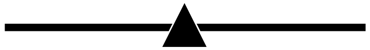
iCub finds the content on the screen interesting.

M= 43.46; SD: 40.89

Item 7



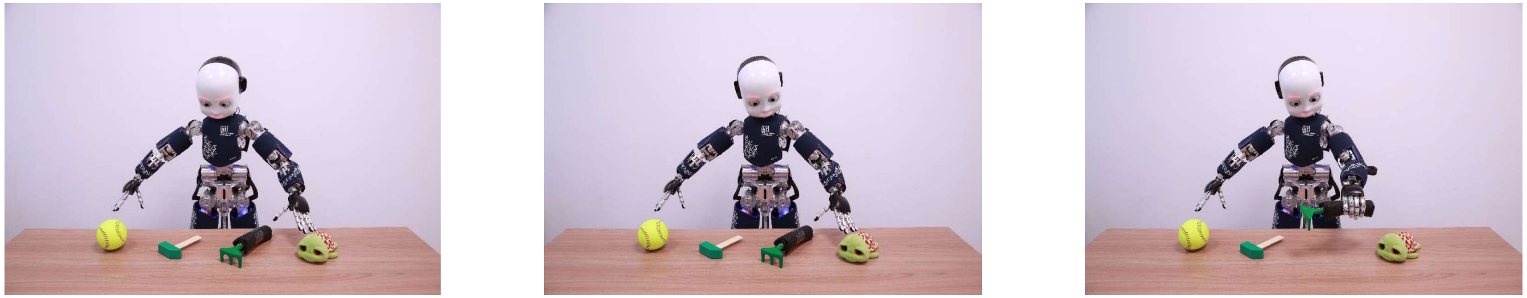
iCub tracks the
ball’s position.



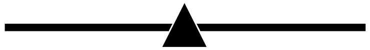
iCub can’t wait to
receive the ball.

M= 33.27; SD: 38.11

Item 8



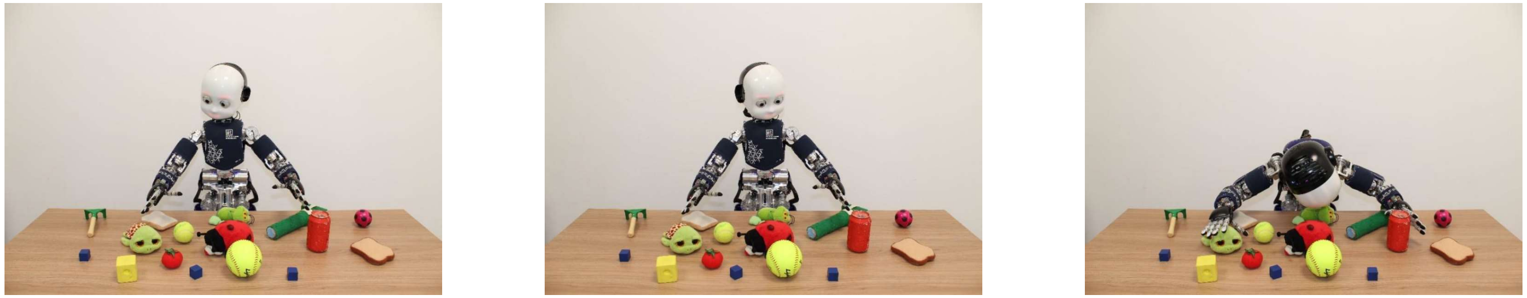
iCub adjusts the force to
the weight of the object.



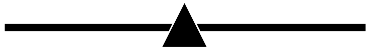
iCub pretends to be
a gardener.

M= 22.84; SD: 30.21

Item 9



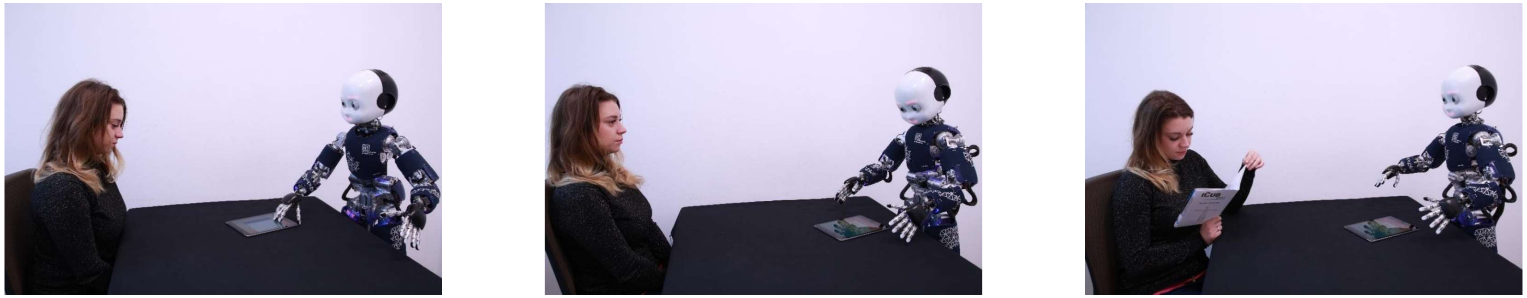
iCub has broken motors.



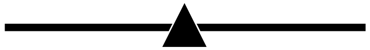
iCub gave up on finding the favorite toy.

M= 55.72; SD: 41.15

Item 10



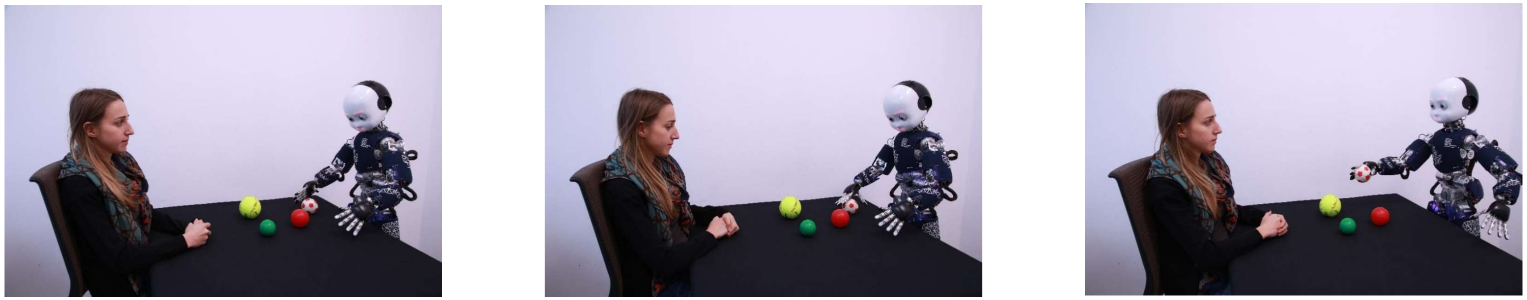
iCub detected differences between the old and the new object in the scene.



iCub wants to read from the book too.

M= 32.35; SD: 37.37

Item 11

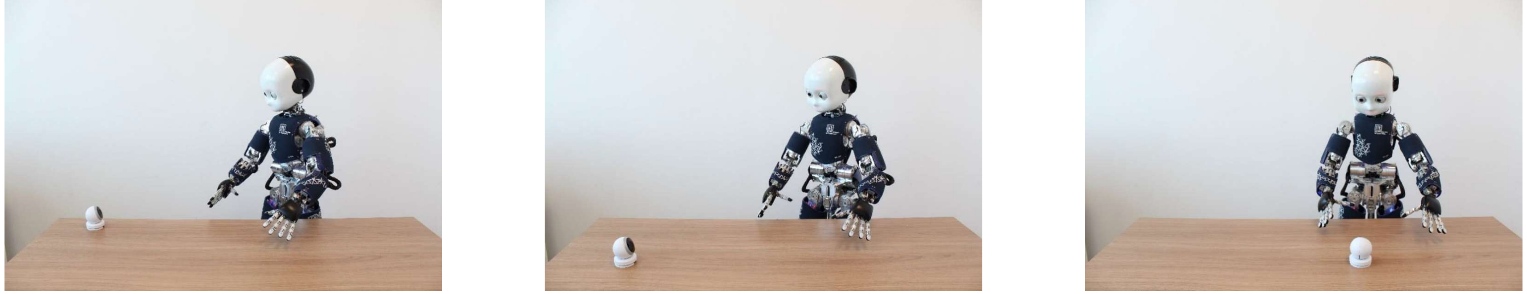


iCub calculates the weight of the balls.

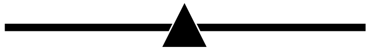
iCub wants to play with the girl.

M= 66.05; SD: 39.47

Item 12



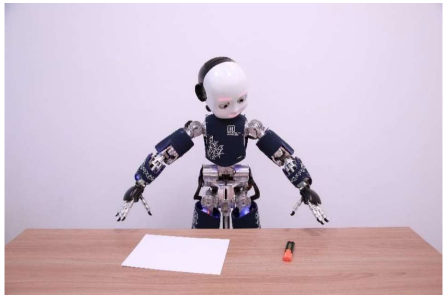
iCub aligns the head
with the webcam.

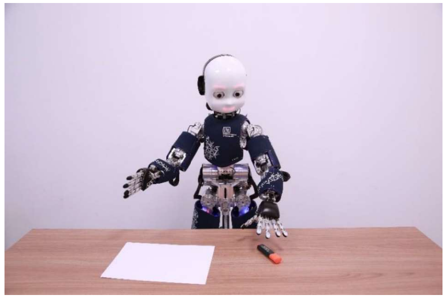


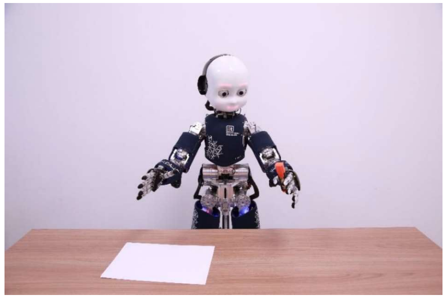
iCub is surprised by
the webcam moving.

M= 28.56; SD: 35.14

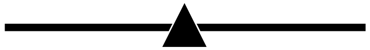
Item 13







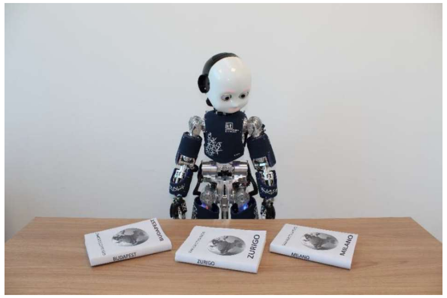
iCub optimizes grip
for small objects.

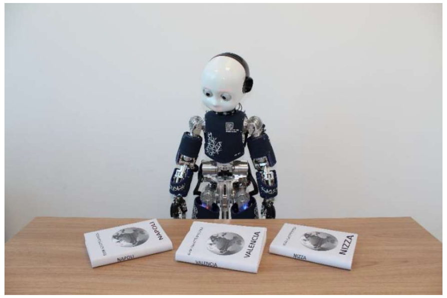


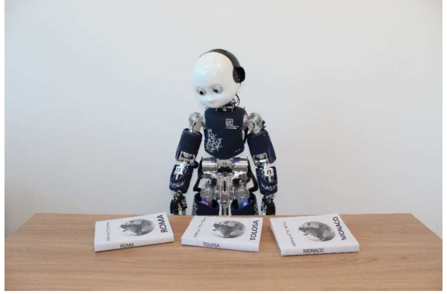
iCub wants to
draw something.

M= 42.65; SD: 40.26

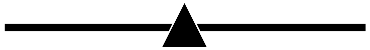
Item 14







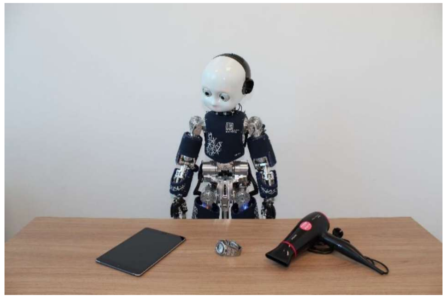
iCub orders Italian
cities alphabetically.



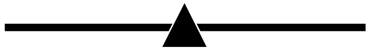
iCub prefers Italian
cities to other cities.

M= 41.85; SD: 40.47

Item 15



iCub easily detects
objects with screens.

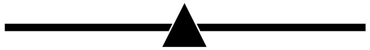


iCub finds digital
technology intriguing.

M= 33.79; SD: 38.01

Item 16

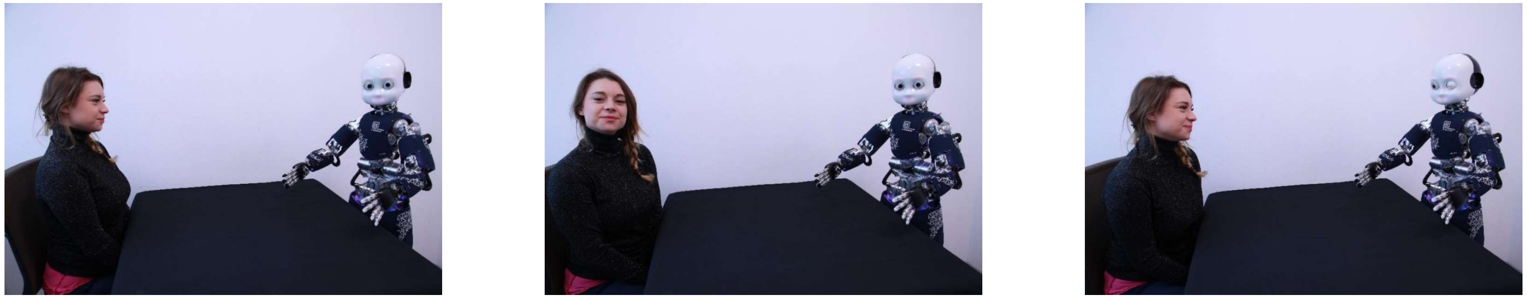
iCub follows the recipe's instructions.



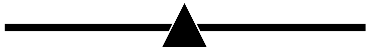
iCub has decided to bake a cake.

M= 33.28; SD: 36.97

Item 17



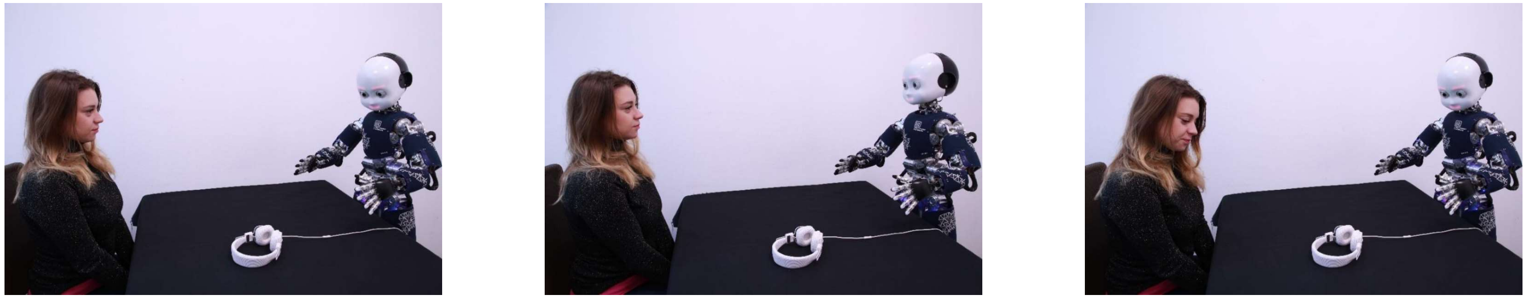
iCub has the
eyelid broken.



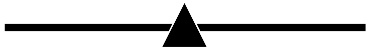
iCub tries to
be funny.

M= 65.04; SD: 38.54

Item 18



iCub measures the distance between the girl and the headphones.

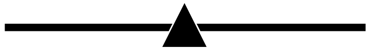


iCub expects that the girl would lend the headphones.

M= 38.09; SD: 38.33

Item 19

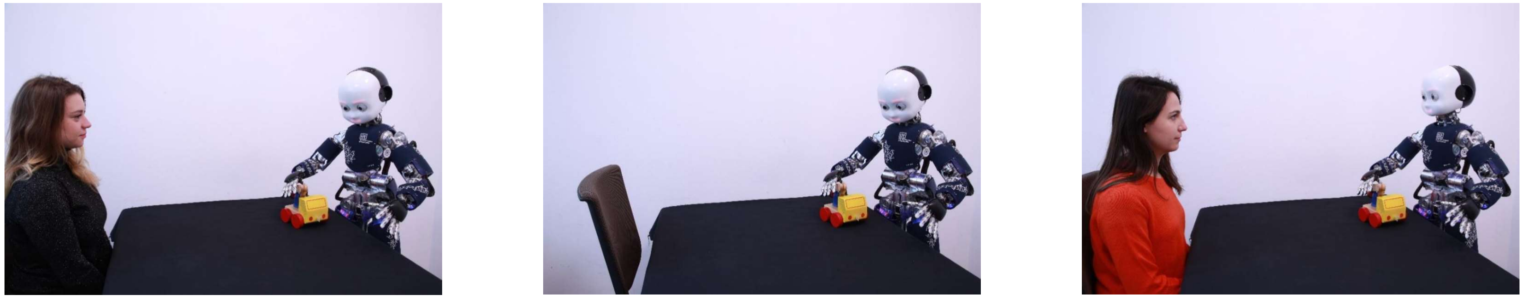
iCub cannot pause the procedure
despite the falling can.



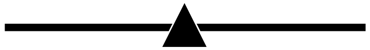
iCub does not intend to
pick up the fallen can

M= 24.71; SD: 31.91

Item 20



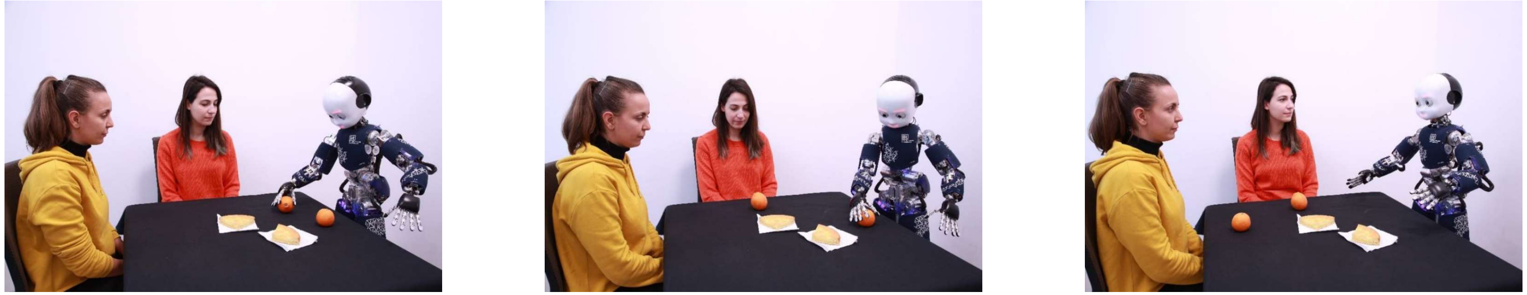
iCub turns the head to bright
colors of the sweater.



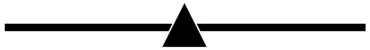
iCub realizes that there is
a new person.

M= 56.58; SD: 39.13

Item 21



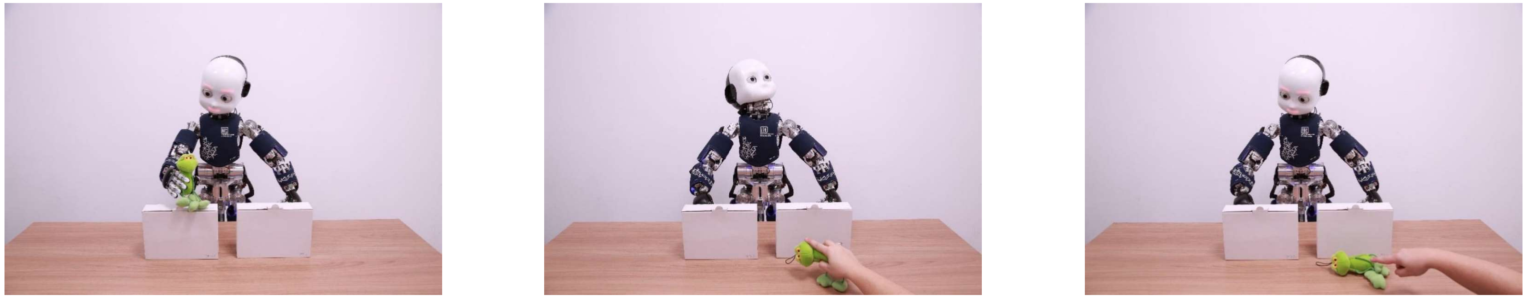
iCub can grasp oranges
easier than cakes.



iCub thinks that the girls
need some healthy food.

M= 24.42; SD: 33.80

Item 22

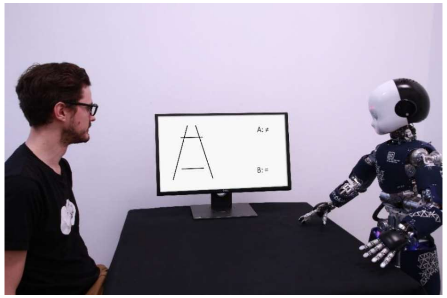


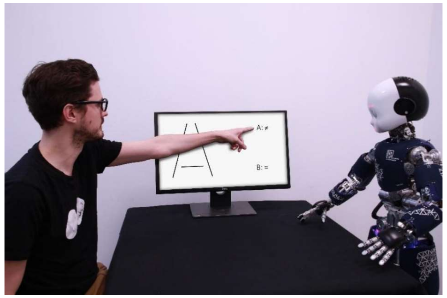
iCub is turning the head to the initial position.

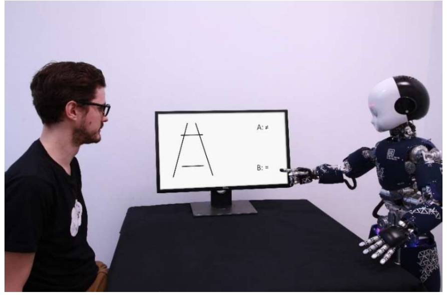
iCub is not interested in the toy anymore.

M= 34.34; SD: 37.02

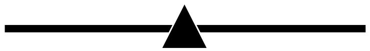
Item 23







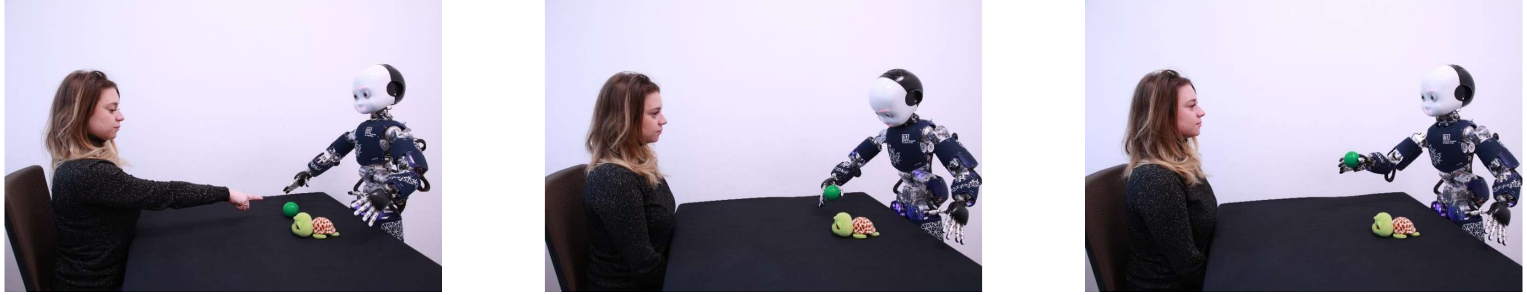
iCub is repeating the
pointing movement.



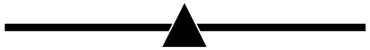
iCub is expressing
an opinion.

M= 52.42; SD: 42.11

Item 24



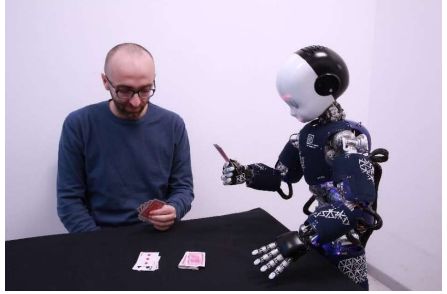
iCub tracked the girl's hand movements.

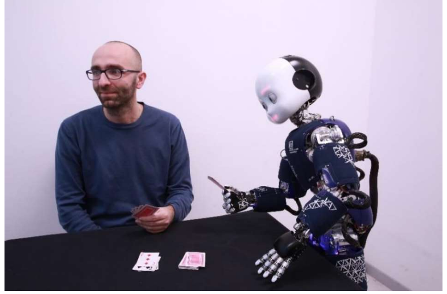


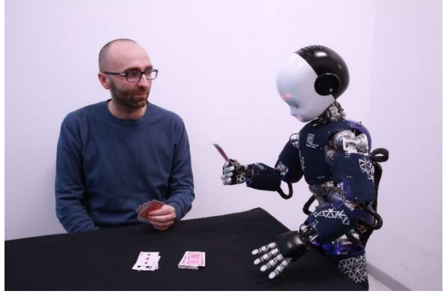
iCub understood that the girl wants the ball.

M= 68.83; SD: 38.41

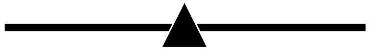
Item 25







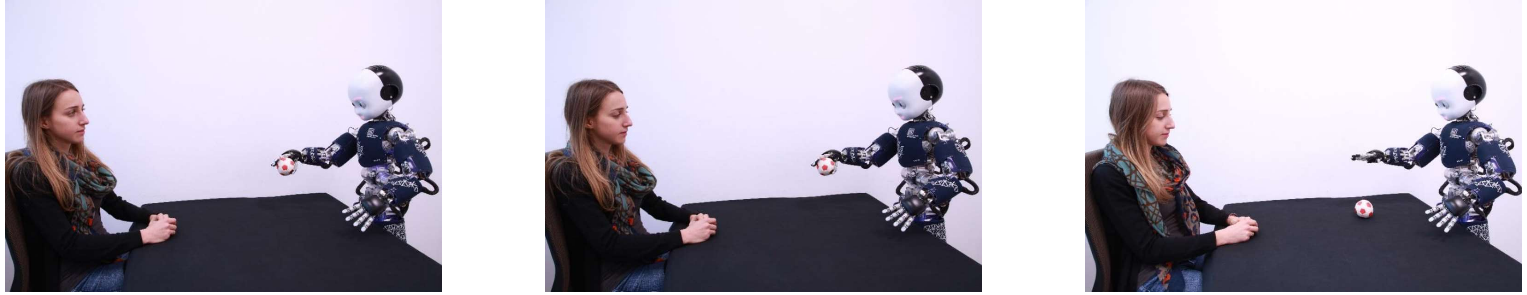
iCub was unbalanced
for a moment.



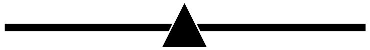
iCub was trying to cheat by
looking at opponent’s cards.

M= 78.16; SD: 30.59

Item 26



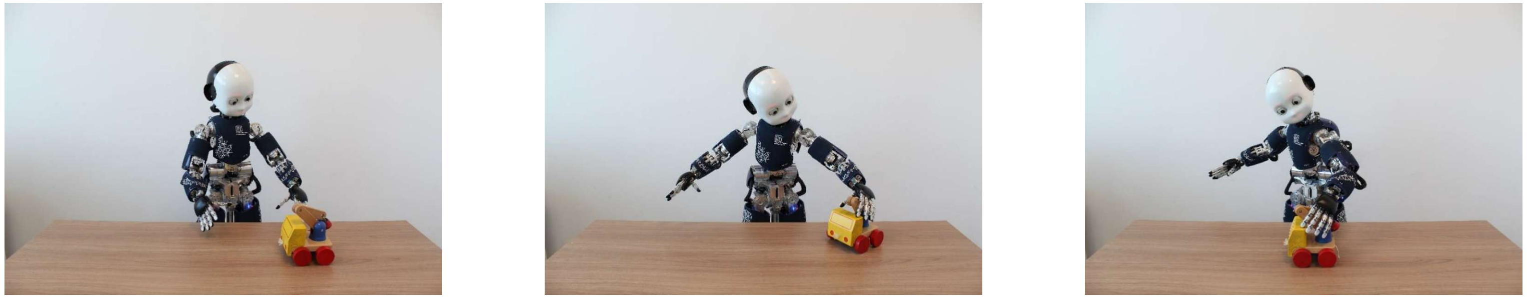
iCub has a failure in
the hand.



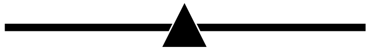
iCub finds it disappointing that
the ball doesn’t bounce back.

M= 57.61; SD: 38.94

Item 27



iCub is optimizing head-arm coordination.



iCub is enjoying playing with the truck.

M= 45.08; SD: 40.81

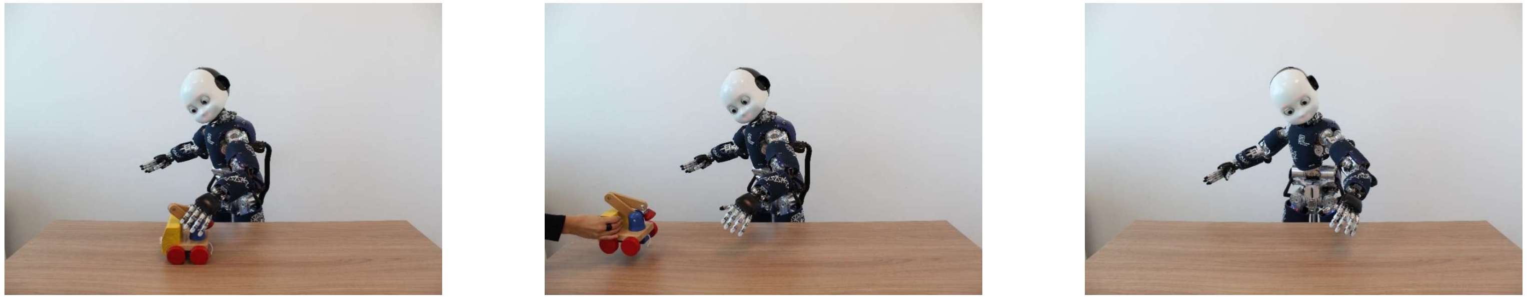
Item 28

iCub classifies cubes
by color.

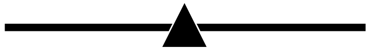
iCub would like to
keep this cube.

M= 10.07; SD: 20.62

Item 29



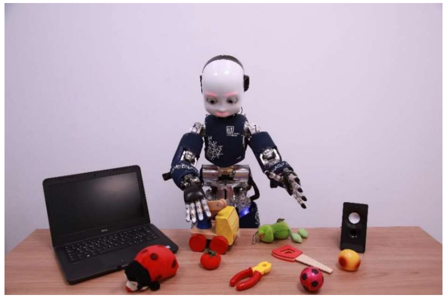
iCub got
stuck.

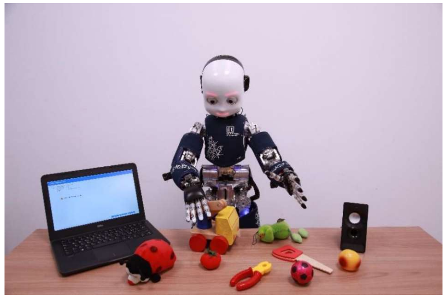


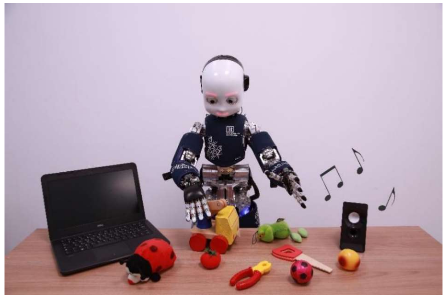
iCub imagines that
the toy is still there.

M= 48.55; SD: 41.41

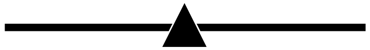
Item 30







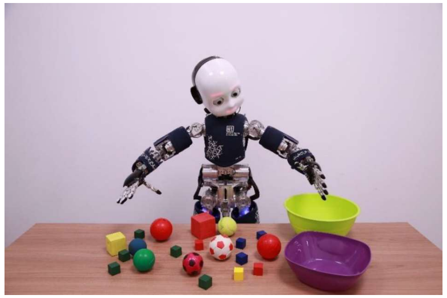
iCub is on stand-by mode.

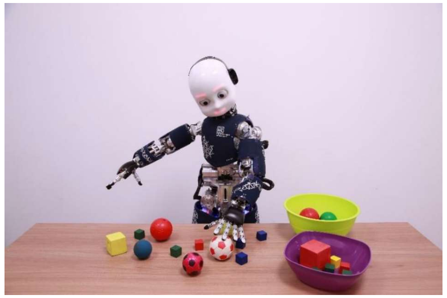


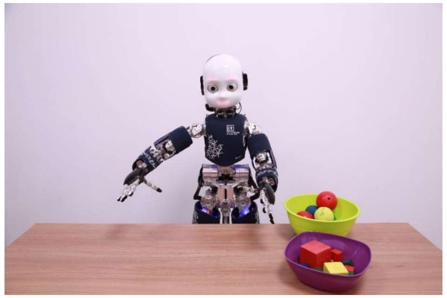
iCub can't decide which toy to play.

M= 25.10; SD: 34.34

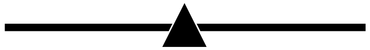
Item 31







iCub places objects in bowls with corresponding shape.

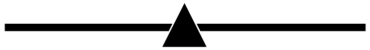


iCub knows that it’s better to clean the table after playing.

M= 11.97; SD: 25.22

Item 32

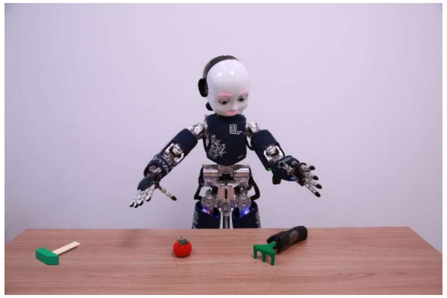
iCub calibrates position of the head
with position of the hand.

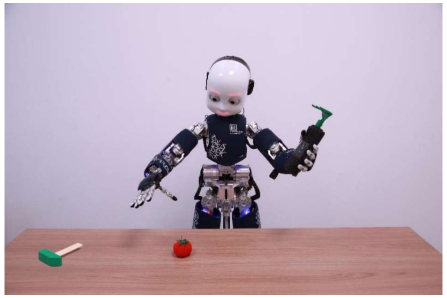


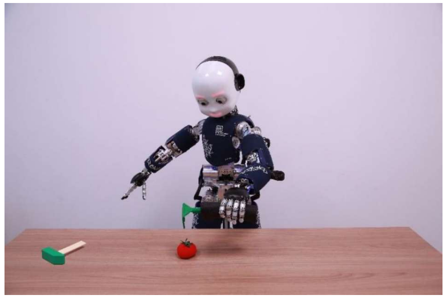
iCub enjoys counting
fingers on each hand.

M= 34.21; SD: 38.33

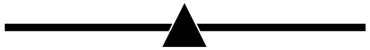
Item 33







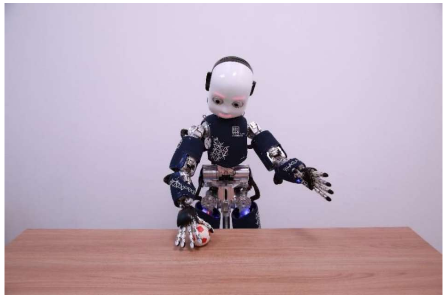
iCub grasped the
closest object.

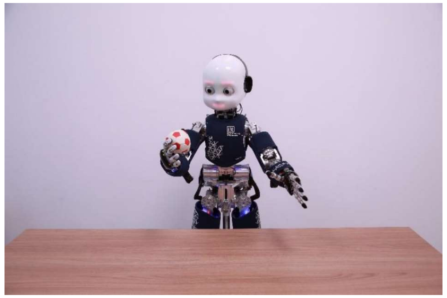


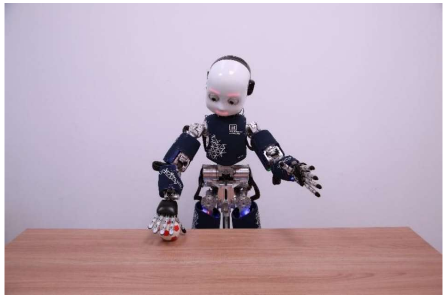
iCub was fascinated
by tool use.

M= 46.40; SD: 41.53

Item 34







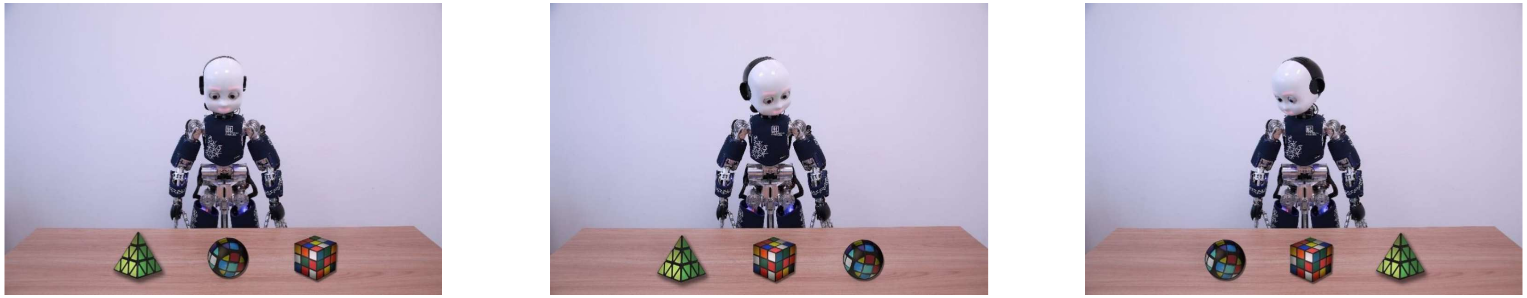
iCub has updated the coordinates
of the ball position.

iCub has decided not to
throw the ball

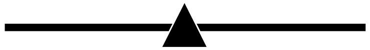
M= 34.40; SD: 36.91

The InStance Questionnaire
Supplementary material 2 ITALIAN

Item 0 (Esempio)



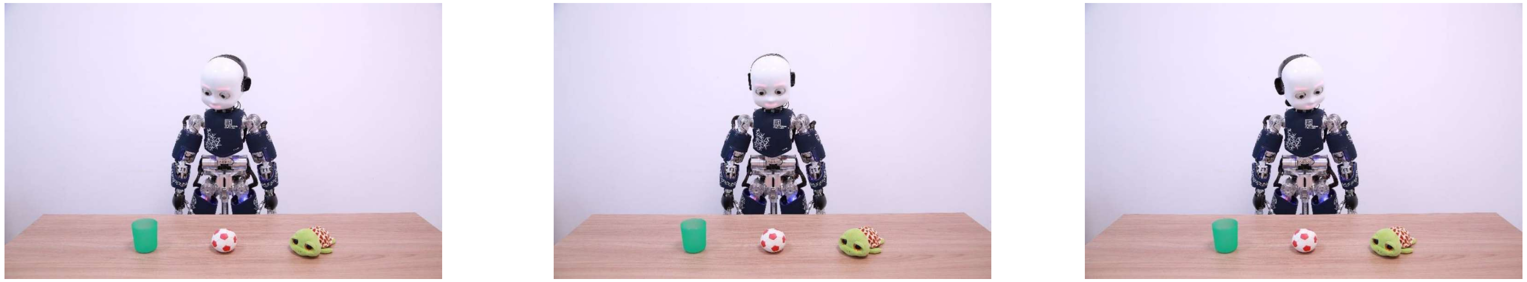
iCub categorizza gli oggetti
in base alla loro forma.



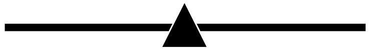
iCub preferisce gli
oggetti sferici.

M= 55.08; SD: 44.39

Item 1



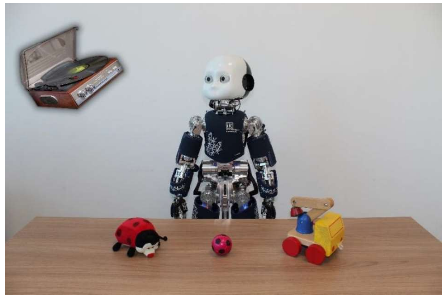
iCub scansiona
l’ambiente.

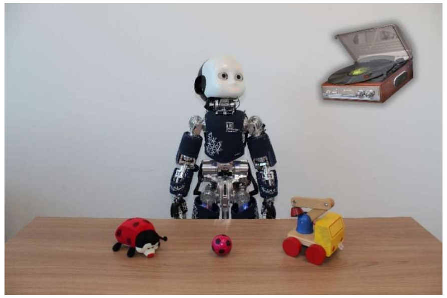


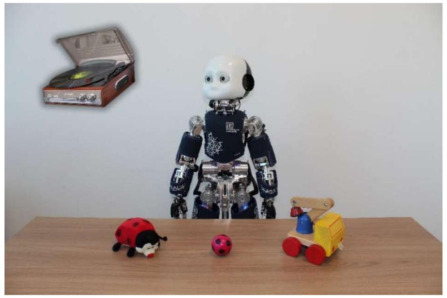
iCub è interessato a
questi oggetti.

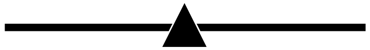
M= 35.25; SD: 39.72

Item 2







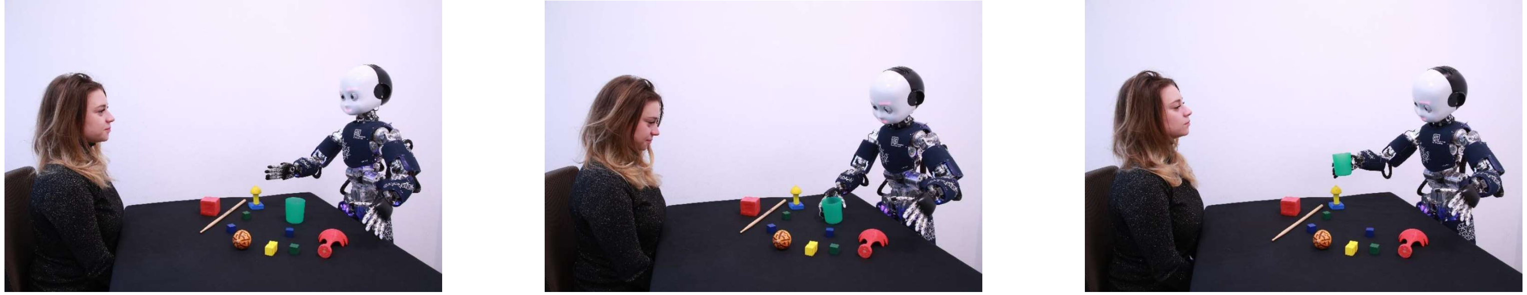


iCub guarda gli oggetti
all’altezza degli occhi.

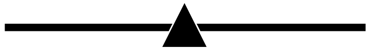
iCub è sorpreso di
vedere l’oggetto in aria.

M= 46.37; SD: 41.91

Item 3



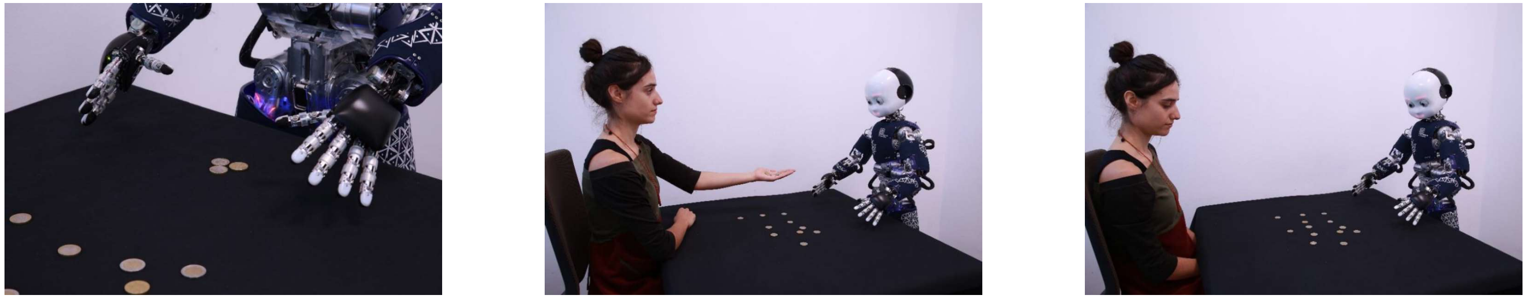
iCub afferra meglio gli oggetti cilindrici.



iCub pensa che alla ragazza piaccia il bicchiere.

M= 50.28; SD: 41.71

Item 4

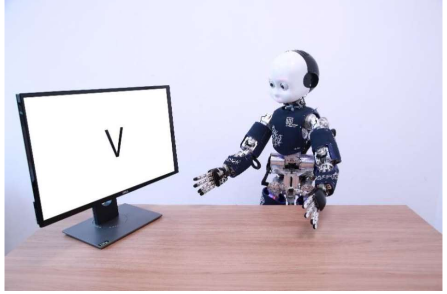


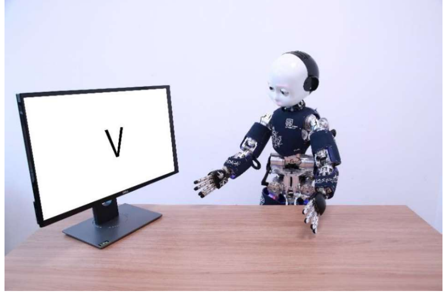
iCub calcola il numero di monete.

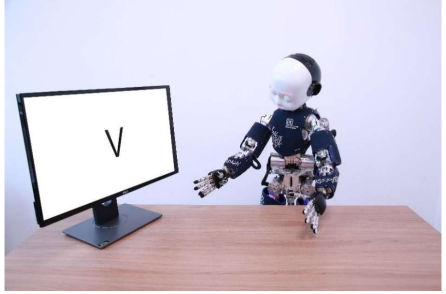
iCub ha deciso di restituire le monete.

M= 31.51; SD: 35.74

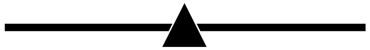
Item 5







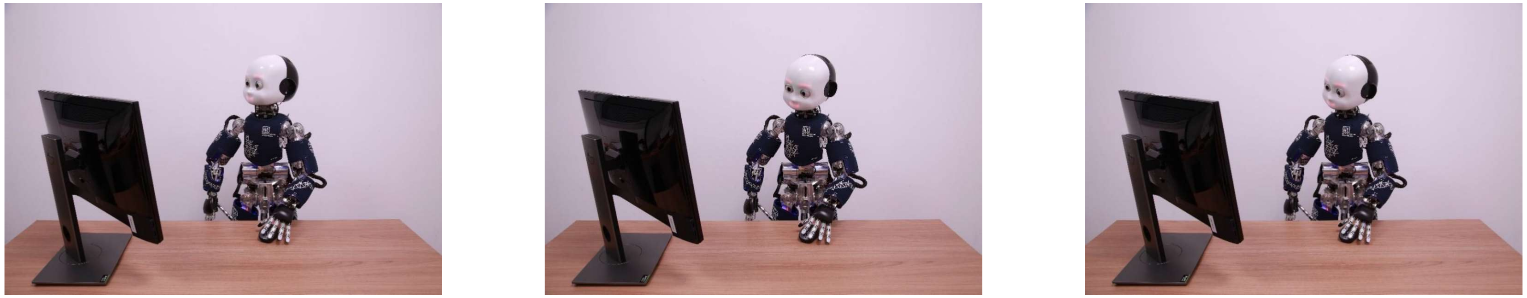
iCub è in fase di
spegnimento.



iCub è
annoiato.

M= 31.45; SD: 38.51

Item 6

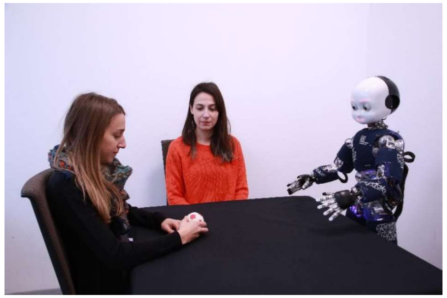


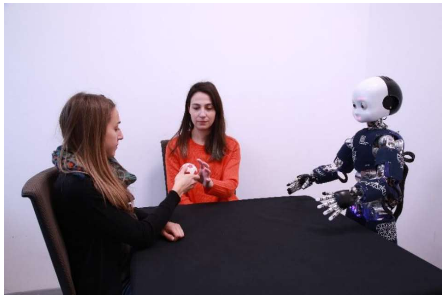
iCub traccia il cursore
del mouse.

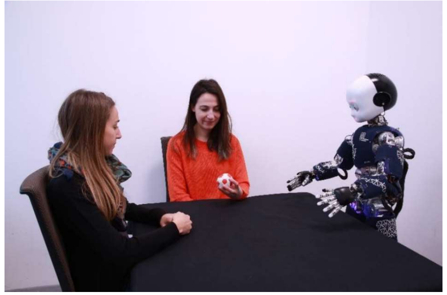
iCub è interessato a quello
che c’è sullo schermo.

M= 43.46; SD: 40.89

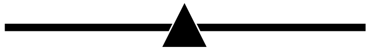
Item 7







iCub traccia la
posizione della palla.

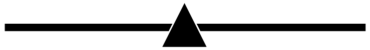


iCub non vede l’ora di
ricevere la palla.

M= 33.27; SD: 38.11

Item 8

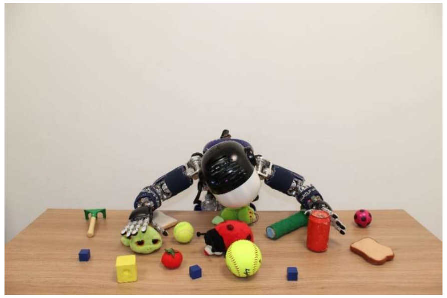
iCub regola la forza in base
al peso dell’oggetto.



iCub finge di essere
un giardiniere.

M= 22.84; SD: 30.21

Item 9

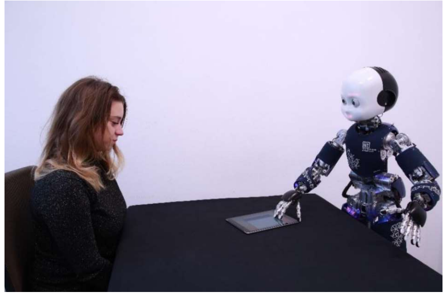


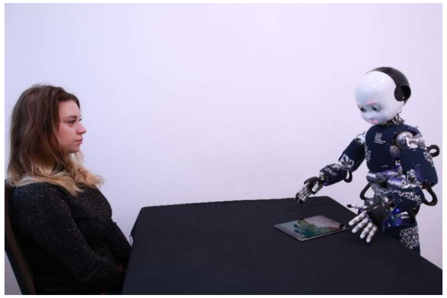
iCub ha i motori
rotti.

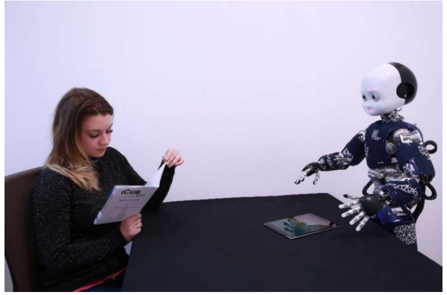
iCub è disperato perché non trova
il suo giocattolo preferito.

M= 55.72; SD: 41.15

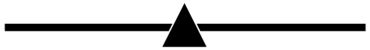
Item 10







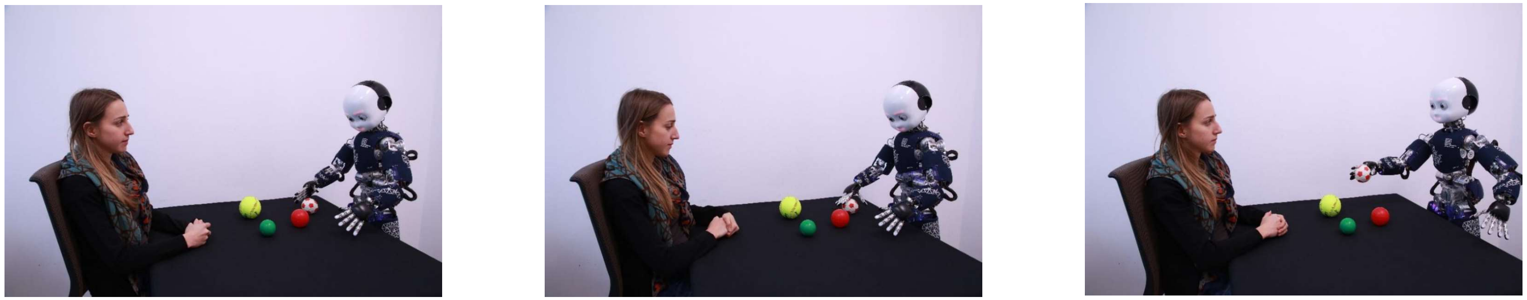
iCub rileva differenze tra il vecchio e il nuovo oggetto in scena.



iCub vuole leggere il libro.

M= 32.35; SD: 37.37

Item 11

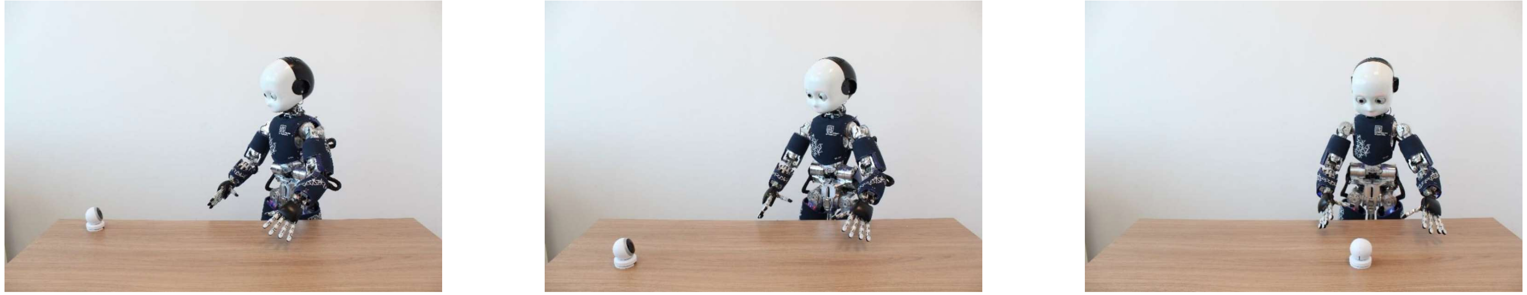


iCub calcola il peso
delle palle.

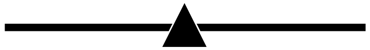
iCub vuole giocare
con la ragazza.

M= 66.05; SD: 39.47

Item 12



iCub allinea la testa con
la webcam.

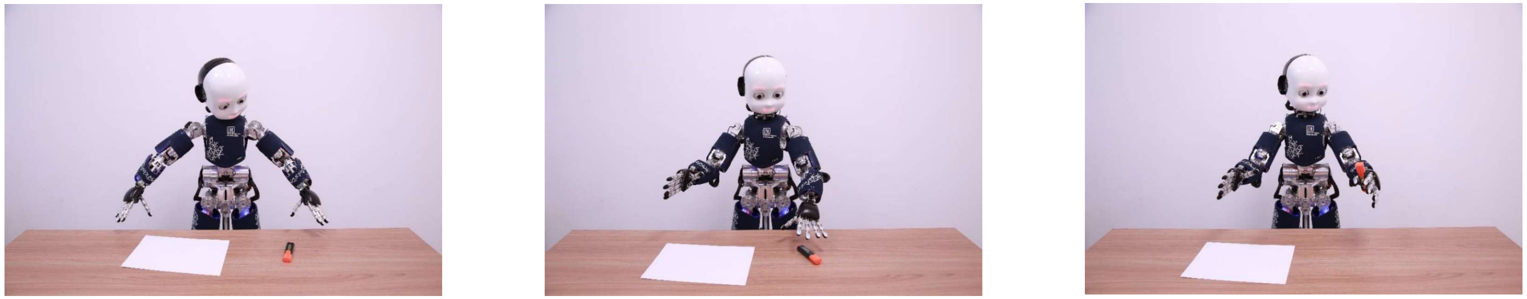


iCub è sorpreso dalla
webcam che si muove.

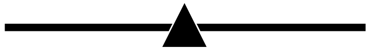
M= 28.56; SD: 35.14

50

Item 13



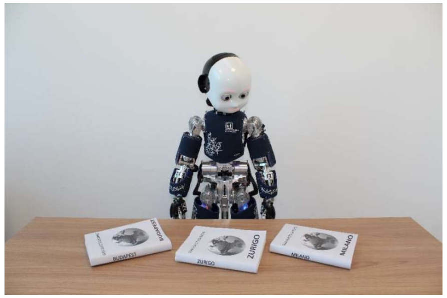
iCub ottimizza la presa
gli oggetti piccoli.

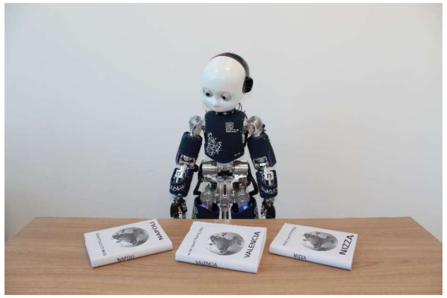


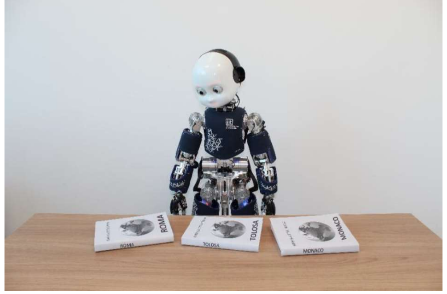
iCub vuole disegnare
qualcosa.

M= 42.65; SD: 40.26

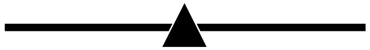
Item 14







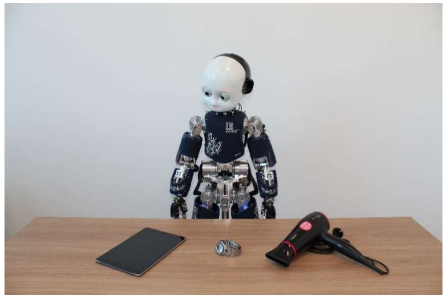
iCub registra le città italiane
in ordine alfabetico.



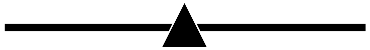
iCub preferisce le
città italiane.

M= 41.85; SD: 40.47

Item 15



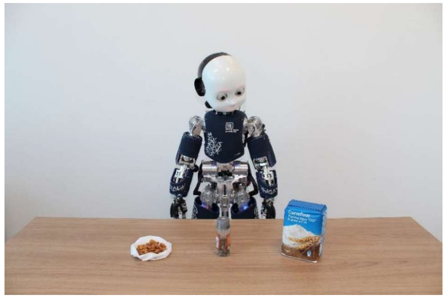
iCub rileva facilmente gli oggetti con uno schermo.

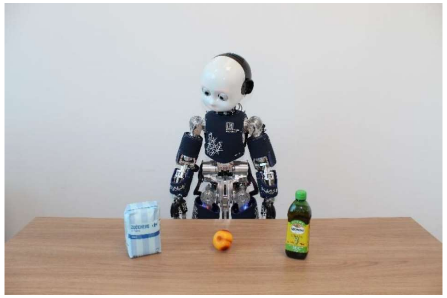


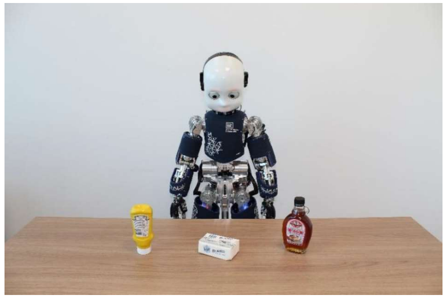
iCub è incuriosito dalla tecnologia digitale.

M= 33.79; SD: 38.01

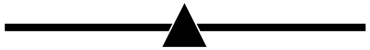
Item 16







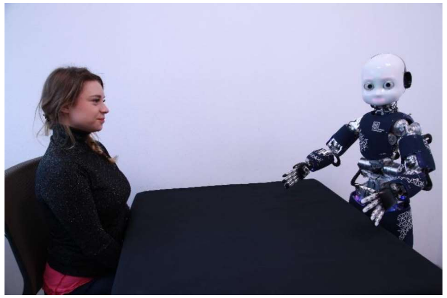
iCub segue le istruzioni
della ricetta.

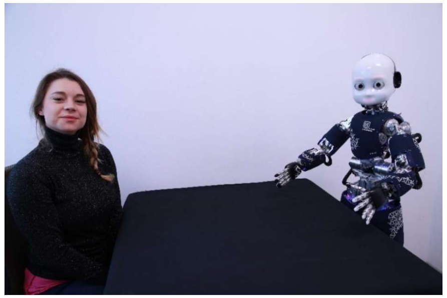


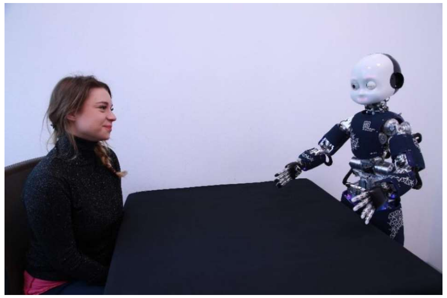
iCub ha deciso di
cucinare una torta.

M= 33.28; SD: 36.97

Item 17





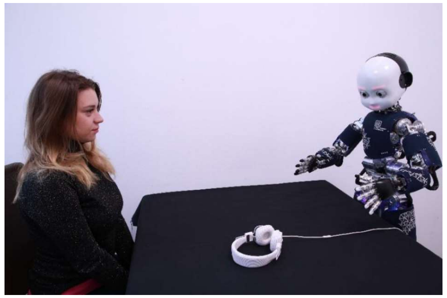


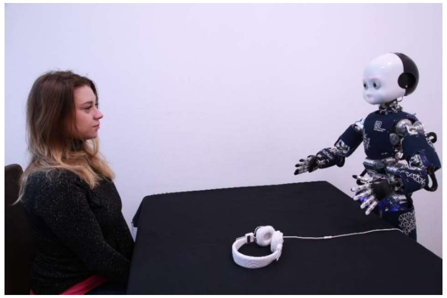
iCub ha i motori
delle palpebre rotti.

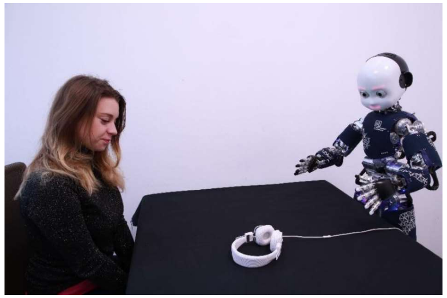
iCub cerca di essere
affascinante.

M= 65.04; SD: 38.54

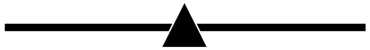
Item 18







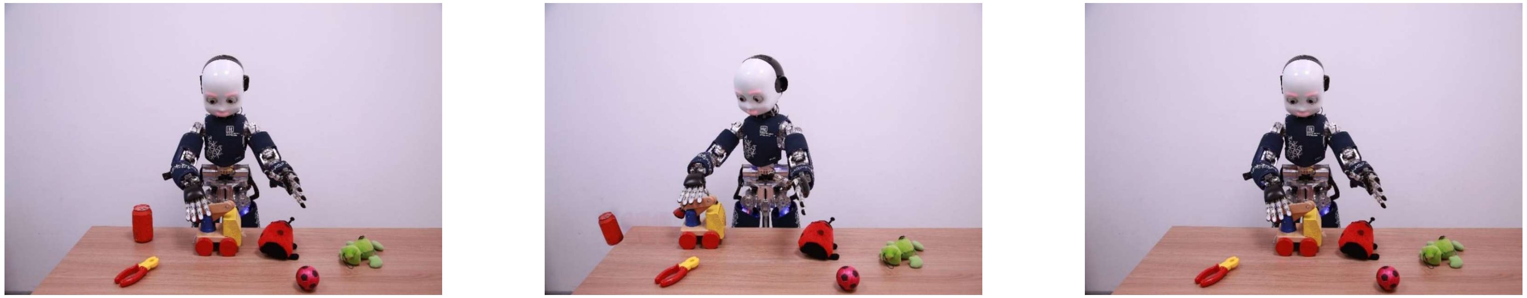
iCub misura la distanza tra la ragazza e le cuffie.



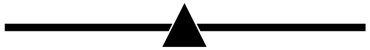
iCub si aspetta che la ragazza dia in prestito le cuffie.

M= 38.09; SD: 38.33

Item 19



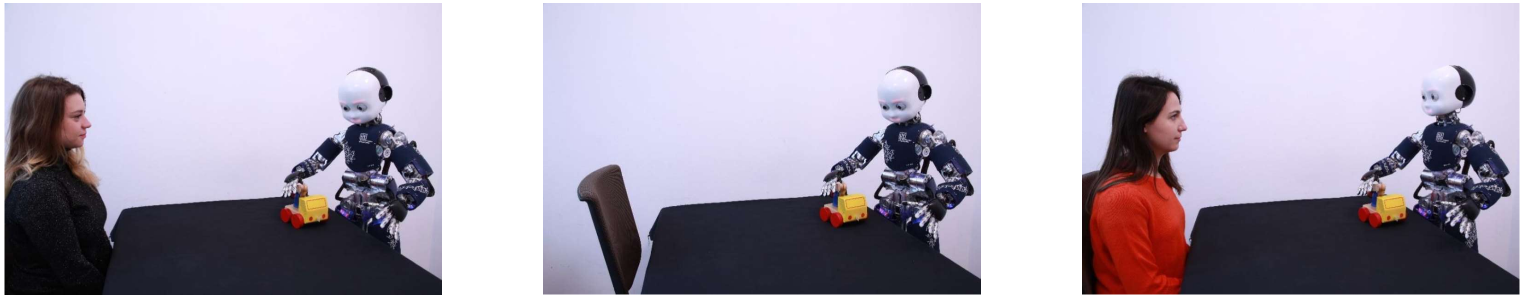
iCub non può bloccare la procedura
nonostante la lattina stia cadendo.



iCub non ha intenzione di
raccogliere la lattina caduta.

M= 24.71; SD: 31.91

Item 20

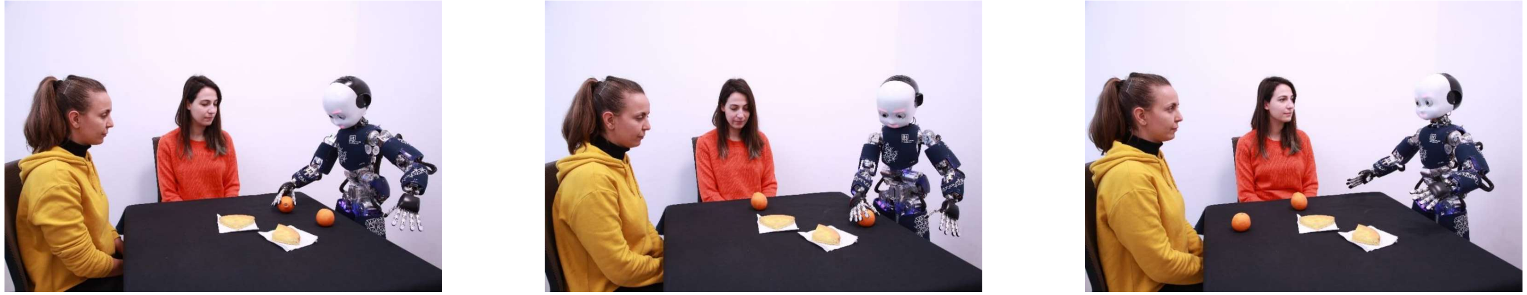


iCub muove la testa verso i
maglioni dai colori luminosi.

iCub è sorpreso che ci sia
una nuova persona.

M= 56.58; SD: 39.13

Item 21

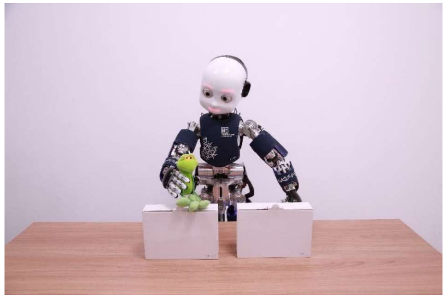


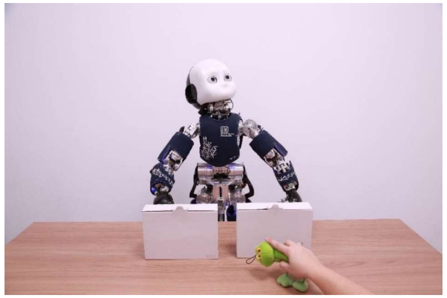
iCub pensa che le ragazze
abbiano bisogno di cibo sano.

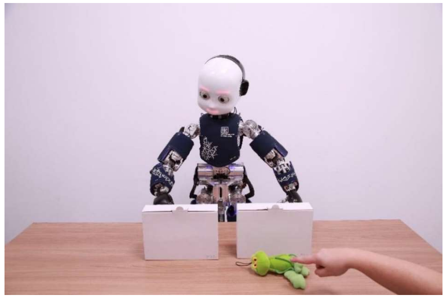
iCub può afferrare le arance
più facilmente della torta.

M= 24.42; SD: 33.80

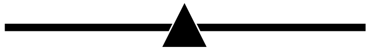
Item 22







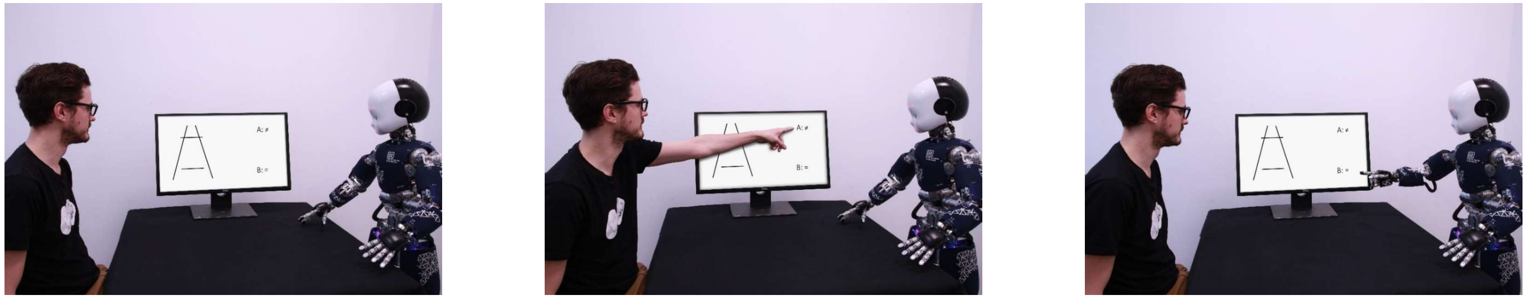
iCub riporta la testa nella
posizione iniziale.



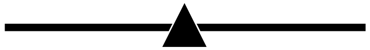
iCub non è più
interessato al giocattolo.

M= 34.34; SD: 37.02

Item 23



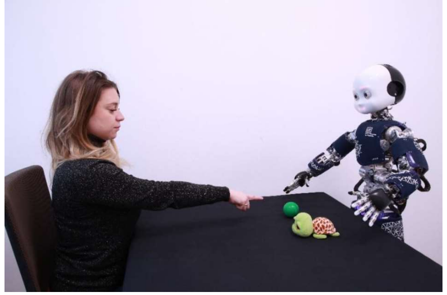
iCub sta ripetendo il
gesto di indicare.

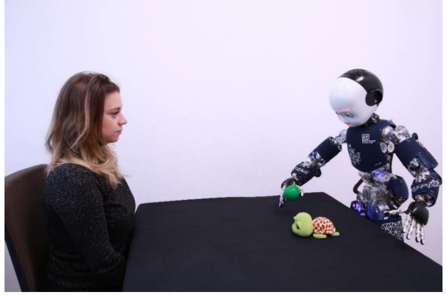


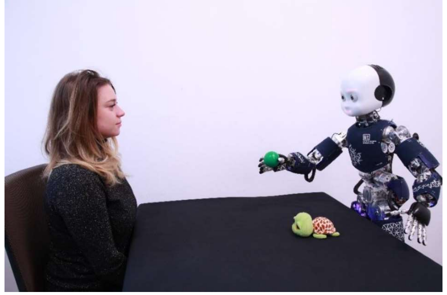
iCub sta esprimendo
un’opinione.

M= 52.42; SD: 42.11

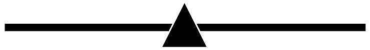
Item 24







iCub traccia i movimenti della mano.

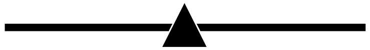


iCub ha capito che la ragazza vuole la palla.

M= 68.83; SD: 38.41

Item 25

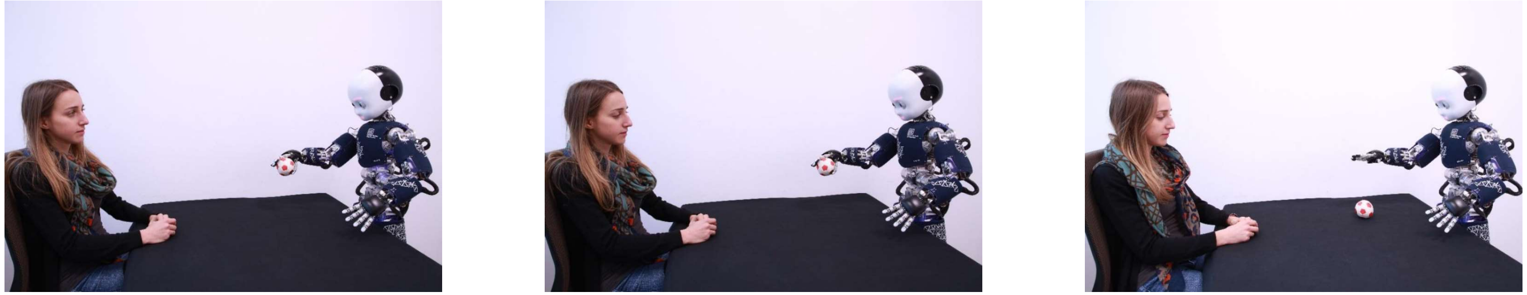
iCub bara guardando le carte
dell’avversario.



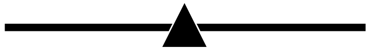
iCub ha perso l'equilibrio
per un momento.

M= 78.16; SD: 30.59

Item 26



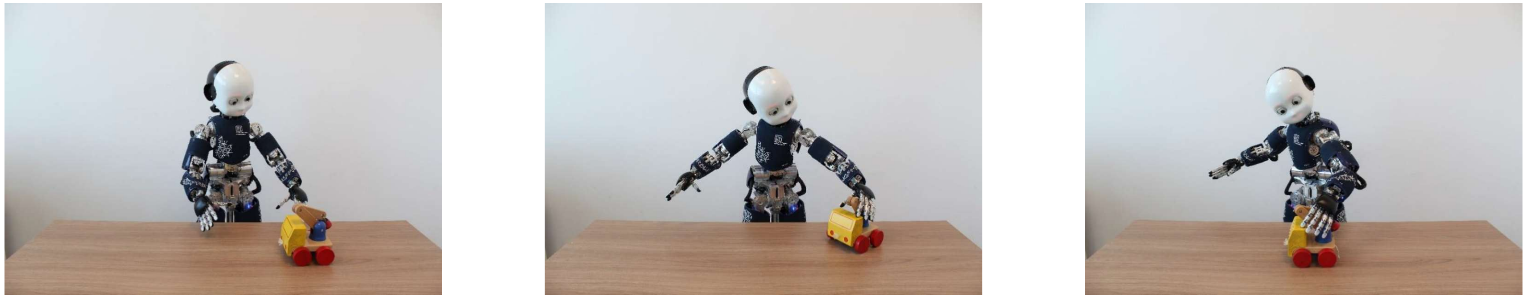
iCub ha un
malfunzionamento alla mano.



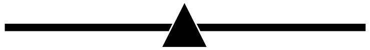
iCub è deluso che la palla non
rimbalzi.

M= 57.61; SD: 38.94

Item 27



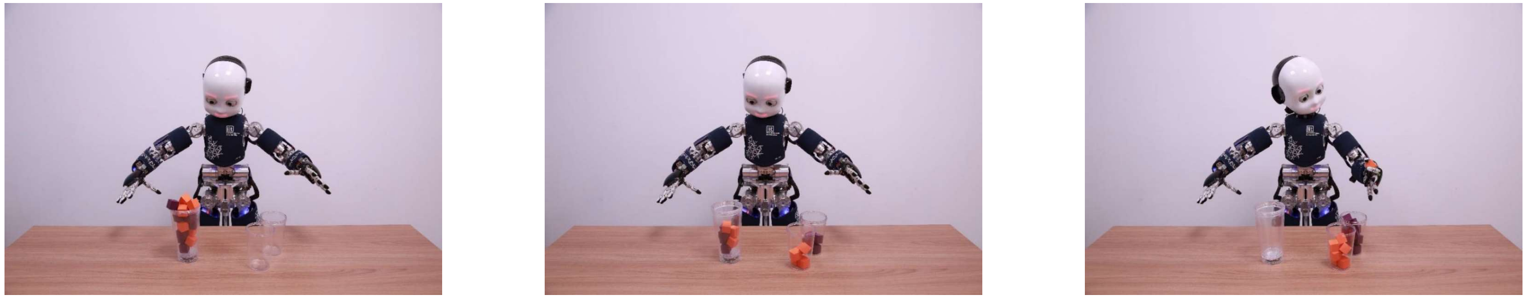
iCub allena la
coordinazione testa-mano.



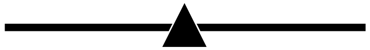
iCub si diverte a giocare
con il camioncino.

M= 45.08; SD: 40.81

Item 28



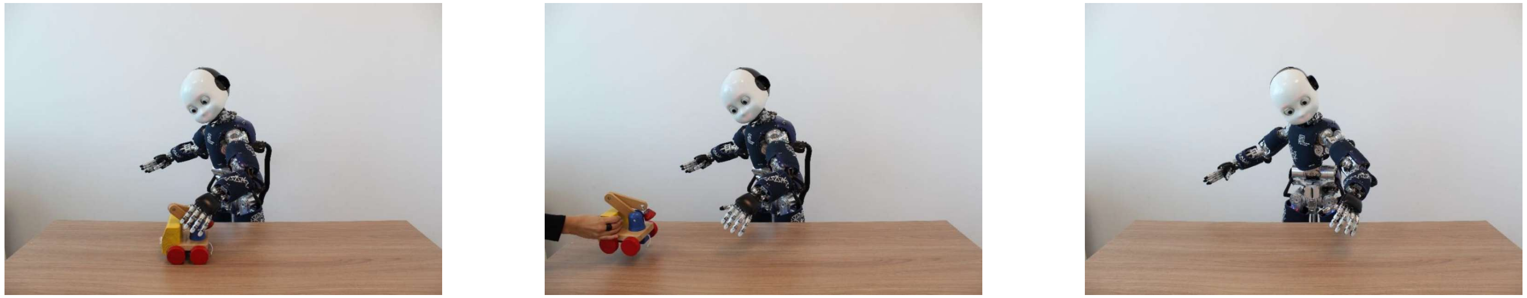
iCub classifica i cubi
in base al colore.



iCub vorrebbe tenere
questo cubo.

M= 10.07; SD: 20.62

Item 29

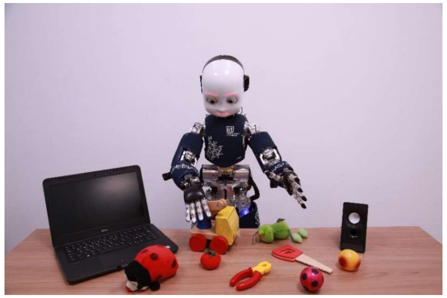


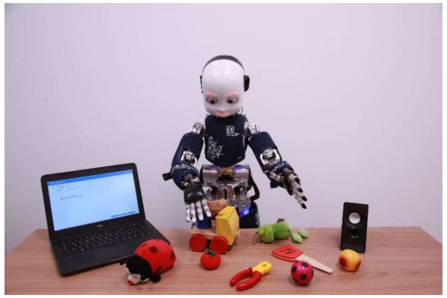
iCub è
bloccato.

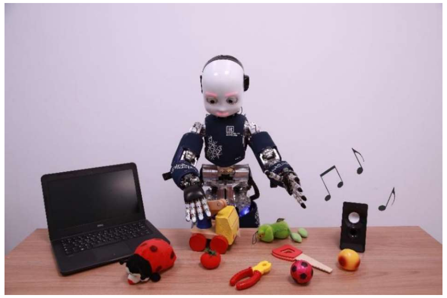
iCub immagina che il
giocattolo sia ancora lì.

M= 48.55; SD: 41.41

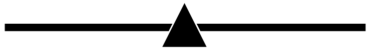
Item 30







iCub è in modalità
stand-by.

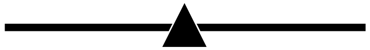


iCub non riesce a scegliere
con quale giocattolo giocare.

M= 25.10; SD: 34.34

Item 31

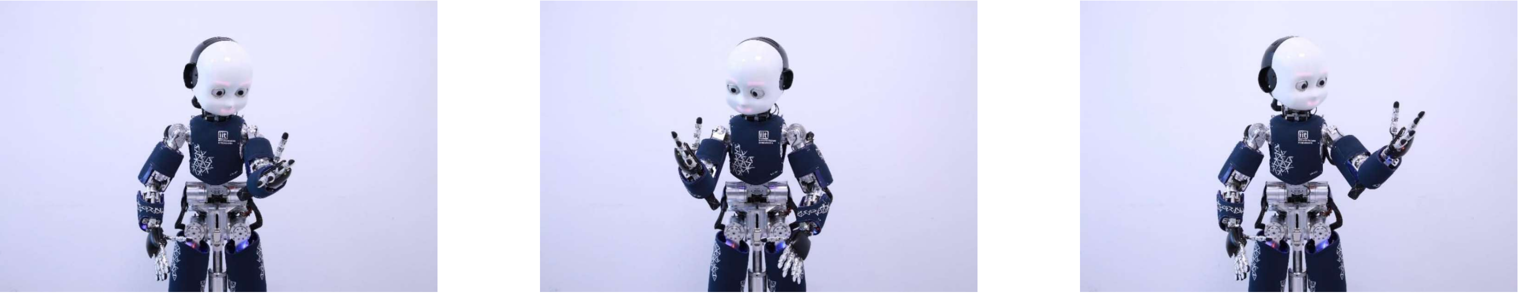
iCub mette gli oggetti nel contenitore con la forma corrispondente.



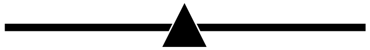
iCub sa che è giusto mettere in ordine il tavolo dopo aver giocato.

M= 11.97; SD: 25.22

Item 32



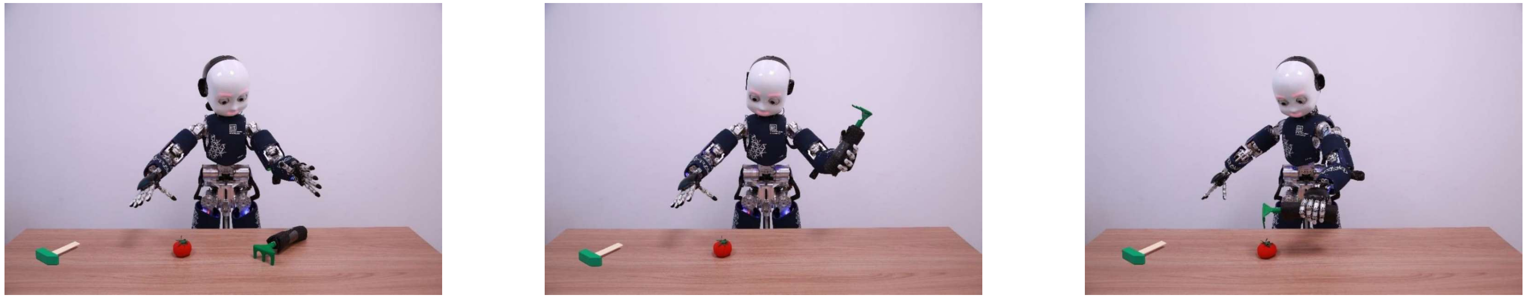
iCub calibra la posizione della testa
con quella della mano.



iCub si diverte contando
le dita su ciascuna mano.

M= 34.21; SD: 38.33

Item 33

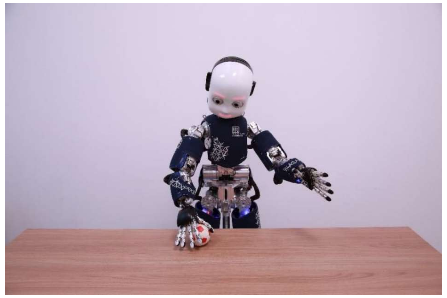


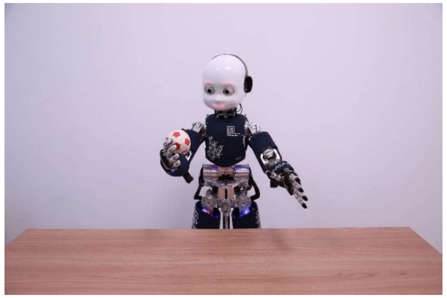
iCub afferra l’oggetto
più vicino.

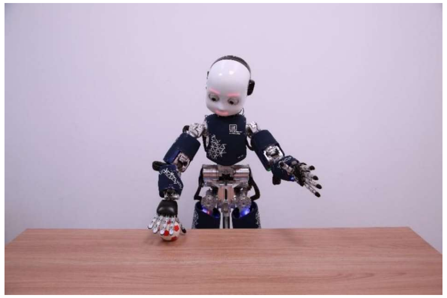
iCub è affascinato
dall’uso degli utensili.

M= 46.40; SD: 41.53

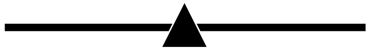
Item 34







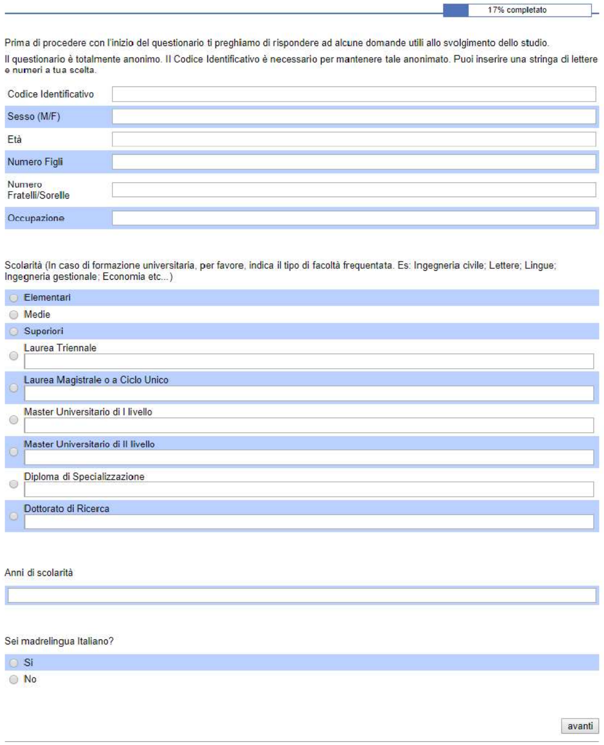
iCub aggiorna le coordinate della posizione della palla.



iCub ha deciso di non lanciare la palla.

M= 34.40; SD: 36.91

Demographics



Prima di procedere con l'inizio del questionario ti preghiamo di rispondere ad alcune domande utili allo svolgimento dello studio.

Il questionario è totalmente anonimo. Il Codice Identificativo è necessario per mantenere tale anonimato. Puoi inserire una stringa di lettere e numeri a tua scelta.

Codice Identificativo	
Sesso (M/F)	
Età	
Numero Figli	
Numero Fratelli/Sorelle	
Occupazione	

Scolarità (In caso di formazione universitaria, per favore, indica il tipo di facoltà frequentata. Es: Ingegneria civile; Lettere; Lingue; Ingegneria gestionale; Economia etc...)

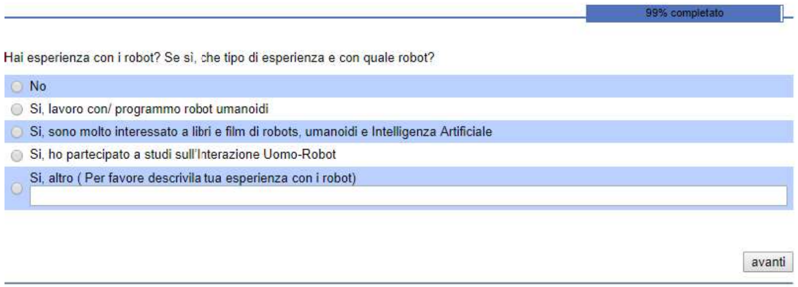
- ○ Elementari
- ○ Medie
- ○ Superiori
- ○ Laurea Triennale
- ○ Laurea Magistrale o a Ciclo Unico
- ○ Master Universitario di I livello
- ○ Master Universitario di II livello
- ○ Diploma di Specializzazione
- ○ Dottorato di Ricerca

Anni di scolarità

Sei madrelingua Italiano?

- ○ Si
- ○ No

Familiarity



Hai esperienza con i robot? Se sì, che tipo di esperienza e con quale robot?

○ No

○ Si, lavoro con/ programmo robot umanoidi

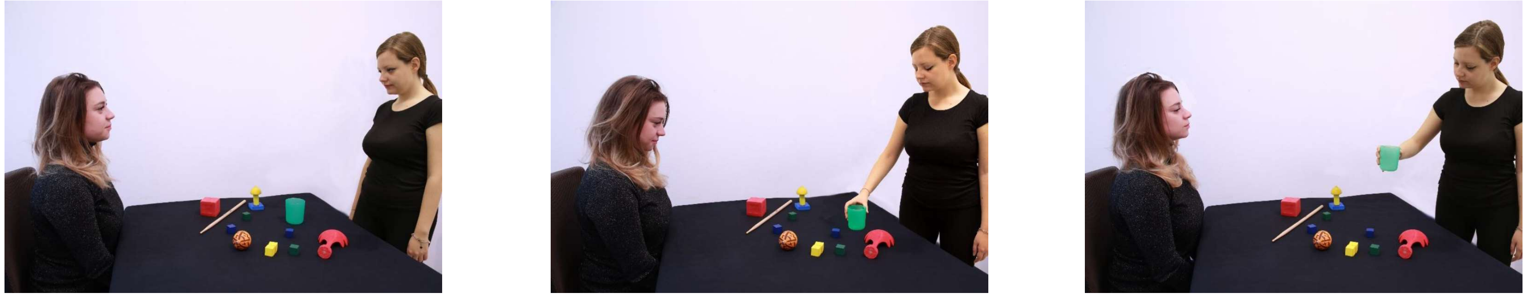
○ Si, sono molto interessato a libri e film di robots, umanoidi e Intelligenza Artificiale

○ Si, ho partecipato a studi sull'Interazione Uomo-Robot

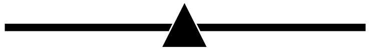
○ Si, altro (Per favore descrivila tua esperienza con i robot)

Human Control Questionnaire
Supplementary material 3 English

Item 3 Control



Paola grasps cylindrical objects best.

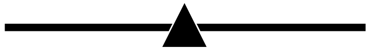


Paola believes that the girl likes the cup.

M= 52.05; SD: 41.61

Item 4 Control

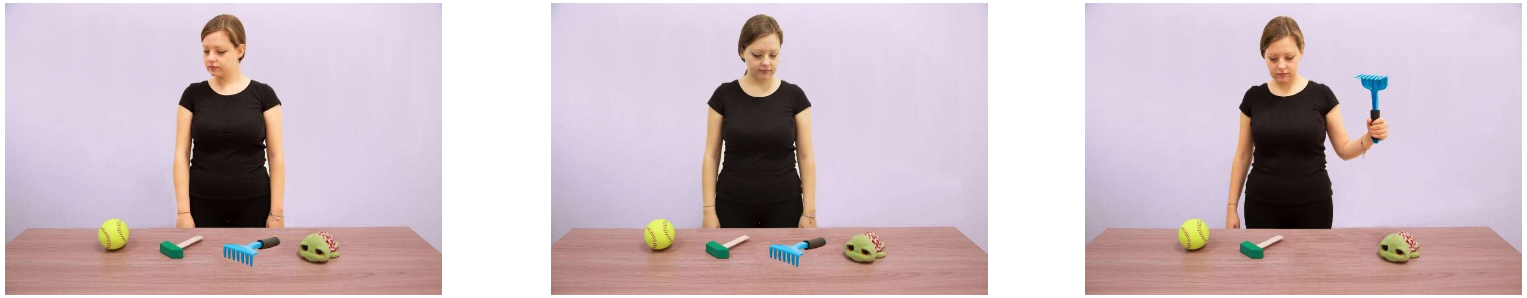
Paola is calculating the number coins.



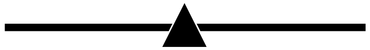
Paola decided to give back the money.

M= 51.47; SD: 42.90

Item 8 Control



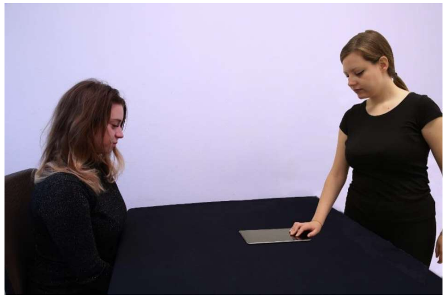
Paola adjusts the force to
the weight of the object.

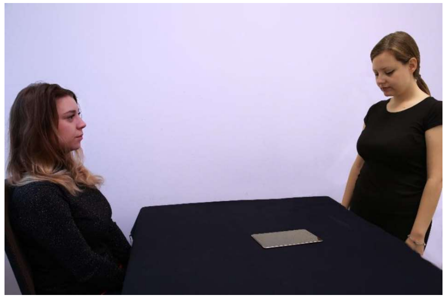


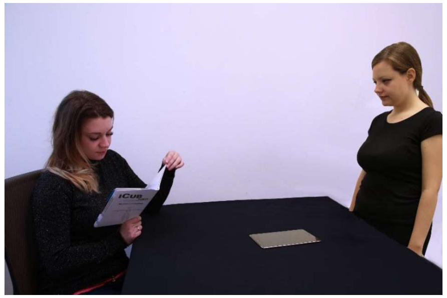
Paola pretends to be
a gardener.

M= 42.25; SD: 40.90

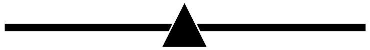
Item 10 Control







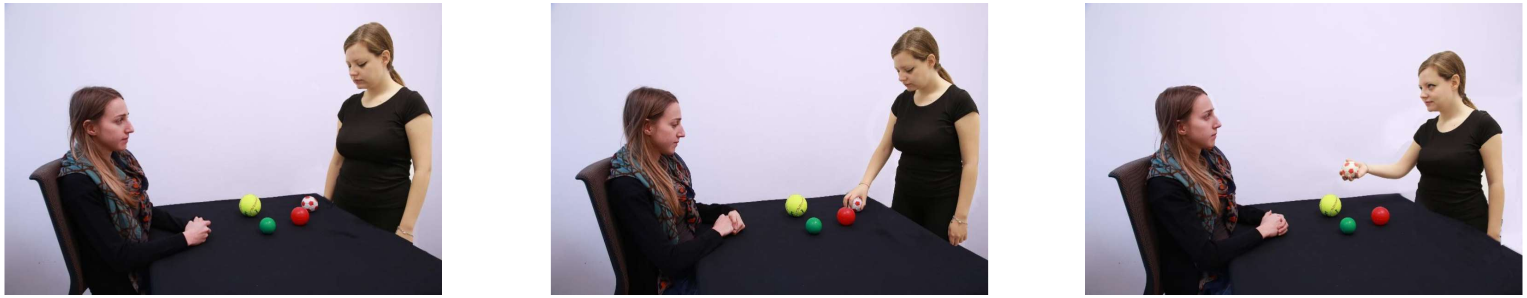
Paola detected differences between the old and the new object in the scene.



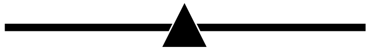
Paola wants to read from the book too.

M= 45.61; SD: 40.99

Item 11 Control



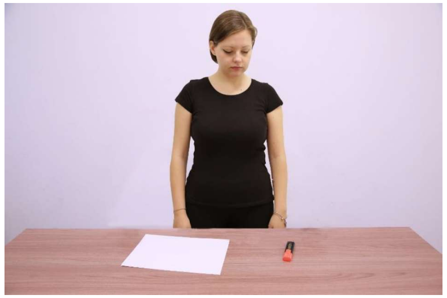
Paola calculates the weight of the balls.

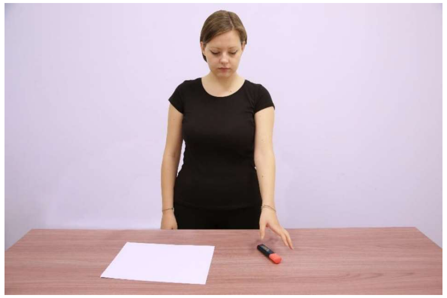


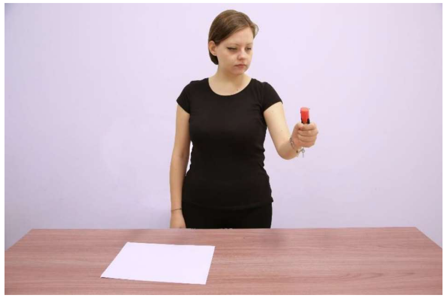
Paola wants to play with the girl.

M= 74.92; SD: 36.36

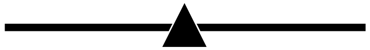
Item 13 Control







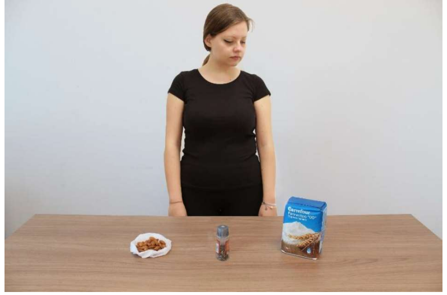
Paola optimizes grip
for small objects.

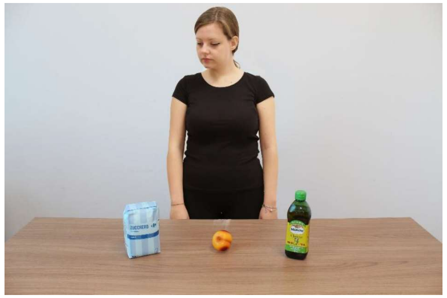


Paola wants to
draw something.

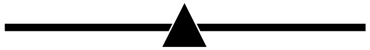
M= 56.47; SD: 42.87

Item 16 Control





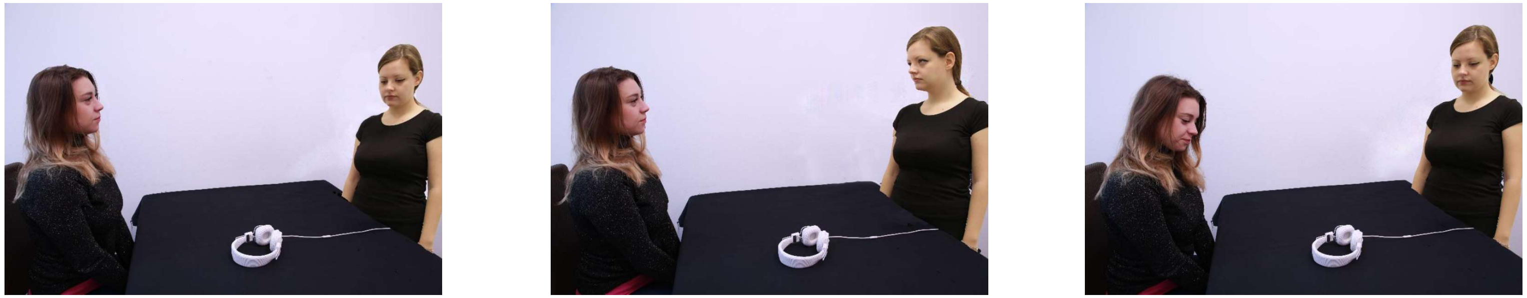
Paola follows the recipe's instructions.



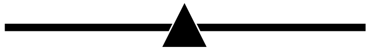
Paola has decided to bake a cake.

M= 51.97; SD: 41.55

Item 18 Control



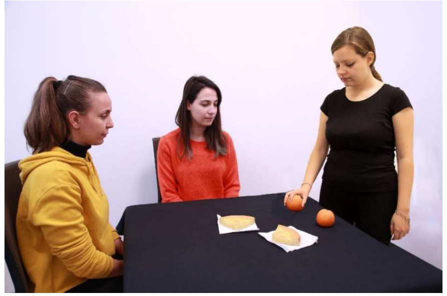
Paola measures the distance between the girl and the headphones.

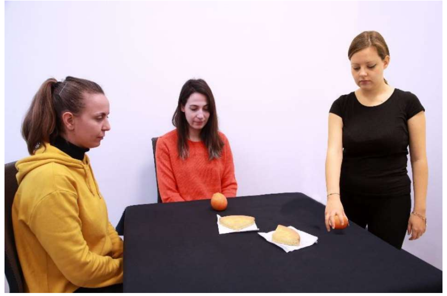


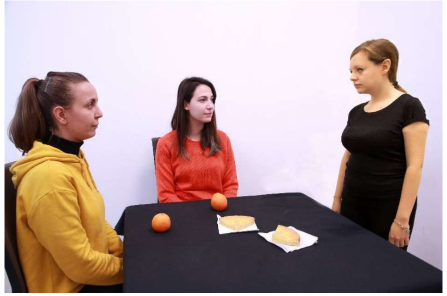
Paola expects that the girl would lend the headphones.

M= 66.82; SD: 37.46

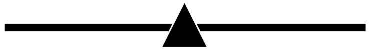
Item 21 Control







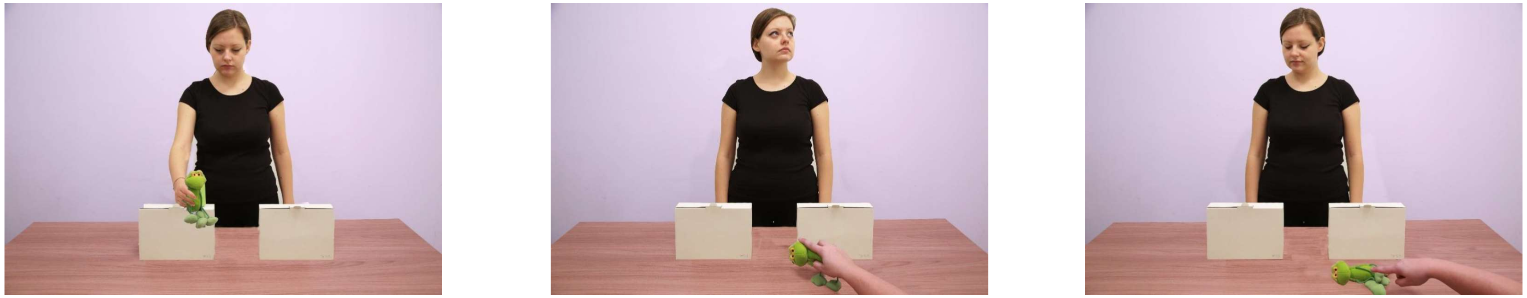
Paola can grasp oranges
easier than cakes.



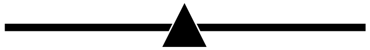
Paola thinks that the girls
need some healthy food.

M= 35.99; SD: 42.48

Item 22 Control



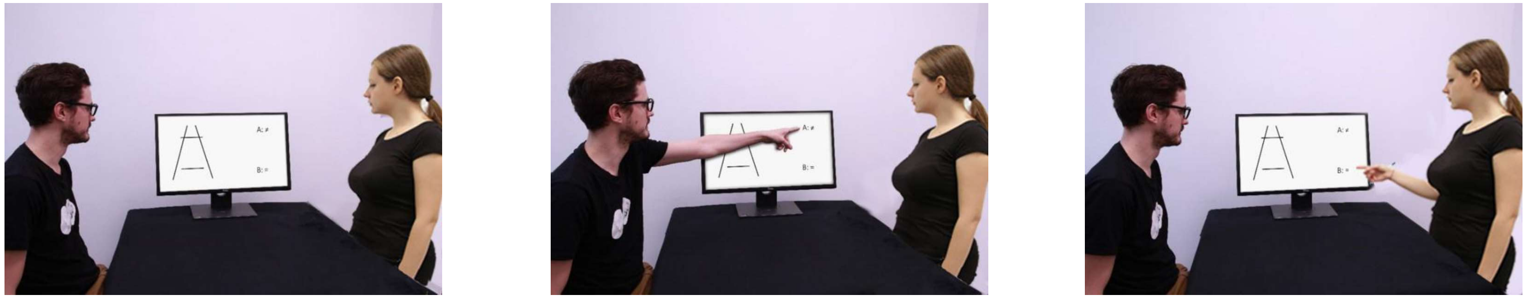
Paola is turning the head to the initial position.



Paola is not interested in the toy anymore.

M= 59.30; SD: 41.65

Item 23 Control

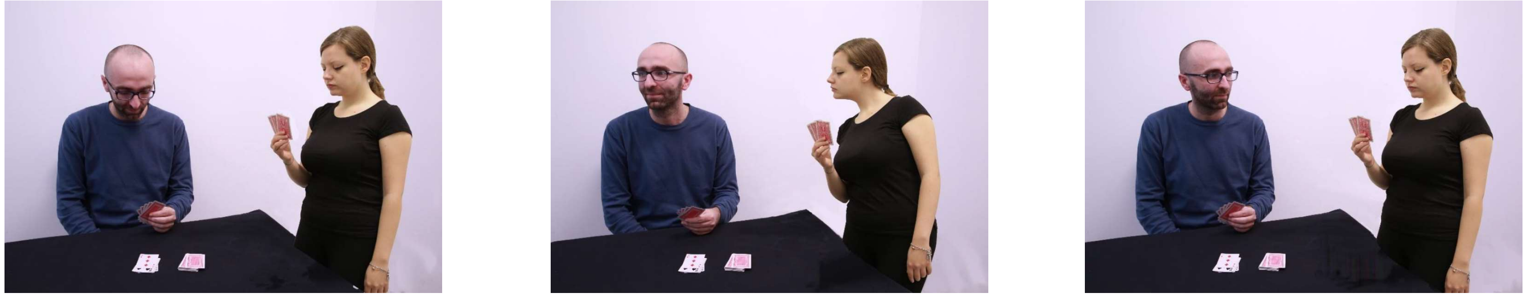


Paola is repeating the pointing movement.

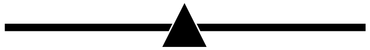
Paola is expressing an opinion.

M= 78.22; SD: 33.44

Item 25 Control



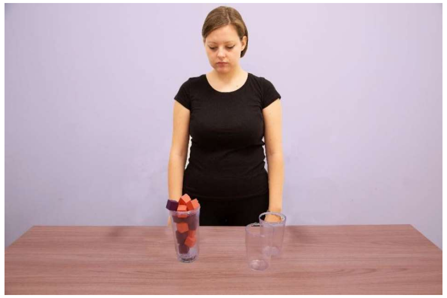
Paola was unbalanced
for a moment.

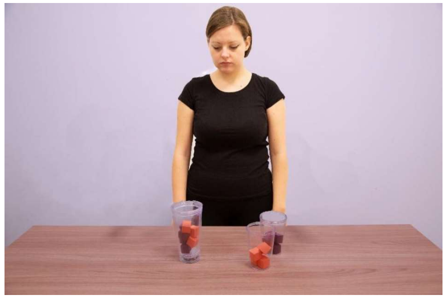


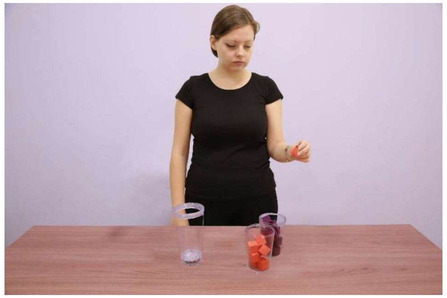
Paola was trying to cheat by
looking at opponent’s cards.

M= 72.58; SD: 37.44

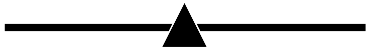
Item 28 Control







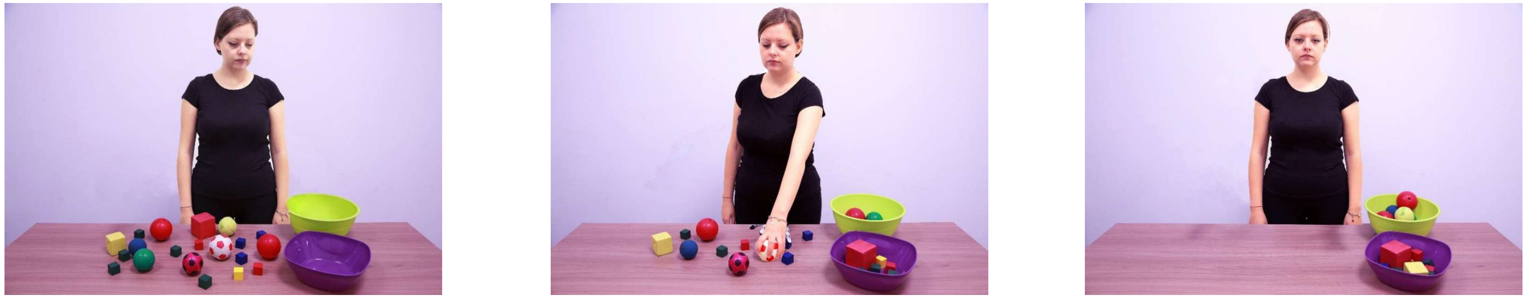
Paola classifies
cubes by color.



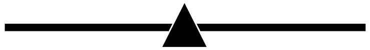
Paola would like
to keep this cube.

M= 43.53; SD: 45.63

Item 31 Control



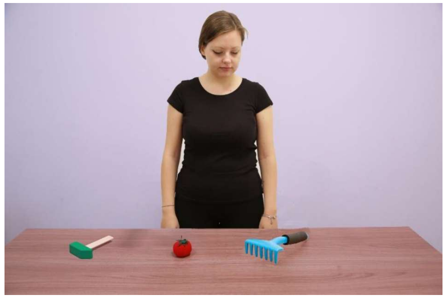
Paola places objects in bowls with corresponding shape.

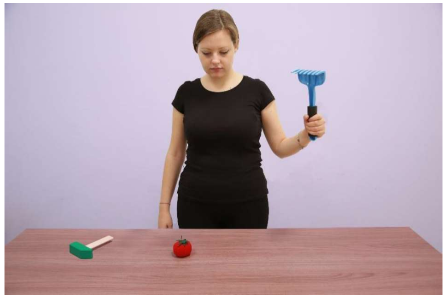


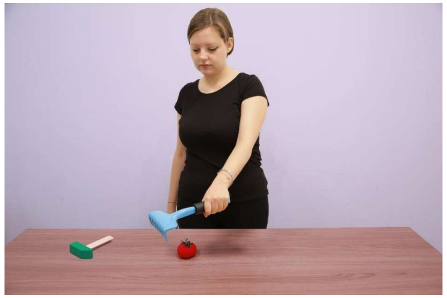
Paola knows that it’s better to clean the table after playing.

M= 53.76; SD: 44.95

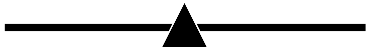
Item 33 Control







Paola grasped the
closest object.

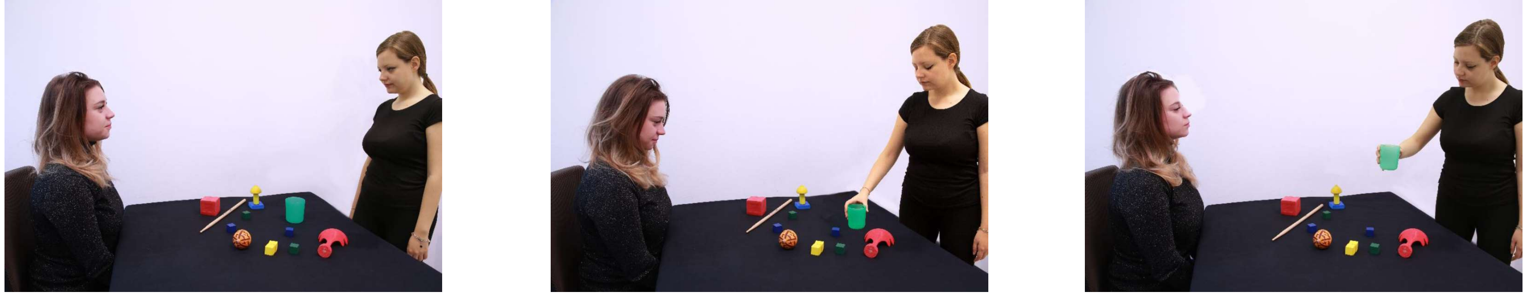


Paola was fascinated
by tool use.

M= 34.33; SD: 40.75

Human Control Questionnaire
Supplementary material 3 ITALIAN

Item 3 Control



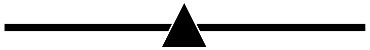
Paola afferra meglio gli oggetti cilindrici.

Paola pensa che alla ragazza piaccia il bicchiere.

M= 52.05; SD: 41.61

Item 4 Control

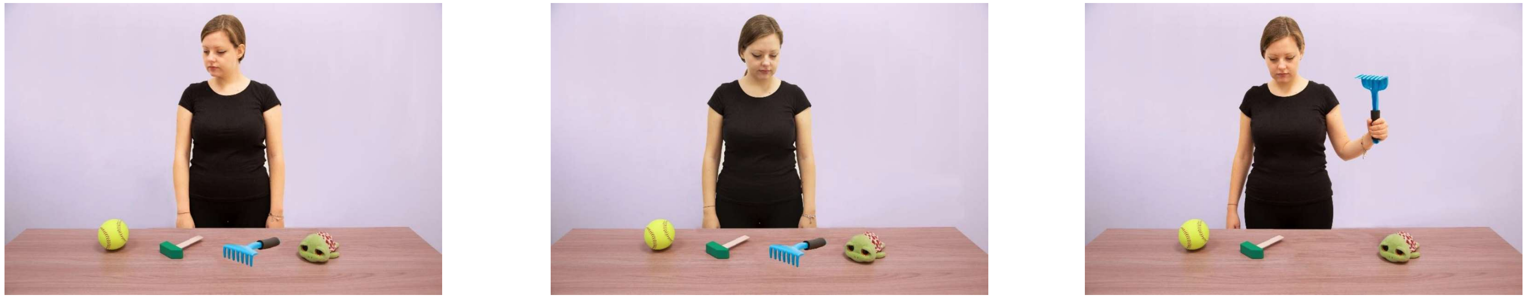
Paola calcola il numero di monete.



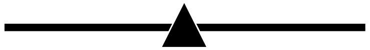
Paola ha deciso di restituire le monete.

M= 51.47; SD: 42.90

Item 8 Control



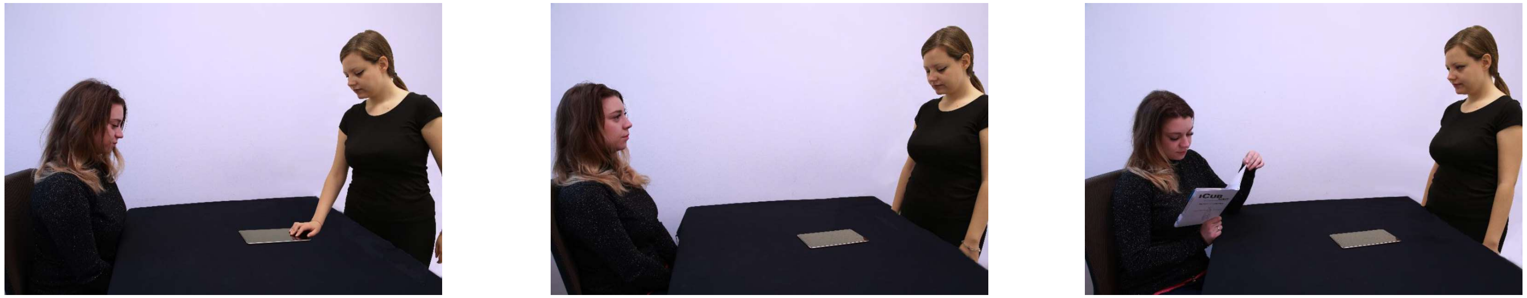
Paola regola la forza in base
al peso dell’oggetto.



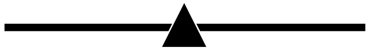
Paola finge di essere
un giardiniere.

M= 42.25; SD: 40.90

Item 10 Control



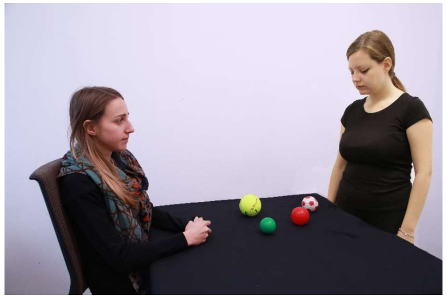
Paola rileva differenze tra il vecchio e il nuovo oggetto in scena.

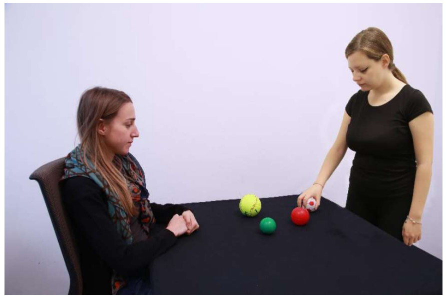


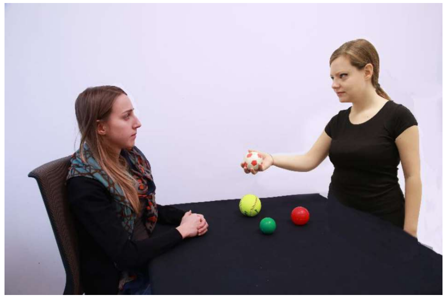
Paola vuole leggere
il libro.

M= 45.61; SD: 40.99

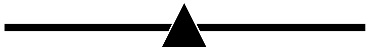
Item 11 Control







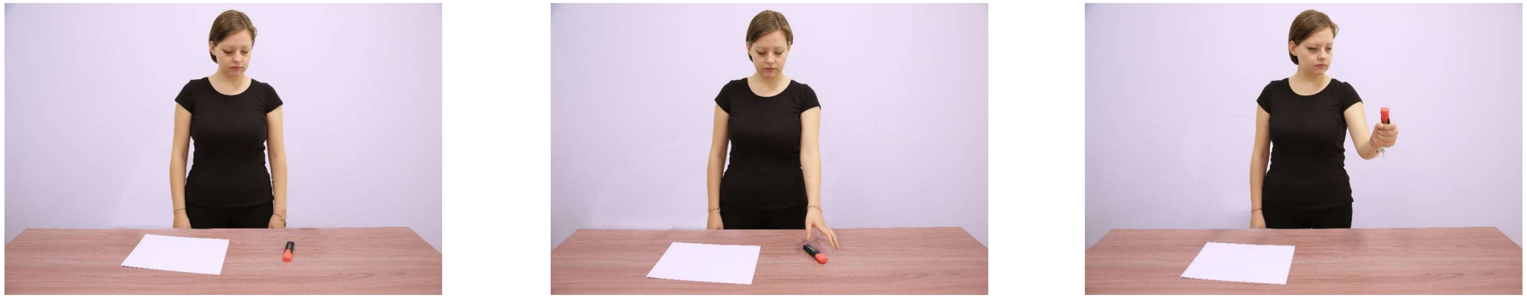
Paola calcola il peso
delle palle.



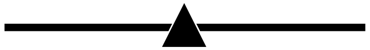
Paola vuole giocare
con la ragazza.

M= 74.92; SD: 36.36

Item 13 Control



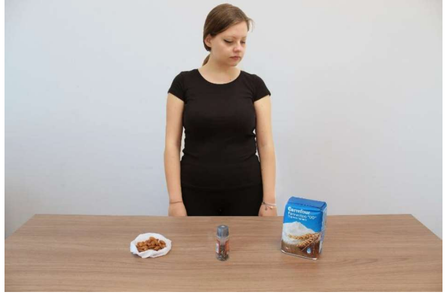
Paola ottimizza la presa
gli oggetti piccoli.

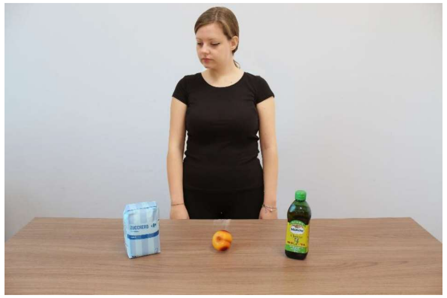


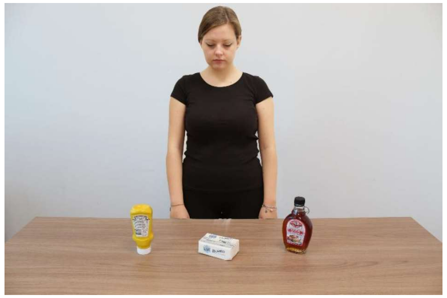
Paola vuole disegnare
qualcosa.

M= 56.47; SD: 42.87

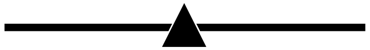
Item 16 Control







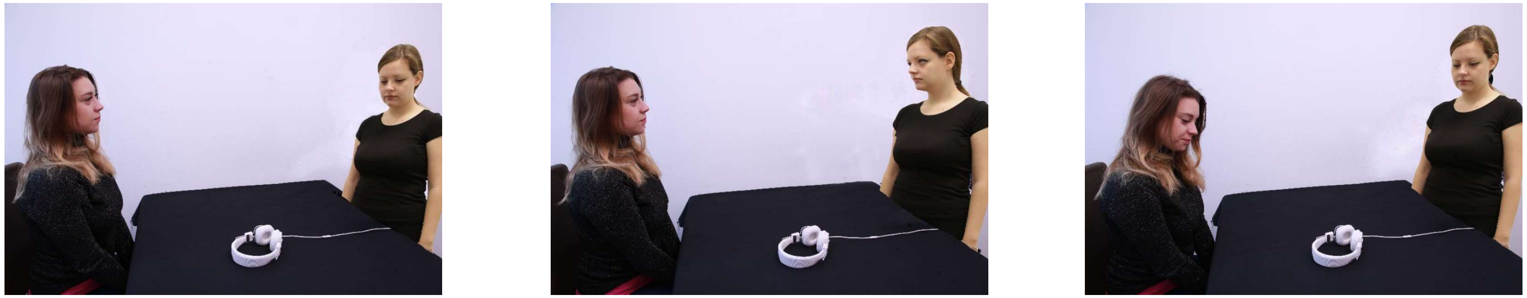
Paola segue le istruzioni
della ricetta.



Paola ha deciso di
cucinare una torta.

M= 51.97; SD: 41.55

Item 18 Control

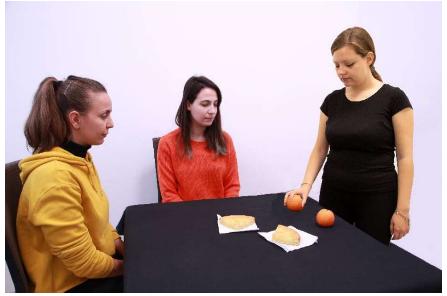


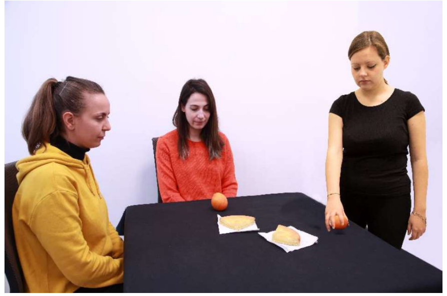
Paola misura la distanza tra la ragazza e le cuffie.

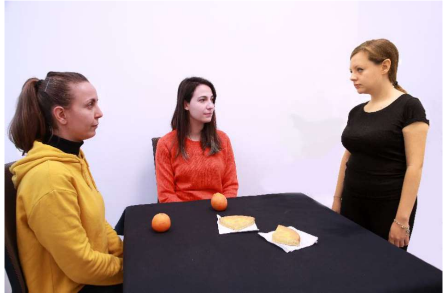
Paola si aspetta che la ragazza dia in prestito le cuffie.

M= 66.82; SD: 37.46

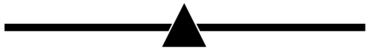
Item 21 Control







Paola pensa che le ragazze
abbiano bisogno di cibo sano.

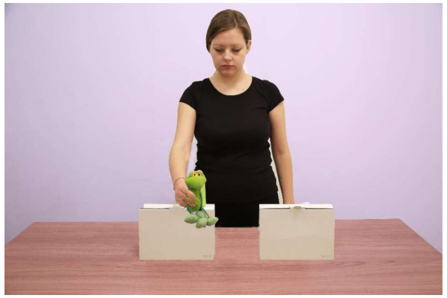


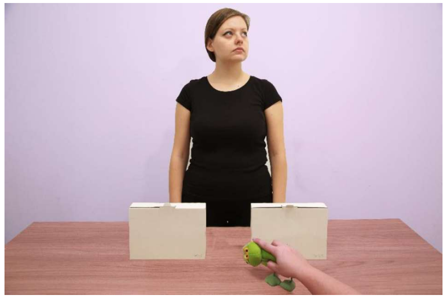
Paola può afferrare le arance
più facilmente della torta.

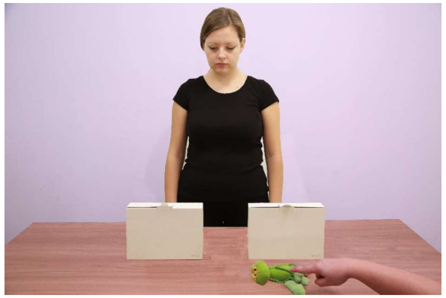
M= 35.99; SD: 42.48

100

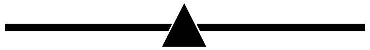
Item 22 Control







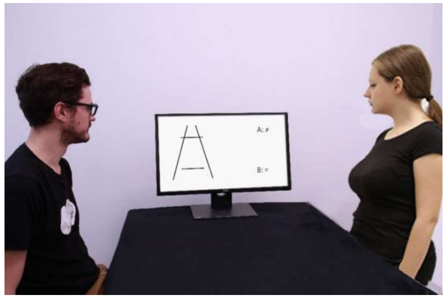
Paola riporta la testa nella
posizione iniziale.

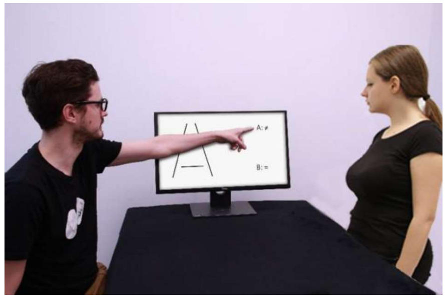


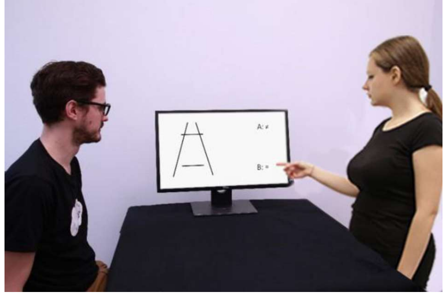
Paola non è più
interessata al giocattolo.

M= 59.30; SD: 41.65

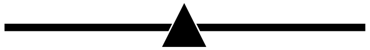
Item 23 Control







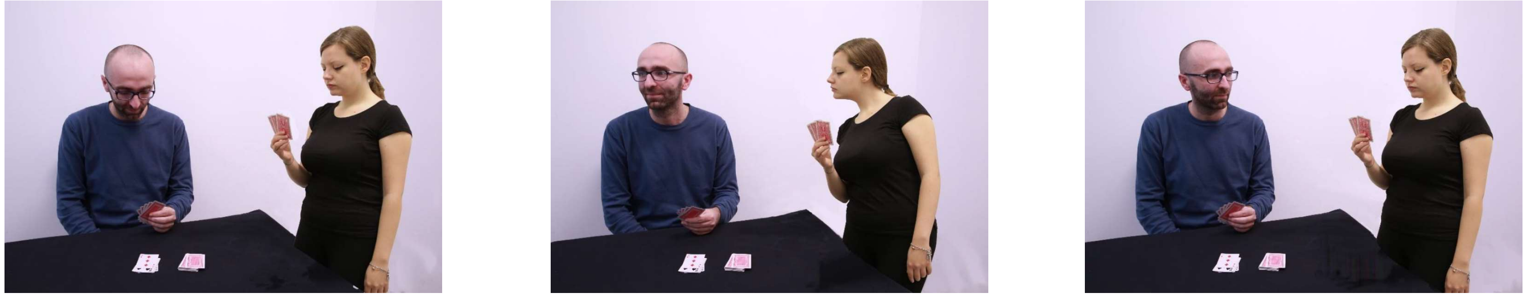
Paola sta ripetendo il
gesto di indicare.



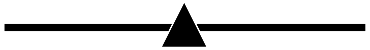
Paola sta esprimendo
un'opinione.

M= 78.22; SD: 33.44

Item 25 Control



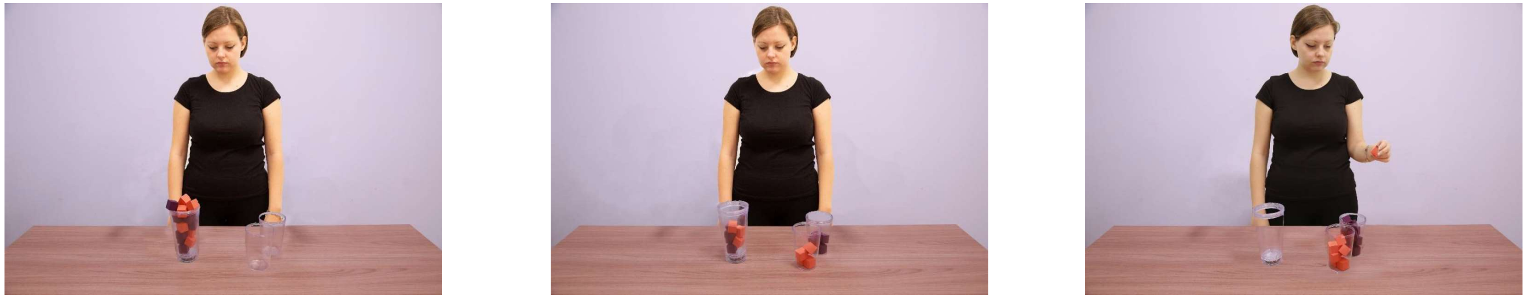
Paola bara guardando le carte dell’avversario.



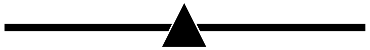
Paola ha perso l'equilibrio per un momento.

M= 72.58; SD: 37.44

Item 28 Control



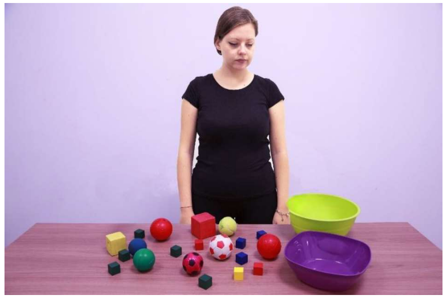
Paola classifica i cubi
in base al colore.

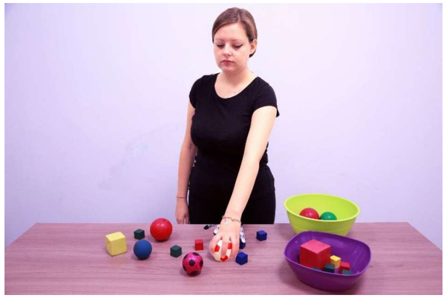


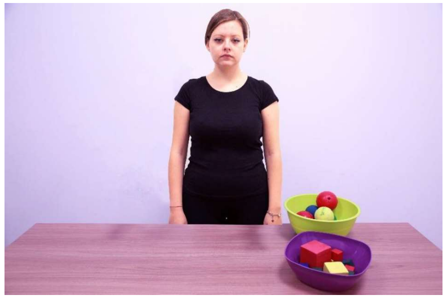
Paola vorrebbe tenere
questo cubo.

M= 43.53; SD: 45.63

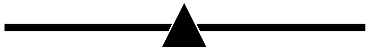
Item 31 Control







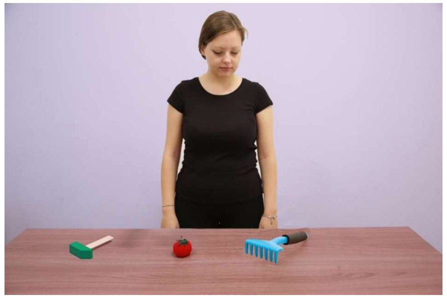
Paola mette gli oggetti nel contenitore
con la forma corrispondente.

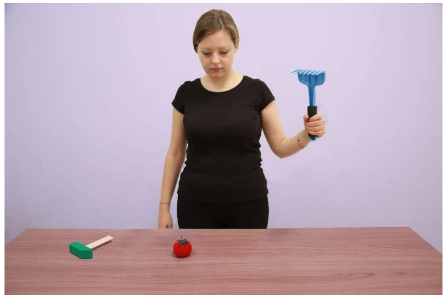


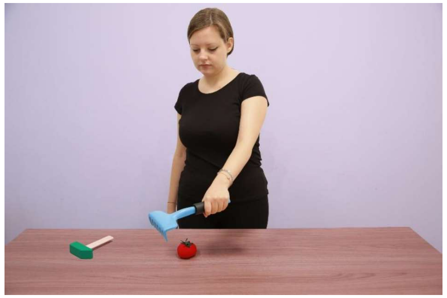
Paola sa che è giusto mettere in
ordine il tavolo dopo aver giocato.

M= 53.76; SD: 44.95

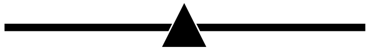
Item 33 Control







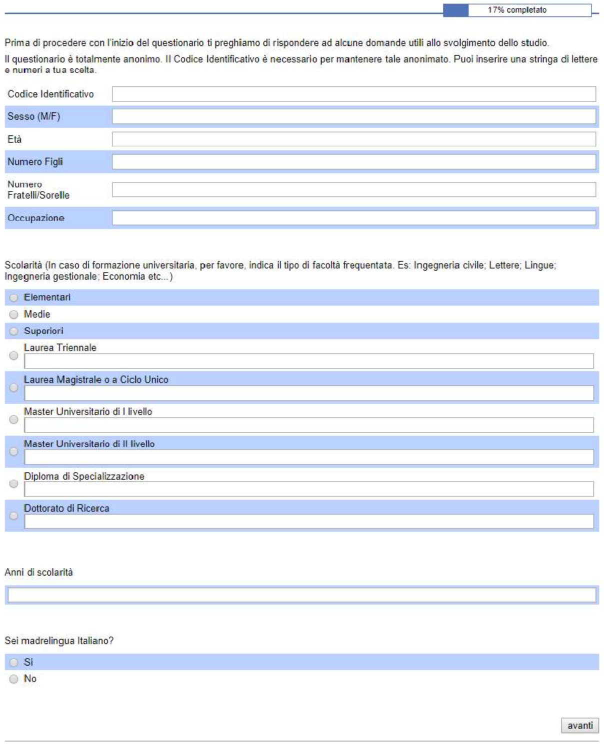
Paola afferra
l’oggetto più vicino.



Paola è affascinata
dall’uso degli utensili.

M= 34.33; SD: 40.75

Demographics



Prima di procedere con l'inizio del questionario ti preghiamo di rispondere ad alcune domande utili allo svolgimento dello studio.

Il questionario è totalmente anonimo. Il Codice Identificativo è necessario per mantenere tale anonimato. Puoi inserire una stringa di lettere e numeri a tua scelta.

Codice Identificativo

Sesso (M/F)

Età

Numero Figli

Numero Fratelli/Sorelle

Occupazione

Scolarità (In caso di formazione universitaria, per favore, indica il tipo di facoltà frequentata. Es: Ingegneria civile; Lettere; Lingue; Ingegneria gestionale; Economia etc...)

○ Elementari

○ Medie

○ Superiori

○ Laurea Triennale

○ Laurea Magistrale o a Ciclo Unico

○ Master Universitario di I livello

○ Master Universitario di II livello

○ Diploma di Specializzazione

○ Dottorato di Ricerca

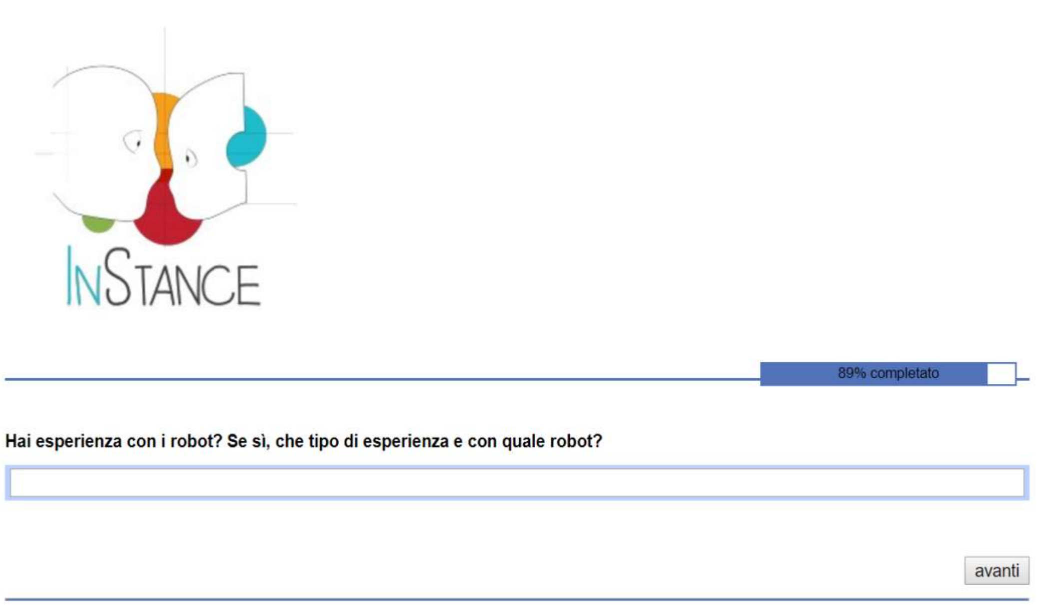
Anni di scolarità

Sei madrelingua Italiano?

○ Si

○ No

Familiarity



Hai esperienza con i robot? Se sì, che tipo di esperienza e con quale robot?

Appendix B

Supplementary Materials Chapter 4

B.1 Plot density of raw response times from both agents

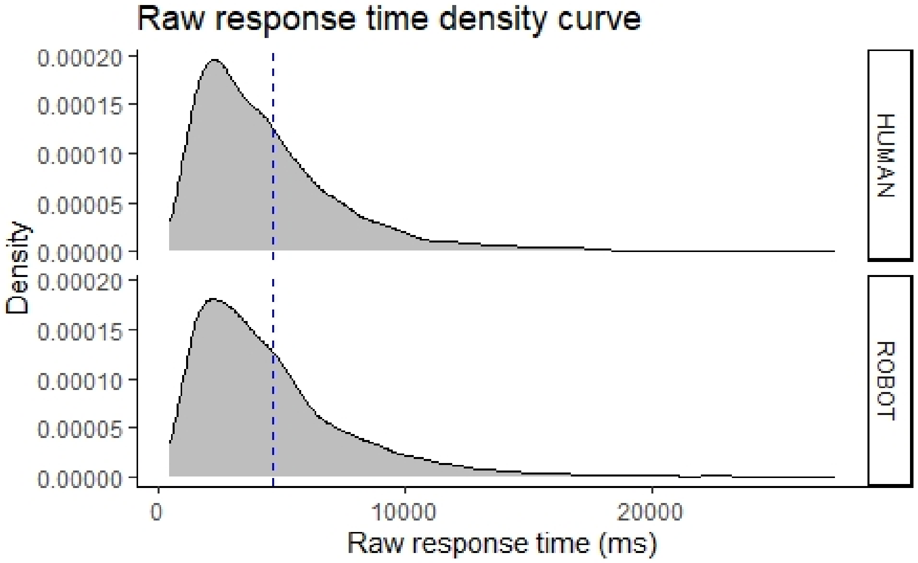
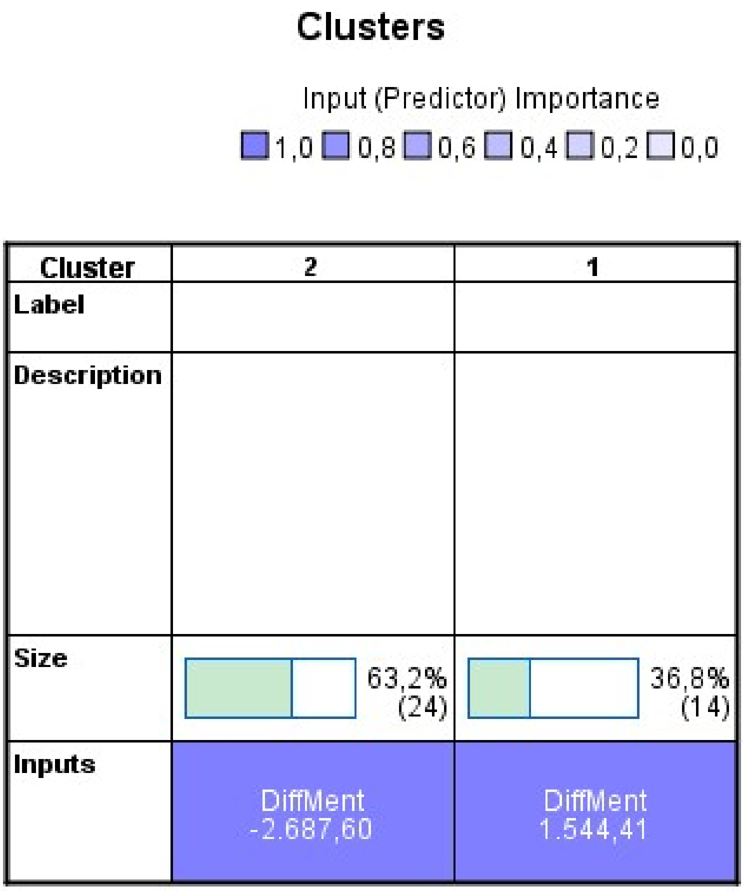


Figure B.1. Density distribution of raw response times from both Human and Robotic agents

B.2 Cluster descriptive statistics and plots

We performed a Two-Step Clustering using SPSS (version 21.0.0, Bacher et al., 2004). The first step consists of quickly creating a pre-grouping of observations into a large number of classes using the BIRCH method (Zhang et al., 1996). As a second step, the classes are gradually merged until a single group is obtained. Euclidean distance can be used if the variables are all quantitative. A log-likelihood based measure is implemented in the case of mixed variables. Once the hierarchy has been developed, each division can be evaluated using the Bayesian Information Criterion (BIC), and an ad hoc process that identifies the appropriate number of classes based on the ratio of BIC changes and the ratio of distance measurements.



Clusters

Input (Predictor) Importance

▇ 1,0 ▇ 0,8 ▇ 0,6 ▇ 0,4 ▇ 0,2 □ 0,0

Cluster	2	1
Label		
Description		
Size	63,2% (24)	36,8% (14)
Inputs	DiffMent -2.687,60	DiffMent 1.544,41

Figure B.2. Clusters composition

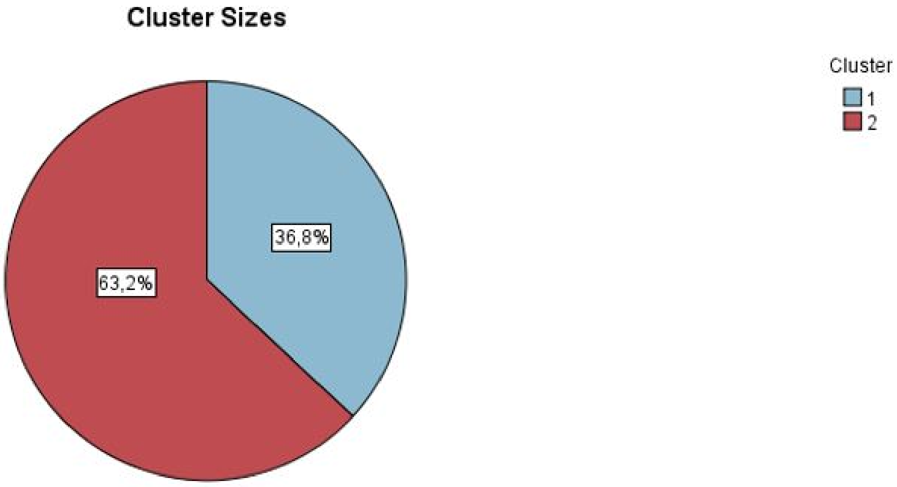


Figure B.3. Cluster sizes

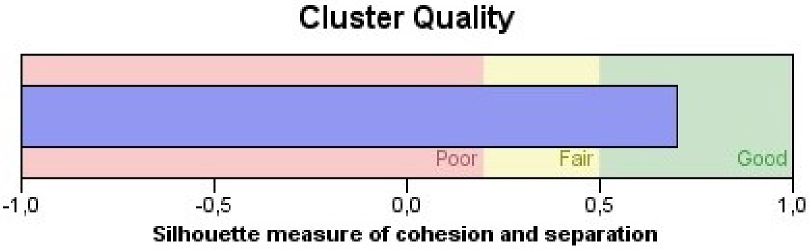


Figure B.4. Silhouette measure of cohesion and separation of the clusters

Size of The Smallest Cluster	Size of the Largest Cluster	Ratio of Sizes (Largest to The Smallest cluster)
14 (36.8%)	24(63.2%)	1.71

Table B.1. Clusters descriptive statistics.

Clusters	Mean	95% C.I.	SD	Minimum	Maximum	Skewness	Kurtosis
Cluster 1	1544.41	[678.54, 2410.28]	1499.64	−134.22	5386.05	1.52	2.25
Cluster 2	−2687.60	[−3428.95, −1946.25]	1755.67	−7730.52	−872.89	−1.54	2.01

Table B.2. Clusters descriptive statistics.

Appendix C

Supplementary Materials Chapter 6

C.1 Supplementary Data

C.1.1 Behavioural data analysis on IST and response times (Mechanistic and Mentalistic biased group)

First, we considered the z-transformed IST score as independent variable and the log-transformed RTs as dependent variable. No statistically significant effect emerged from the model [b = .001, t (<.001) = 0.488, p = .626].

C.2 IST individual score and pupil results including the unbiased group

The first model (GLMM) aimed at investigating the relationship between pupil size and the selected description in the IST (the binomial attribution of intentional or mechanistic behaviour). In this first GLMM we included participants as random effect. Our fixed effects were: 1) mean pupil size; 2) robot behaviour previously observed; 3) participants' general bias at the IST as independent variable in a full factorial design, while we considered the selected attribution to explain the item (by considering as mechanistic a score < 50 and mentalistic a score > 50) as dependent variable. Because of this, the distribution of the GLMM is binomial. Results showed that the interaction effect RobotBehaviour * Bias emerged as statistically significant [$\chi^2(2)$ = 15.537, p = <.001]. We further investigated the contrast between mentalistic and mechanistic attribution with planned pairwise comparisons (Tukey's HSD correction for multiple comparisons): mechanistic group: z= 2.424, p= 0.01; mentalistic group: z= 2.932, p= 0.003; unbiased group: z= -2.031, p= 0.042. Results showed that participants in each group significantly differed in the mentalistic and mechanistic attribution in the IST. Specifically, both biased groups chose more often an attribution congruent with the behaviour previously observed on the robot (more mechanistic attribution after watching machine-like behaviour and vice-versa). On the other hand, the unbiased group showed the opposite pattern, i.e., choosing more often the mechanistic attribution after incongruent (human-like) behaviour and the mentalistic attribution after mechanistic behaviour.

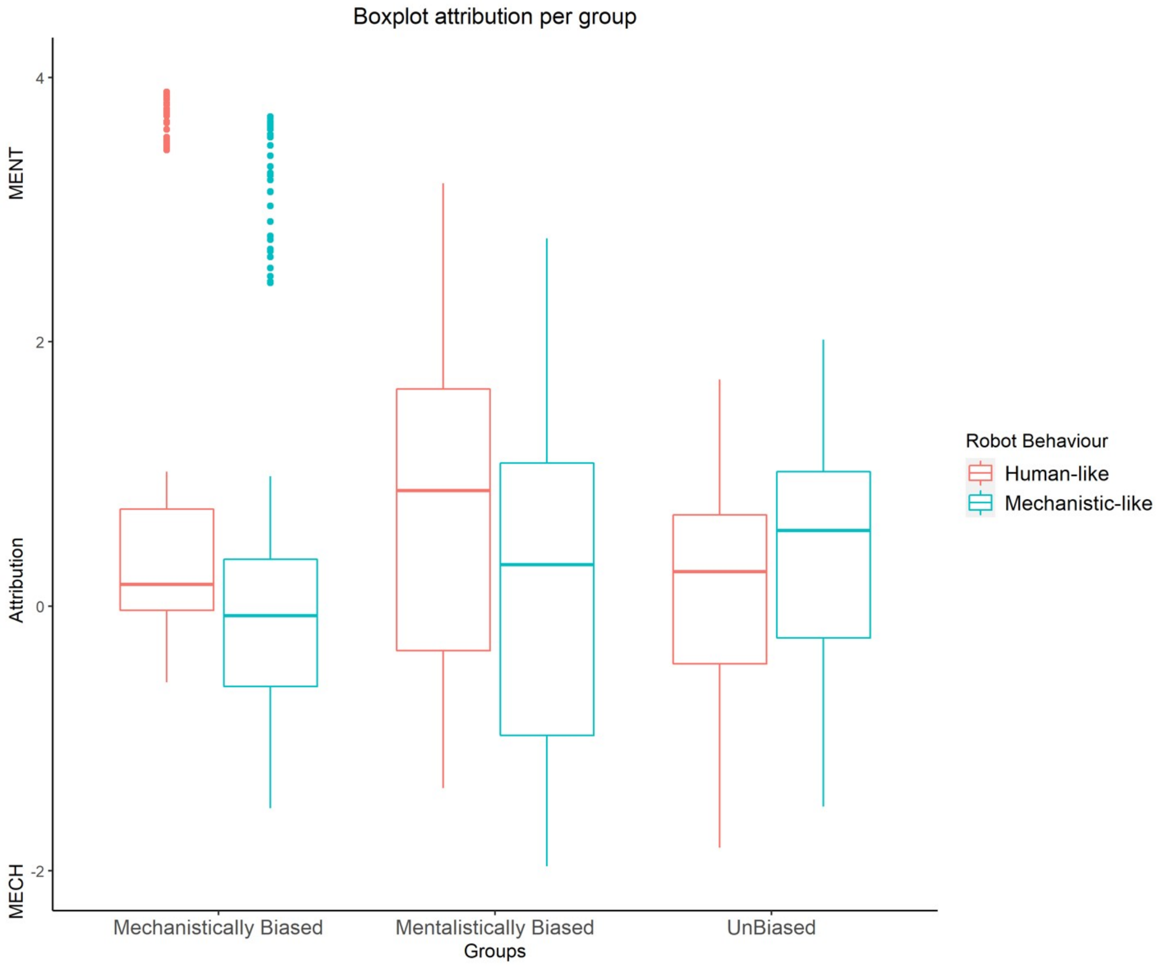


Figure C.1. GLMM: Boxplot showing the statistically significant effect RobotBehaviour * Bias on attribution.

The interaction effect between RobotBehaviour * Bias * mean pupil size was statistically significant [[$\chi^2(2)$ = 8.62, p = .013]. To investigate the interaction between RobotBehaviour * Bias * mean pupil size, we tested the RobotBehaviour * mean pupil size interaction in three separate GLMMs, one for each bias group: mechanistic group [$\chi^2(1)$ = 7.701 p = .006]; mentalistic group [$\chi^2(1)$ = 3.001, p =.083]; unbiased group ($\chi^2(1)$ = 1.064, p = .302). These results show that mechanistically-biased participants showed a greater pupil dilation for attributions congruent with the robot behaviour [b= -9.28, z = -2.755, p =.005]: when attributing a mechanistic description after the observation of the robot behaving in a machine-like way and when attributing a mentalistic score after the observation of the robot behaving in a human-like way. On the other hand, mentalistic-biased participants, showed a tendency, although insignificant, towards greater pupil sizes for mentalistic attributions, relative to mechanistic attributions, regardless of the robot behaviour [b= -4.45, z = -1.73, p = .083, Fig. C.2]. The unbiased group of participants tended to show opposite effects than the mechanistically-biased sample, but the modulatory effects on pupil dilation have not reached the level of significance [b= 2.60, z = 1.03, p = .302, Fig. C.2].

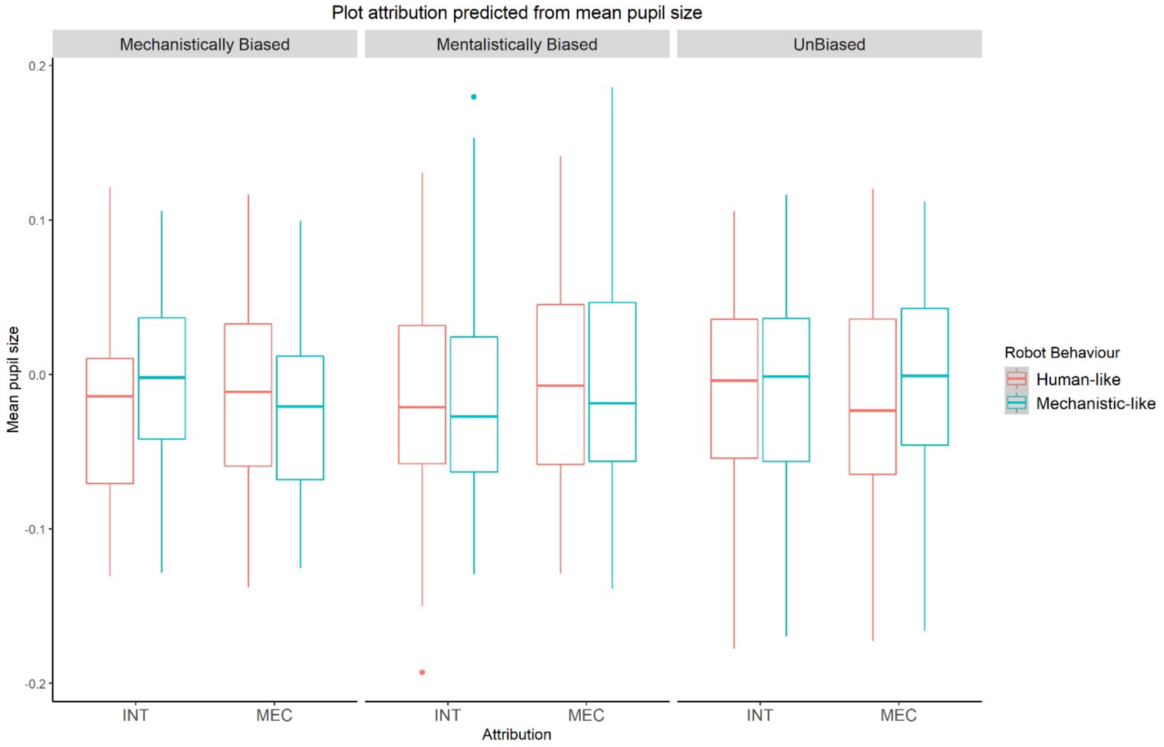


Figure C.2. GLMM on mechanistic group (N= 9), mentalistic group (N= 12) and unbiased group (N= 13): mechanistic bias group show the interaction effect between attribution and mean pupil size. No statistically significant effect on attribution and pupil size on mentalistic bias group and unbiased.

C.3 Response time and pupil size analysis

The second model aimed at investigating the relationship between pupil dilation and response times. Here, we considered the response times transformed on a logarithmic scale as dependent variable and mean pupil size, robot behaviour previously observed and participants' bias as independent variable. To create database with variables for this analysis we excluded 8 trials as speed outliers (> 20 sec). Results showed a significant three-way interaction between mean pupil size * Bias * Robot Behaviour [b = 2.401, t (2269.69) = 3.275, p = .001] (Supp. Fig.C.3). To investigate the interaction between pupil size * Bias * RobotBehaviour, we tested the two-way interaction between pupil size * RobotBehaviour for each bias group in three separate models: mechanistic group [b = -2.079, t (597.56) = -3.766, p = < .001]; mentalistic group [b = -0.297, t (802.489) = 0.662, p = 0.508]; unbiased group [b = -0.173, t (870.765) = -0.378, p = 0.706]. As shown in Figure C.2, participants with a mechanistic bias showed an inverse relation between pupil dilation and response times after the robot showing a behaviour congruent with their bias: the greater the pupil dilation, the faster they were in choosing the attribution. This relationship was not significant after they observed the non-congruent behaviour. On the other hand, as shown in Fig. C.3, mentalistically-biased participants showed faster response times and a larger mean pupil size independently of the previously observed robot's behaviour (main effect of pupil: [b = -1.31, t (809.16) = -3.948, p = <.001], main effect of robot behaviour: [b = 0.06, t (799.06) = 2.256, p = .024]. Moreover, unbiased participants showed a main effect of pupil [b = -1.70., t (876.60) = -4.927, p = < .001] but not a main effect of robot behaviour [b = -0.01, t (876.48) = -0.717, p = .473] nor the interaction effect [b = -0.17., t (870.76) = -0.378, p = .706] A tentative interpretation of these results on pupil dilation in relation to participants' response times may be the following:

Mechanistically-biased participants: although the main analysis shows certain degree of cognitive flexibility of the mechanistically inclined participants (see main text), this flexibility might have had a cost in terms of cognitive resources, due to the additional resources devoted to integration of subtle behavioural cues from the robot. This cognitive cost may be reflected in participants' response time (Spears and Haslam, 1997)

Mentalistically-biased participants: a larger pupil dilation was associated with faster RTs overall. Perhaps this group of participants showed increased engagement of cognitive resources in the mentalistic descriptions overall, as they were trying to adhere to their initial bias, and respond as fast as possible in line with the bias, but at the same time, "making sense" of an artificial agent.

Unbiased participants: showed no modulation of the pupil size related to their attributions, but a larger pupil dilation was associated with faster RTs overall. This is in line with previous literature (Van Kempen et al., 2019) on the relationship between pupil size (and associated commitment of cognitive resources) and speed of responding. However, this is not necessarily related to a mentalistic or a mechanistic bias.

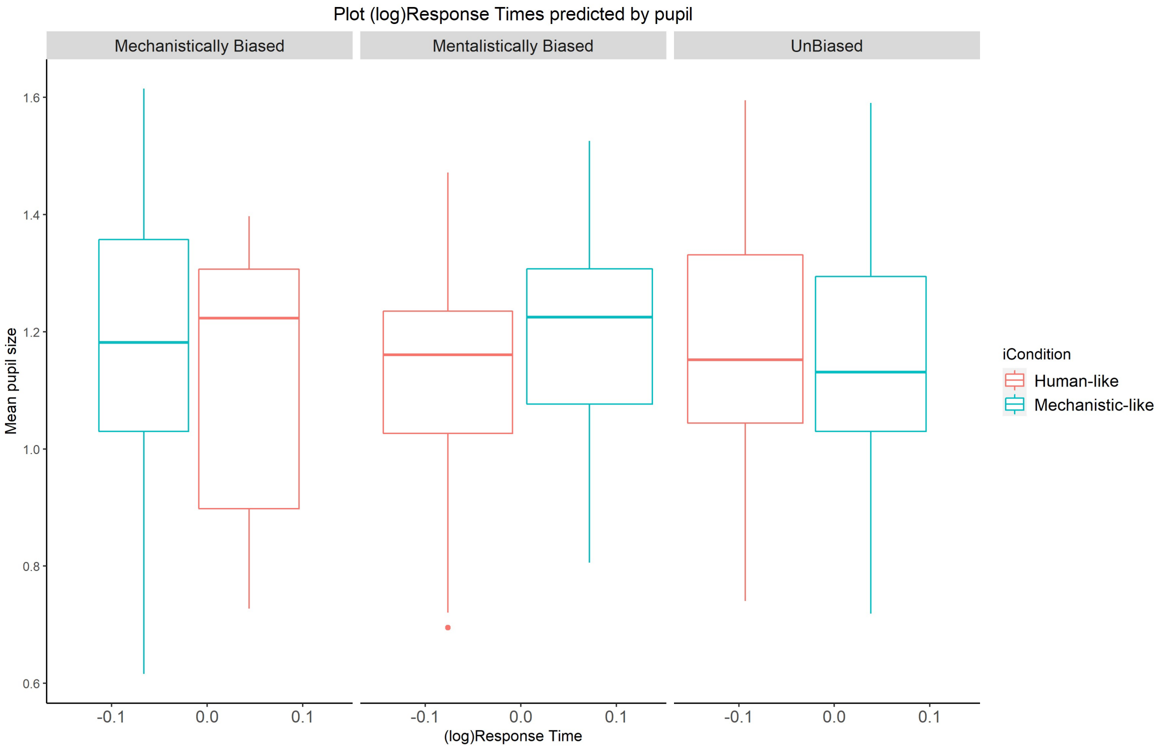


Figure C.3. LMM on mechanistic group (N=9), mentalistic group (N=12) and unbiased group (N=13): the mechanistic bias group shows the interaction effect between response times (log-corrected) and mean pupil size. The mentalistic bias group shows statistically significant main effect of pupil size and robot behaviour, nut no interaction. The unbiased group shows a statistically significant main effect of pupil size but no main effect of robot behaviour, nor interaction.

C.4 Table with models parametes

Fixed effects	Estimate	Std Error	z value	Pr(> |z|)	Significance
(Intercept)	0.658	0.474	1.388	0.165	
MeanPupilDiameter	2.296	2.374	0.967	0.333	
RobotBehaviour	−0.537	0.187	−2.864	0.004	**
Bias	0.060	0.626	0.096	0.923	
MeanPupilDiameter:RobotBehaviour	−9.291	3.372	−2.755	0.005	**
MeanPupilDiameter:Bias	−1.510	3.076	−0.491	0.623	
RobotBehaviour:Bias	−0.01	0.254	−0.040	0.967	
MeanPupilDiameter:RobotBehaviour:Bias	4.858	4.239	1.146	0.251	

Table C.1. Results of Generalized Linear Mixed Effect Model on the total sample. Signifiance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Fixed effects Mechanistic biased group	Estimate	Std Error	z value	Pr(> |z|)	Significance
(Intercept)	0.641	0.438	1.461	0.143	
MeanPupilDiameter	2.305	2.373	0.972	0.331	
RobotBehaviour	−0.53	0.187	−2.862	0.004	**
MeanPupilDiameter:RobotBehaviour	−9.284	3.368	−2.757	0.005	**

Table C.2. Results of Generalized Linear Mixed Effect Model on the Mechanistic Biased sample. Signifiance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Fixed effects Mentalistic Biased group	Estimate	Std Error	z value	Pr(> |z|)	Significance
(Intercept)	0.721	0.425	1.698	0.089	.
MeanPupilDiameter	0.778	1.962	0.397	0.691	
RobotBehaviour	−0.549	0.172	−3.180	0.001	**
MeanPupilDiameter:RobotBehaviour	−4.453	2.575	−1.730	0.083	**

Table C.3. Results of Generalized Linear Mixed Effect Model on the Mentalistic Biased sample. Signifiance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Fixed effects	Estimate	Std Error	df	t value	Pr(> |z|)	Significance
(Intercept)	1.296	0.046	$2,458.943$	2.808	< 0.001	***
z_ISS^2	−0.146	0.015	$141,999.699$	−9.737	< 0.001	***

Table C.4. Results of Linear Mixed Effect Model on the total sample. z_ISS^2 : z corrected quadratic InStance Score. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Appendix D

Supplementary Materials Chapter 8

D.1 Inclusion Criteria for data collection

Participants were selection according to the following criteria:

1. fluency and high proficiency in English;
2. previous countries that were included in other cultural studies in HRI (Lim et al., 2020);
3. availability of a sufficient number of participants that would fulfil point 1 and 2 on the online platform used to collected data (Prolific.co). This criterion, for example, lead to the exclusion of Japanese and Chinese participants as the first country did not have (at the time of data collection) a sufficient number of participants registered, and the latter did not have enough participants fluent in English.

D.2 Path Model Parameters

	label	est	se	z	p	CI (lower)	CI (upper)	std (lv)	std (all)	std (nox)
HRIES_SOC	CV_COLL	0.078	0.039	1.976	0.048	0.002	0.154	0.078	0.086	0.086
	CV_LONG	0.076	0.043	1.786	0.074	-0.007	0.162	0.076	0.085	0.085
	CV_MAS	0.026	0.044	0.601	0.548	-0.063	0.109	0.026	0.029	0.029
	CV_POW	0.034	0.136	0.251	0.802	-0.226	0.427	0.034	0.040	0.040
	CV_UNC	0.041	0.044	0.937	0.349	-0.044	0.127	0.041	0.045	0.045
	CV_OTHER	0.035	0.041	0.860	0.390	-0.043	0.117	0.035	0.039	0.039
HRIES_AN	CV_COLL	0.120	0.044	2.737	0.006	0.034	0.206	0.120	0.123	0.123
	CV_LONG	0.038	0.048	0.787	0.431	-0.050	0.138	0.038	0.039	0.039
	CV_MAS	0.113	0.055	2.047	0.041	-0.019	0.198	0.113	0.116	0.116
	CV_POW	0.023	0.261	0.088	0.930	0.007	0.833	0.023	0.025	0.025
	CV_UNC	-0.024	0.046	-0.520	0.603	-0.116	0.065	-0.024	-0.024	-0.024
	CV_OTHER	-0.029	0.045	-0.660	0.509	-0.112	0.063	-0.029	-0.030	-0.030
HRIES_AG	CV_COLL	0.148	0.037	4.009	$< .001$	0.075	0.219	0.148	0.162	0.162
	CV_LONG	0.060	0.045	1.328	0.184	-0.026	0.153	0.060	0.065	0.065
	CV_MAS	0.048	0.043	1.119	0.263	-0.045	0.125	0.048	0.053	0.053
	CV_POW	-0.030	0.153	-0.199	0.842	-0.189	0.455	-0.030	-0.035	-0.035
	CV_UNC	0.009	0.047	0.194	0.847	-0.083	0.100	0.009	0.010	0.010
	CV_OTHER	0.008	0.038	0.203	0.839	-0.066	0.085	0.008	0.008	0.008
HRIES_DIST	CV_COLL	0.029	0.048	0.611	0.541	-0.068	0.121	0.029	0.030	0.030
	CV_LONG	-0.087	0.054	-1.625	0.104	-0.182	0.028	-0.087	-0.088	-0.088
	CV_MAS	0.074	0.060	1.227	0.220	-0.082	0.153	0.074	0.075	0.075
	CV_POW	-0.028	0.364	-0.077	0.939	-0.045	1.003	-0.028	-0.030	-0.030
	CV_UNC	-0.021	0.051	-0.412	0.680	-0.126	0.073	-0.021	-0.021	-0.021
	CV_OTHER	-0.050	0.046	-1.071	0.284	-0.130	0.050	-0.050	-0.050	-0.050
IST	CV_COLL	0.004	0.039	0.098	0.922	-0.073	0.082	0.004	0.004	0.004

	label	est	se	z	p	CI (lower)	CI (upper)	std (lv)	std (all)	std (nox)
	CV_LONG	-0.091	0.046	-1.997	0.046	-0.176	0.002	-0.091	-0.092	-0.092
	CV_MAS	0.056	0.045	1.268	0.205	-0.041	0.136	0.056	0.057	0.057
	CV_POW	0.024	0.157	0.155	0.877	-0.157	0.507	0.024	0.026	0.026
	CV_UNC_Z	0.009	0.048	0.191	0.849	-0.085	0.102	0.009	0.009	0.009
	CV_OTHER	-0.008	0.040	-0.205	0.838	-0.082	0.073	-0.008	-0.008	-0.008
	HRIES_SOC	0.262	0.062	4.256	< .001	0.141	0.383	0.262	0.237	0.237
	HRIES_AN	0.106	0.052	2.046	0.041	0.003	0.206	0.106	0.105	0.105
	HRIES_AG	0.151	0.054	2.781	0.005	0.045	0.257	0.151	0.139	0.139
	HRIES_DIST	0.107	0.043	2.487	0.013	0.020	0.190	0.107	0.106	0.106
CV_COLL	CV_LONG	0.143	0.047	3.010	0.003	0.051	0.235	0.143	0.148	0.148
	CV_MAS	0.027	0.045	0.603	0.547	-0.060	0.117	0.027	0.028	0.028
	CV_POW	0.044	0.032	1.358	0.174	-0.002	0.119	0.044	0.043	0.043
	CV_UNC	0.119	0.043	2.743	0.006	0.035	0.204	0.119	0.125	0.125
	CV_OTHER	0.048	0.040	1.201	0.230	-0.030	0.126	0.048	0.050	0.050
CV_LONG	CV_MAS	0.083	0.044	1.898	0.058	-0.003	0.168	0.083	0.085	0.085
	CV_POW	-0.042	0.019	-2.176	0.030	-0.084	-0.010	-0.042	-0.041	-0.041
	CV_UNC	0.425	0.044	9.740	< .001	0.342	0.512	0.425	0.447	0.447
	CV_OTHER	0.235	0.047	4.944	< .001	0.143	0.330	0.235	0.244	0.244
CV_MAS	CV_POW	0.067	0.046	1.453	0.146	-0.035	0.131	0.067	0.066	0.066
	CV_UNC	0.058	0.041	1.406	0.160	-0.023	0.138	0.058	0.060	0.060
	CV_OTHER	-0.212	0.041	-5.229	< .001	-0.293	-0.134	-0.212	-0.219	-0.219
CV_POW	CV_UNC	0.002	0.010	0.194	0.846	-0.017	0.023	0.002	0.002	0.002
	CV_OTHER	-0.029	0.031	-0.948	0.343	-0.074	0.040	-0.029	-0.029	-0.029
CV_UNC	CV_OTHER	0.208	0.042	4.886	< .001	0.124	0.290	0.208	0.218	0.218
HRIES_SOC	HRIES_AN	0.464	0.039	11.890	< .001	0.384	0.535	0.464	0.566	0.566

	label	est	se	z	p	CI (lower)	CI (upper)	std (lv)	std (all)	std (nox)
	HRIES_AG	0.363	0.036	10.071	< .001	0.289	0.430	0.363	0.475	0.475
	HRIES_DIST	-0.178	0.042	-4.203	< .001	-0.265	-0.096	-0.178	-0.214	-0.214
HRIES_AN	HRIES_AG	0.343	0.036	9.423	< .001	0.269	0.412	0.343	0.413	0.413
	HRIES_DIST	-0.134	0.042	-3.173	0.002	-0.219	-0.052	-0.134	-0.148	-0.148
HRIES_AG	HRIES_DIST	-0.072	0.037	-1.963	0.050	-0.145	-8.553e-4	-0.072	-0.086	-0.086
IST	VT	0.045	0.038	1.166	0.244	-0.032	0.119	0.045	0.050	0.050
HRIES_SOC	HRIES_SOC	0.755	0.049	15.311	< .001	0.654	0.846	0.755	0.970	0.970
HRIES_AN	HRIES_AN	0.891	0.043	20.605	< .001	0.794	0.964	0.891	0.965	0.965
HRIES_AG	HRIES_AG	0.775	0.047	16.414	< .001	0.674	0.861	0.775	0.960	0.960
HRIES_DIST	HRIES_DIST	0.920	0.046	19.823	< .001	0.814	0.996	0.920	0.979	0.979
IST	IST	0.796	0.044	17.915	< .001	0.694	0.868	0.796	0.836	0.836
CV_COLL	CV_COLL	0.965	0.060	16.134	< .001	0.849	1.084	0.965	1.000	1.000
CV_LONG	CV_LONG	0.961	0.071	13.523	< .001	0.827	1.106	0.961	1.000	1.000
CV_MAS	CV_MAS	0.979	0.052	18.739	< .001	0.877	1.082	0.979	1.000	1.000
CV_POW	CV_POW	1.056	0.996	1.059	0.289	0.038	3.073	1.056	1.000	1.000
CV_UNC	CV_UNC	0.943	0.058	16.192	< .001	0.832	1.059	0.943	1.000	1.000
CV_OTHER	CV_OTHER	0.958	0.061	15.762	< .001	0.841	1.078	0.958	1.000	1.000
VT	VT	1.001	0.071	14.122	< .001	0.867	1.140	1.001	1.000	1.000

Table D.1. Parameter Estimates Path Model. All data were z-score corrected.

D.3 Model Parameters of Post-Hoc model comparison

No indirect effect of Collectivism on the HRIES mediated by the IST (bootC.I. included 0, see Table D1).

Table D.2. Indirect effects Post-Hoc model comparison

									95% C.I.	
					Estimate	Std. Error	z-value	p	Lower	Upper
CVS_COLL	→	IST	→	HRIES_SOC	0.017	0.013	1.308	0.191	-0.008	0.043
CVS_COLL	→	IST	→	HRIES_AN	0.016	0.012	1.302	0.193	-0.007	0.041
CVS_COLL	→	IST	→	HRIES_AG	0.014	0.011	1.301	0.193	-0.007	0.037
CVS_COLL	→	IST	→	HRIES_DIST	0.002	0.003	0.768	0.443	-0.002	0.013

Note. Delta method standard errors, bias-corrected percentile bootstrap confidence intervals, ML estimator.