UNIVERSITY OF GENOVA

PHD PROGRAM IN BIOENGINEERING AND ROBOTICS

The distracted robot: what happens when artificial agents behave like us

by

Davide Ghiglino

Thesis submitted for the degree *Doctor of Philosophy* (33° cycle)

December 2020

Agnieszka Wykowska	Supervisor
Cesco Willemse	Co-Supervisor
Giorgio Cannata	Head of the PhD program

Thesis Jury:

Carmen Usai, Università degli Studi di Genova	Internal examiner
Eva Wiese, George Mason University	External examiner
Thierry Chaminade, Université Aix-Marseille	External examiner

Dibris

Department of Informatics, Bioengineering, Robotics and System Engineering

"*If you're not failing every now and again,*
it's a sign you're not doing anything very innovative"

- Woody Allen

To my supervisor, my colleagues, my family, and friends,
for being my crutch when I failed and my pride when I succeeded

Declaration

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

Davide Ghiglino
December 2020

Acknowledgments

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 authors. The European Commission or its services cannot be held responsible for any use that may be made of the information it contains.

Abstract

In everyday life, we are frequently exposed to different smart technologies. From our smartphones to avatars in computer games, and soon perhaps humanoid robots, we are surrounded by artificial agents created to interact with us. Already during the design phase of an artificial agent, engineers often endow it with functions aimed to promote the interaction and engagement with it, ranging from its "communicative" abilities to the movements it produces. Still, whether an artificial agent that can behave like a human could boost the spontaneity and naturalness of interaction is still an open question. Even during the interaction with conspecifics, humans rely partially on motion cues when they need to infer the mental states underpinning behavior. Similar processes may be activated during the interaction with embodied artificial agents, such as humanoid robots. At the same time, a humanoid robot that can faithfully reproduce human-like behavior may undermine the interaction, causing a shift in attribution: from being endearing to being uncanny. Furthermore, it is still not clear whether individual biases and prior knowledge related to artificial agents can override perceptual evidence of human-like traits.

A relatively new area of research emerged in the context of investigating individuals' reactions towards robots, widely referred to as Human-Robot Interaction (HRI). HRI is a multidisciplinary community that comprises psychologists, neuroscientists, philosophers as well as roboticists, and engineers. However, HRI research has been often based on explicit measures (i.e. self-report questionnaires, a-posteriori interviews), while more implicit social cognitive processes that are elicited during the interaction with artificial agents took second place behind more qualitative and anecdotal results. The present work aims to demonstrate the usefulness of combining the systematic approach of cognitive neuroscience with HRI paradigms to further investigate social cognition processes evoked by artificial agents.

Thus, this thesis aimed at exploring human sensitivity to anthropomorphic characteristics of a humanoid robot's (i.e. iCub robot) behavior, based on motion cues, under different conditions of prior knowledge. To meet this aim, we manipulated the human-likeness of the behaviors displayed by the robot and the explicitness of instructions provided to the participants, in both screen-based and real-time interaction scenarios. Furthermore, we explored some of the individual differences that affect general attitudes towards robots, and the attribution of human-likeness consequently.

Index

SECTION I - INTRODUCTION[1]

[1] Part of this section has been published as Ghiglino, D., & Wykowska, A. (2020). When robots (pretend to) think. In Artificial Intelligence (pp. 49-74). Mentis.

1.1 Mental activity and mindreading

Mental activity is a fascinating mystery that humans have tried to unveil since the very beginning of philosophy. According to the Cambridge dictionary the verb "to think" refers to activities such as "to consider", "to meditate", "to remember", "to wish", "to desire" or "to intend". There is one common feature that links all these verbs: reference to the content, the direction toward an object (Brentano, 1874), or the feature of "aboutness" (Burgos, 2007; Millikan, 1984) – something that has been highlighted by philosophers, as the key feature of mental states (Brentano, 1874), reflected in language structures of "propositional attitudes" (Russell, 1905) inherently containing "aboutness" (I believe that… I worry that…I hope that…). According to Searle (1980), what differentiates a natural human thought from a computer program can be illustrated by the famous "Chinese room" thought experiment: although a computer might be able to perfectly operate on the syntax and symbols of a language, it will not understand the meaning, the "aboutness" related to all mental processes of humans.

In terms of etymology, in languages originated from Latin, the verb "to think" is derived from "pensum", which used to be the quantity of wool assigned to the weavers in Roman times. It described a simple material, which could be transformed into something complex through work and dedication. Whereas for Latins the activity of thinking was connected with practical and manual activities, ancient Greeks had already developed and entrenched the concept of "νοῦς" (noûs), which referred to the intellectual perception of reality from the "thinker" perspective. Since Anaxagoras, philosophers, psychologists, neuroscientists, and researchers tried to explain what it means "to think" and what are the underpinning mechanisms involved in such an activity. Reviewing theories of thinking is beyond the aims of this chapter, but it is interesting to point out that after more than two thousand years, the debate on the materialistic or abstract nature of thinking is still heated. According to neuroscientific accounts, the activity of thinking is considered to be the result of the elaboration of information collected by our sensory system. Therefore, from a biological point of view, the mind itself can be treated as a result of brain activity, which can be explained in terms of electrochemical signals. On the other hand, certain philosophical perspectives postulate that this explanation seems to be too reductive. Various degrees of mind-body dualism or non-reductionism have been proposed since Descartes. Fodor, for example, pointed out that it is impossible to infer high-level properties exhibited by a system just from knowing lower-level properties (Fodor, 1975). Similarly, Putnam claims that micro-properties of molecules or atoms are not sufficient to explain macro-properties of a behavior (Putnam, 1975). The debate about what constitutes thinking and whether the mind is

reducible to brain activity continues to be extremely prominent in the philosophy of mind, but it is outside the scope of this chapter.

Despite the lack of agreement about the nature of thinking, attribution of thoughts to others has undeniably practical consequences in everyday life. When we are interacting with others, we probably do not ponder over the etymology or the meaning of the concept of "thinking", but we still formulate hypotheses, predictions, expectations, and, more broadly, representations of the others' goals, desires, and intentions, and behaviors following from those. We "think" spontaneously about others' and our own mental states. In order to survive in the complexity of the world we are living, our species phylogenetically developed flexible strategies of understanding others' thoughts, strategies that adapt to the situation that is being experienced at a given moment and/or to the interaction partner. In the last fifty years, growing interest in social cognition led researchers to investigate new questions related to the attribution of thoughts to others. Social cognition refers to the cognitive processes that underlie social interactions, including perceiving, interpreting, and generating responses to intentions, disposition, and behaviors of others (Green, Horan, 2010). The ability to infer and predict intentions, thoughts, desires, intuitions, behavioral reactions, plans, and beliefs of other people is a crucial facet of social cognition (Frith, Frith, 2012) and is often referred to as "mindreading" or "mentalization". Mindreading is a concept developed to describe the process of understanding or predicting other people's behaviors in terms of their thoughts, feelings, wishes, beliefs, or desires. It represents all the processes that allow us to understand behavior in terms of underpinning internal states and making behavior meaningful (Fonagy, 2018). This ability enables individuals to assess or understand other's mental states in a specific situation and, thus, to interpret and anticipate their behaviors (Bèrubè, 2013).

Mindreading is activated through the attribution of mental states to others, which may differ from one's own (Korkmaz, 2011; Sabbagh, 2004). Some authors postulated that such ability is driven by innate neural mechanisms dedicated to mental state reasoning (Fodor, 1983; Sholl, Leslie, 1999; Karmiloff-Smith et al., 1995). Other authors claimed that that reasoning about others' mental states requires the capability to empathize with someone else, and that this depends on personal experience (Lillard, Kavanaugh, 2014; Taylor, Carlson, 1997). The study of these mechanisms constantly requires a revision of models at multiple levels, which are valuable as far as each captures different phenomena, and no single level can be eliminated (Schaafsma et al., 2015). Despite the differences in such models, the ability to read another agent's mind seems to be recruited every time we predict the behavior of other agents. This requires tracking and representing: (1) the agent that displays the behavior, (2) particular bits of information from the environment that surrounds the agent, and (3) the

relation between the two (based on cues relevant to the mental state of the agent) (Baron-Cohen, 2013). Then, we use those representations to predict and/or interpret the action(s) of the agent. It is now widely recognized that mindreading presents universal features and its pattern of development in humans is remarkably similar across different cultures. For example, it has been shown that understanding false belief emerges in children of the Baka, a preliterate tribe in Cameroon (Avis and Harris, 1991), at a similar age to children living in the Western world. A meta-analysis of children's false belief studies provides parallel developmental trajectories of mindreading abilities in Chinese and North American children, coupled with differences of approximately two years in acquisition timing across communities (Liu et al., 2008). These data support the idea that mindreading abilities constitute a human "universal". However, specific, experiential factors (i.e. peculiar educational practices, social habits) may impact temporal aspects of this ability (Brüne & Brüne-Cohrs, 2006). According to some anthropologists, primitive forms of mindreading abilities can be retraced in other mammals. Jolly and Humphery, in 1966 and 1976,theorized one of the best-known evolutionary hypotheses for mindreading (i.e. "Machiavellian intelligence hypothesis") according to which: "the social environment might have been a significant selective pressure for primate intelligence" (see Byrne & Whiten, 1997). Primates show a "surplus" of intelligence that overcomes the immediate survival needs, like eating, avoiding predators, feeding offspring, etc. According to the Machiavellian intelligence hypothesis, this surplus intelligence should be advantageous for social manipulation, deception, and cooperation (Speber, 2002). This suggests a slightly independent evolutionary history of mindreading abilities from that of language.

Notably, more recent computational models of mindreading and language suggest a strong co-evolution of such abilities. If, from one side, language relies on mindreading for recognizing communicative intentions, mindreading abilities profit from language in turn, for expressing mental states explicitly, and for transmitting the knowledge of such mental states to others. Given this interdependence, it has been hypothesized that language and mindreading have co-evolved, due to the social pressure characterizing most mammals' environment (Kapron-king, Kirby, & Woensdregt, M, 2020). Indeed, primates are essentially social animals, and group living certainly confers adaptive advantages on the individual such as better protection from predation and food sharing (Alexander, 1987). On the other hand, group living incurs the cost of directly competing for resources and sexual partners. This situation may have created specific selective pressures in primates to evolve 'social intelligence' (Whiten, 2000). Crucial in the context of primate group living with strong mutual dependency and complex interactions is the ability of individuals to identify others who cooperate and, even more importantly, who try to defect (Brune, 2006). That is, if an individual trusts that

cooperation will be reciprocated, cheating could be an even more successful strategy for another subject. Thus, to counteract cheating one must be able to detect deception (Trivers, 1971). Premack and Woodruff (1978), studying chimpanzee social abilities, inferred that an agent display mindreading abilities every time it imputes mental states to himself and others. A system of inferences of this kind is properly viewed as a theory (aka "Theory of Mind", ToM), because such states are not directly observable, but used to make predictions about the behavior of others.

Nowadays, this evolutionary mindreading model appears to be limited (Martin, 2016), especially because it does not give any clues about the nature of the construct. In 1978, Dennett suggested the use of "false-belief" tasks to study the mechanisms underlying mindreading. Developmental psychologists rapidly adopted mindreading models to explain the acquisition of a 'mental perspective' in children (Wimmer and Perner, 1983). Some authors hypothesized that the capacity to understand that others can hold beliefs that might differ from one's own is acquired at age 5–6 years (Perner and Wimmer, 1985). This acquisition includes the capacity to distinguish between mental and physical objects, interpret the gaze of others, and understand their mental states, desires, and intentions. More recent studies showed that precursors of these abilities may be acquired around 2-3 years (e.g. Perner, 2001). According to Goldman (1992), pretense and pretend could be considered as key factors for the development of such skills, and provide useful insights to the neuropsychological approach to mindreading. Indeed, pretense and pretend plays enable the infant to decouple and construct secondary representations, and has been renamed as 'meta-representation' (Leslie, 1987). From a neuropsychological perspective, the brain may possess a "mind-reading" system, composed of an "Intentionality Detector" (ID), an "Eye Direction Detector" (EDD), and a "Shared Attention Mechanism" (Baron-Cohen, 1995). The ID is proposed to be a perceptual device that interprets 'primitive volitional' mental states such as goals and desires (i.e. "the agent wants"). These are seen as basic mental states required for making basic sense of the movements of all organisms in the environment (Goldman, 2006). The second innate mechanism (EDD) has three basic functions of detecting the presence of eye-like stimuli, computing whether eyes are directed towards it or towards another direction, and inferring from the observation that the organism's eyes are directed at something else that it actually sees something (Macrae, 2002). Finally, the SAM is considered to be a higher-order skill that allows the individual to form what is called 'triadic representation' (Langavant, 2011). Triadic representations conceptualize relations between an Agent, the Self, and an Object (which can be another agent). The SAM builds triadic representations by perceiving the perceptual state of another agent and computes shared attention by comparing another agent's perceptual state with the self's current perceptual state (Nader-Grosbois, 2011). These claims

go well beyond the tepid functional proposal made by Premack and Woodruff (1978). Problems that before 1978 would have been deemed as falling under categories such as metacognition, attribution theory or "Piagetian" developmental studies are now being called "mindreading research" (Flavell and Miller, 1998: 853); even putative deficits in mindreading have become an explanatory concept for autism, schizophrenia, and related disorders (Baron-Cohen, Leslie and Frith, 1985; Frith and Frith, 2003).

1.2 Ascribing a "mind" to artificial agents

1.2.1 The intentional stance and mindreading

The advent of new technologies - seemingly smart artificial agents, such as Apple's Siri, Amazon's Alexa or Google Home assistant is giving researchers new tools to test mindreading models, pushing the cognitive flexibility of the human social brain from the natural domain to towards the artificial. Such technologies seem to display a certain degree of "artificial thinking" that resembles (but only to a very limited extent) human thought. But do artificial agents lead people to treat them like "thoughtful" agents? Artificial intelligence that these agents seem to display might enhance humans' tendency to anthropomorphize agents for which they lack specific knowledge and, as a cascade effect, this might lead towards maindreading. Whether it is possible, or desirable, to attribute mental states toward artificial agents is still an open question. However, addressing this point is important not only for theoretical purposes but also for practical and ethical consideration about the design and development of social artificial agents.Already in the seventies, Dennett proposed an account addressing how humans understand and predict the behavior of various systems (Dennett, 1971). The author postulated the existence of different strategies, called stances. According to Dennett, when we are interacting with inert objects, such as a leaf falling from a tree, we explain the behavior we might see in terms of physical laws. A leaf falling to the ground is just an effect of the gravity pull. Thus, we adopt the physical stance to understand the behavior of a physical system and predict its consequences. The very same model can be applied to a ball rolling down the street, whose behavior is explained in terms of acceleration, gravity pull, attrition, etc. However, when the object we are perceiving is a complex artifact, such as an airplane landing at the airfield or a car moving through a street, relying on physical information might not be an efficient (or accessible) strategy to explain and predict the behavior we are observing. To predict efficiently the behavior displayed by such complex entities we might need to rely on their functionality and design, using what Dennett

calls the design stance. The fact that a car is moving is not explicable just in terms of inertial motion, but it becomes clearer as we start thinking about the design of the motor or the brake. When we are interacting with very complex agents, such as humans, neither the physical nor the design stance can provide an efficient model of explanation. For example, if a person talking to us and repetitively moves her head in the direction of a clock, neither the physical proprieties of the agent nor the anatomy underpinning the neck movement might be a reliable source of information to make efficient predictions regarding the person's behavior. In the depicted scenario, the behavior displayed by the agent might be due to her boredom or her intention to leave us as soon as possible. Therefore, to understand and predict her behavior, we need to include in our model a representation of the agent's internal states, adopting what Dennett calls intentional stance. Adopting the intentional stance means to treat the acting agent as an intentional being, who is aware of their behavior and who acts to maximize the likelihood of achieving a pre-set goal. In other words, adopting this attitude allows us to treat the behavior as a consequence of a mental act, which is directed toward an object. Here, the terms "intentionality" is related to Brentano's conceptualization, since adopting an intentional stance implies the understanding of an immanent interconnection between mental phenomena and objective contents towards which the phenomena are directed (the "aboutness") (see Mayer-Hillebrand, 1951; Jacquette, 1991). It is important to point out that the adoption of such strategies does not necessarily imply the ascription of a mind to the perceived agent. For example, when we are interacting with agents that display artificial intelligence, the most efficient way to interact with them might be the adoption of the intentional stance. When we need to ask our smart assistant about the weather condition or the latest news, the easiest way to communicate our request is to treat it as if it was an intentional being. This does not imply that we represent it as a "true believer" (Dennett, 1981), equipped with mental states. The intentional stance seem to be adopted by default when interacting with other humans, but it might allow individuals to deal with unknown entities as well, or with artifacts whose behavior is ambiguous or impregnable. This tendency to attribute intentionality towards unknown entities might depend on the natural tendency humans have to attribute anthropomorphic traits towards entities for which they lack a specific knowledge. Indeed, the adoption of these three stances depends on the knowledge, the experience, and the representation that humans build around the acting agent. This means that when such knowledge and representation changes, also the strategy we are adopting toward the agent might change. This flexibility is essential since our cognitive system needs to constantly deal with an overwhelming amount of information coming from the environment, relying on a limited pool of resources. Therefore, when a certain explanation model is inefficient, resources need to be allocated toward another model that can grant

rapid adaptation of the entire system. On the other hand, when a simpler and less cognitively demanding model is efficient to explain and predict an agent's behavior, resources shall be shifted toward this latter model.

Therefore, we can define the "intentional stance" as the disposition to treat an entity as a rational agent, possessing mental states, which can be used to interpret and predict the behavior it displays (Frith & Frith, 2006). In this sense, the intentional stance can be considered a crucial component of mindreading. Indeed, mindreading is an extremely complex cognitive function that refers to the entire process of ascribing and reasoning about an agent's mental states. Dennett himself, in 1987, described the intentional stance as follows: "Here is how it works: first you decide to treat the object whose behavior 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". In other words, the behavior of an agent is the principal source of information for making attributions about their mental states, but such information is only used in this way when we adopt an intentional stance. This implies that the intentional stance could be adopted towards any artifact as long as it fulfils the stance's assumption that it is a "rational" agent. Therefore, the main requirement of a system for being treated as intentional is displaying a behavior that could be well explained by adopting the intentional stance. This does not necessarily require perceiving the system itself to have mental states (Marchesi et al., 2019). Thus, we might even consider the intentional stance as one of the key "prerequisites" for mindreading (Abu-Akel et al., 2020). Specifically, while the intentional stance corresponds to the general attitude that an individual takes towards explaining others' behaviors, mindreading refers to explaining a particular behavior in a particular context, with reference to the underpinning mental states.

1.2.2 Empirical approaches to studying intentional stance towards artificial agents

As mentioned above, new technologies provided researchers with new tools to explore in more complex interactive scenarios the question of attribution of intentionality to artificial agents. In particular, robots are of interest, as they are embodied, thus introducing physical (and potentially social) presence in the environment, they can manipulate the environment, thus introducing artificial agency. Yet, they still offer excellent experimental control, as they can be programmed to behave in

an exactly specified and controlled way. In this context, it is therefore promising to examine whether and when humans adopt an intentional stance to robots.

Some authors claim that implicit social signals lead humans to perceive automatically an agent behavior as the reflection of mental operations (for a review see Frith, Frith, 2008). We speculate that designing artificial agents that can closely imitate humans' behavior, even at the level of implicit social signals, would induce humans to perceive them as intentional agents. Some recent studies suggest that the perception of the intentionality of action might be an automatic and immediate process. In 2007, Terada et al. investigated whether such an automatic process happens when participants are exposed to reactive movements of non-humanoid artificial agents (a chair or a cube with wheels). In their experimental setup, participants were observed while they were freely interacting with an artificial agent in an empty room. The movement of the artificial agent was remotely controlled and could be periodic or reactive. While in the first case, the movement of the artificial agent was random, in the reactive condition the movement was happening immediately after the movement of a participant. At the end of the interaction, participants were asked to fill questionnaires in order to evaluate whether the movement of the agent was perceived as intentional or not. Results showed that the physical appearance of the agent (chair vs cube) did not affect the attribution of intentionality, while the movement profile (periodic vs reactive) played a significant role in such attributions. In their experiment, authors demonstrated that when the behavior of an artificial agent is contingent on human behavior, participants tend to evaluate it as goal-directed. In their study, the authors claim that enhancing the sense of intentionality perceived toward a robot is the key to smoothing communication between these agents and humans. The same authors demonstrated in a follow-up experiment (Terada et al., 2008) that other factors might affect intentionality attribution toward artificial agents. For example, context and participant's expectations might have an impact on the attribution of intentionality. In line with this, Wiese, Wykowska et al. (2012) showed that what mattered in evoking engagement in joint attention with a robot avatar was the belief that participants held regarding its agency, rather than the actual physical characteristics of the agent. The authors manipulated the instruction that they have given to the participants of their study. In one condition, participants were instructed that the robot face they were observing was controlled by a human, while in another condition, they received an instruction that it was just a pre-programmed automaton. Joint attention (measured in the form of gaze cueing effects, Friesen & Kingstone, 1998; Driver et al., 1999) was evoked to a larger extent when participants believed the robot's behavior was controlled by a human, relative to the belief that it was only preprogrammed. Interestingly, the physical features of the observed avatar were identical across conditions. In a

follow-up EEG study (Wykowska, Wiese, et al., 2014), the authors showed that this effect is reflected at early stages of attentional processing while Özdem et al. (2017) identified – in an fMRI study – neural correlates of such belief manipulation.

In a recent study, Thellman et al. (2017) asked a large number of people (N=93) with various backgrounds to interpret and explain behaviors displayed by two different agents (i.e. a human and a humanoid robot) involved in every-day activities (such as cleaning the floor or cooking). Specifically, the authors presented to their participants' pictures depicting the two agents doing the same actions and proposing, for each scenario, a possible explanation of the behavior. All explanations were referring to mental states both when the depicted agent was a robot, and when the depicted agent was a human. After observing each scenario, participants were asked to rate the explanation in terms of plausibility and to rate the general behavior displayed by the agent in terms of intentionality, desirability, and controllability. Results of their experiment showed that there was no difference between the two agents in terms of plausibility, intentionality, and desirability ratings between the conditions. However, when the behavior was displayed by the human agent, participants evaluated that the behavior was controlled by the agents themselves. On the contrary, when the behavior was displayed by the artificial agents, the behavior seemed controlled by external causes or was evaluated as programmed. Results of this experiment suggest that humans might adopt the intentional stance to some extent towards artificial agents as well. Yet, the existence of some differences in ratings supports the idea that different strategies are used to interpret artificial agents and human behaviors.

Recently, Kamide et al. (2015) used the "Anshin" questionnaire to explore the reactions of more than nine hundred Japanese participants to videos depicting several robots. They discovered that aspects like perceived humanness, proficiency, and comfortability with a robot changed from robot to robot and from behavior to behavior. Specifically, they noticed that usually, the more the humanoid resembled a human in physical appearance, the higher participants were evaluating the comfort of the interaction and the proficiency of behaviors. In some situations, they found that exposure to more machine-like robots made humans feel stressed, but their results confirm that attributions towards artificial agents can fluctuate according to several factors. It is important to point out, however, that both Thellmann and colleagues as well as Kamide and colleagues used pictures or videos as stimuli in their experiments. Such "offline" experiments might not be sufficient to capture the social mechanisms involved during "live" interactions. In this context, it is crucial to address the question of whether behavioral parameters of an observed agent in more "live" naturalistic protocols affect the attribution of intentionality. In line with this hypothesis, Weiss and Bartneck (2015) pointed out that evaluations of likeability, anthropomorphism, and human-likeness of a robot using their Godspeed

questionnaire are dramatically influenced by the interaction with a robot rather than by its physical appearance.

In a recent study, Wykowska et al. (2015) designed two experiments that might help to clarify this aspect. In the first experiment, the authors showed their participants a non-humanoid robot (mechanical arms attached to a picture depicting the face of a human). Participants were asked to observe the behaviors of the robot, trying to identify which one of those behaviors was pre-programmed and which one was remotely controlled by a confederate located in another room. In a second experiment, the same design was adopted, but the artificial agent was a humanoid robot NAO (SoftBank Robotics). In addition to the explicit discrimination task, the authors implemented in both experiments a visual letter discrimination paradigm, to collect an implicit measure of engagement. While participants were observing the behavior of the robot and rating its agency, they were also asked to discriminate between two letters ("T" or "F") appearing on a screen located behind the robot. Results of both experiments suggest that humans are quite sensitive in detecting subtle differences between movements displayed by an artificial agent. Indeed, participants were able to discriminate accurately when a movement was remotely controlled by a confederate and when it was pre-programmed. Furthermore, this study provides information about the implicit aspects of social interaction with artificial agents. The results suggested that when artificial agents show a certain degree of anthropomorphism, implicit social processes of engagement (in joint attention, in the case of this study) are activated.

The recruitment of mindreading abilities to understand and predict the behavior displayed by an artificial agent seems to be modulated also by the knowledge that individuals possess regarding that specific agent. From an anthropological perspective, the limited knowledge our ancestors developed around natural phenomena, like the rain or the earthquakes, might be the reason why they attributed such phenomena to the "God's mind". When knowledge has become more developed and accessible, such interpretations have been substituted with physical explanations of the phenomena. Thus, it seems that when we are interacting or observing the behavior of an unknown (or not understandable) entity or phenomenon, we refer to the intentional model, as this might be the most available and default model that we possess. In fact, we are probably most experienced with intentional explanations, rather than scientific or technical explanations of various systems we observe in the environment. However, when our knowledge about an entity changes, the strategy we adopt to understand its behaviors changes as well. Similar reasoning can be applied to robots: if a person with no formal education who had never seen a robot before is exposed to a human-like robot, s/he might be more likely to adopt the default intentional stance towards the robot, as compared to a

person who is experienced with robotics, or has acquired sufficient level of general education (physics and engineering included). This suggests that it is important to take into consideration levels of education and experience with robots and perhaps also various age groups when testing attitudes towards robots.

Several factors might modulate individuals' tendency to ascribe a mind towards robots. Indeed, in human-human interactions, countless factors affect the way we perceive and understand our conspecifics' behavior, from personality traits to mood (Lopes et al., 2005; Cozolino, 2014). For example, it has been widely demonstrated that introverted people show difficulties in communicating (Sallinen-Kuparinen et al., 1991; Schoemaker, Kalverboer 1994) and that people in a depressed mood tend to avoid social interaction (Young, Leyton 2002; Erber et al., 1996). Internal dispositions and states seem to be strictly connected with the attitude we show during our everyday life interactions. Therefore, we can speculate that as it happens for mindreading abilities (Platek et al., 2003; Nichols, Stich, 2003), intentional stance might be influenced also by self-awareness and self-reflective abilities. Few studies demonstrated that self-processing abilities affect mental state attribution to others, and, specifically, that psychiatric populations with deficits in self-awareness show impairments in the attribution of mental states to other agents (Williams, 2010; Moriguchi et al., 2006). Studying whether such influences affect interactions with robots as well might help in the development of new explanatory models about human social abilities. Although human-computer interaction (HCI) and human-robot interaction (HRI) studies are focused on designing artificial interfaces that can satisfy the needs of humans, understanding the way humans interact with artificial agents can also improve our knowledge about human cognition in general. Theories and models of human cognition can be tested using such technologies, offering new challenges (Dautenhahn, 2007). Cognitive architecture models, for example, are largely used to develop artificial intelligence (Kelley, 2006; Metta et al., 2010). A case of particular interest might be the humanoid robot iCub, mentioned in the studies that we reported in previous sections. Its developers and designers took inspiration from models of cognitive architectures and implemented similar functions in the robot (Vernon et al., 2007). Specifically, the synthetic architecture that is fitted to the iCub is composed of components such as a distributed multi-functional perceptual-motor network, a system of inhibitory modulation, and a system of action simulation. These systems originated from research and models proposed by neuroscientists, such as the presence in the brain of specific networks responsible for action selection (Chevalier, Deniau, 1990; Deniau, Chevalier 1985) or models about action simulation (Shanahan, 2005). Implementing a robot with such functions help researchers to clarify the ecological validity of their models, providing new insights into the functioning of the human brain.

1.2.2.1 Impact on Human-Robot Interaction

It is in this context that our theoretical questions about adopting the intentional stance, characteristics of robot behavior, or inter-individual differences need to be addressed. One of the most interesting questions is whether endowing robots with intentionality is crucial for evoking adoption of the intentional stance, or rather whether emulation of human-like behavior is sufficient. This question is analogous to the issue of the Turing test (Turing, 1950): Is it possible to produce responses in the Turing test that would make participants believe they are interacting with an intelligent agent, even though it is just an algorithm? Or rather, is it indispensable to have genuine intelligence in order to pass the Turing test. In simple words, is it possible to "fake" intelligence or intentionality thanks to very well-designed behavioral characteristics? At present, researchers are still far from developing artificial agents that can be perceived as human-like. Although some technologies provided with artificial intelligence might be able to pass the Turing test, humans are still fully aware of the boundaries between artificial agents and living beings (Kahn et al., 2006). The open question is whether it is only a matter of technological advances to "fake" intentionality, or is it in principle impossible.

Evoking the adoption of the intentional stance towards artificial agents might be desirable in certain applications. For example, pathological traits, such as autistic traits, severely compromise social interaction and a multitude of studies demonstrated that individuals diagnosed with autism spectrum disorder (ASD) are often unable to deal with the complexity of human social interaction (Wing, Gould, 1979; Frith, 2003; McConnell, 2002). Specifically, children diagnosed with ASD often display less interest in behaviors aimed at generating and maintaining social interaction (Vivanti, Nuske, 2017). The severity, onset time, and specificity of such deficits captured the attention of clinicians and neuroscientists, stimulating the development of innovative evidence-based therapeutic approaches (Yates, Couteur, 2016; Fernandez et al., 2018). Recent literature suggests that robots might be efficiently used as clinical tools to enhance the social competencies of children affected with social impairments (Diehl et al, 2012; Feil-Seifer, Mataric, 2010; Kajopoulos et al., 2015; Carlson et al., 2018). Several studies demonstrated that the advantages of using such artificial agents rely on their physical appearance and the extensive control that clinicians and scientists can exert on their behaviors (Scassellati et al., 2012; Robins et al., 2005; Liu et al., 2008). Robots may represent a safe, predictable, and coherent environment to train social interaction skills in people that display difficulties in interaction with other humans (Dautenhahn, Werry 2004). It seems that artificial agents can significantly improve therapeutic efficiency in patients affected with social impairments. For

example, in 2013, Zheng et al. demonstrated that after only four sessions with a specific rehabilitation protocol carried out with a robot, children with ASD improve their social skills (Zheng et al., 2013). Concerns might be raised about the development of a certain social dependence that the patient can develop toward assistive technology. However, some studies demonstrated that learning acquired during the activities with robots is automatically transferred to everyday-life activities and interaction with other humans (Francois et al., 2009; Robins et al., 2005). We speculate that robot behaviors that evoke the adoption of intentional stance shall be even more efficient in training social skills that can be then transferred to human interactions. This is because adopting an intentional stance towards other humans is a natural process in the social interactions of typically developing individuals. Difficulty in effortless adopting of the intentional stance towards others is likely part of the problem in social skill deficiencies (Griffin & Dennett, 2008). Therefore, evoking an intentional stance in therapies with robots should have beneficial effects for transferring the skills trained with robots to interactions with humans.

Apart from the application of social robots to healthcare, social robots are supposed to have a potential application also in elderly care. Robots that are designed for elderly care provide support in everyday-life activities (eating, bathing, etc.), mobility, housekeeping, and monitoring the health condition of the user. Several researchers focused their attention on consequences caused by the introduction of assistive robots (Pollak et al., 2002; Graf et al., 2004) and assistive smart environments (Bahadori et al., 2003) in elderly care. Results showed that the introduction of robots in elderly care does not only improve the quality of life of people in terms of perceived autonomy but also in terms of psychological wellbeing (Broekens et al., 2009; Sharkey, Sharkey 2012). This is because such robots can provide social feedback to the user, and can be treated as “companions”. Companion robots seem to increase positive mood while reducing the perception of loneliness and stress (for a review, see Broekens et al., 2009 and Bemelmans et al., 2012). Results emphasize the usefulness of robots that can induce some degree of social attunement. These applications will be needed in the near future, considering that the elderly population is constantly increasing, while the availability of healthcare professionals is decreasing (WHO, Investing in the health workforce enables stronger health systems, in the Fact sheet. 2007: Belgrade, Copenhagen). To maximize the efficiency of healthcare professionals, repetitive and frustrating activities can be assigned to assistive robots. In parallel, this would allow lower institutionalization of medical care and higher autonomy for individuals. Also in the case of elderly care, robots that evoke adoption of intentional stance might prove more beneficial than those that not, due to the potential higher degree of bonding. However, some studies pointed out the risks that such social devices might entail. In the last decades, the main

application of robots was industry. Recently, however, domestic robots started entering our homes. Some authors question the safety granted by new technologies, in terms of producing reliable and non-harmful behavior (Denning et al., 2009; Sharkey & Sharkey, 2010), others point out the issues related to psychological risks, such as social isolation (Sparrow, 2006; Hampton et al., 2009) or addiction to technology (Veruggio, 2005). Finally, one of the most discussed controversial aspects related to robots is the issue of privacy. Privacy is an ancient concept, developed by ancient Greeks as a distinction between public and private life. Nowadays, privacy is conceived as the possibility of controlling one's data. Recently, privacy management has caught the attention of media and legislative authorities. although no robot or artificial intelligence was involved in the 2018 scandal of Facebook, Facebook admitted that information of eighty-seven million users was given to Cambridge Analytica without their permission. If we think about assistive and domestic robots, one of the necessary requirements for them to function adequately in their role is collecting, processing, and potentially recording data from users. Some of the commercial house assistants, like google home, need to save information gathered during the interaction with the user in the cloud. In the context of the intentional stance, perhaps also in this area, if robots induce adoption of the intentional stance, they might be more likely to collect more personal or intimate information from the user, relative to robots that would be treated only as mechanistic artifacts. This is because humans might be more inclined to reveal their intimate information towards agents that are perceived as intentional (to make it more evident: we are certainly more likely to confide in another human than in an artifact such as a coffee machine). Thus, designing robots that evoke the adoption of the intentional stance in users might have consequences for the type of data that is recorded and potentially stored. Companies producing robots for daily home use must guarantee privacy protection of such information, to avoid third parties accessing those data without the consent of the user.

In this context, it is crucial to inform users about all the aspects involved in interacting with technology. The more we aim to introduce technology in everyday life, the more we need to inform users about the potential risks, as well as benefits provided by having access to such tools. Giving as much information as possible to users, and educating them on how to use technology would on the one hand circumvent certain fears produced often by popular culture (e.g., the fear of robots "taking over the world") and would reduce the unnecessary hype related to AI and robotics that is based on misconceptions about what AI is capable of. Some authors, for instance, claimed that assistive robots are unethical since their effectiveness depends on deceiving users creating an illusion of companionship (Sparrow, 2002; Turkle et al., 2006; Wallach, 2009). However, it is crucial to avoid creating illusions and misconceptions related to new technologies. We argue that adopting an

intentional stance towards artificial agents might be beneficial in some contexts, and is not necessarily something to be afraid of, given how automatic and default this mechanism is (Heider, Simmel, 1944; Gray et al., 2007; Wiese, Metta, Wykowska, 2017). However, it is generally important to provide sufficient information about robot technologies to all users, to avoid potential risks of misconceptions, misuse of private data as well as unnecessary fears related to fantasies created by popular culture.

Studying the conditions and consequences of implementing human-like behaviors on artificial agents that can potentially induce the adoption of intentional stance is fascinating from a theoretical perspective, and extremely important for the future of our societies. From the theoretical perspective, it can be informative concerning the mechanisms of human social cognition – how flexible is our socio-cognitive system. How far can the human brain go in extending mechanisms of social cognition from natural to artificial agents? What consequences does adopting the intentional stance have for other (perhaps more implicit) mechanisms of social cognition? Can we develop models of intentionality for artificial systems? Is endowing artificial systems with intentionality necessary for evoking adoption of intentional stance or is it sufficient to emulate human-like behaviors?

From the perspective of applications of robots and future societal impact, it is important to discuss whether and when treating artificial agents as intentional systems is desirable. It is also important to keep in mind that proper information for the end-users regarding the capabilities of the artificial systems can avoid misconceptions and hype on the one hand and fears and risks on the other. All these issues need to be discussed not only among the scientific community but also with the general public. This would allow strengthening people's awareness about the technology, avoiding misuse and misjudgment.

1.2.2.2 The development of the InStance questionnaire[2]

In order to address the question of whether humans adopt the intentional stance toward a robot, Marchesi et al. recently created a tool (the Intentional Stance Questionnaire, ISQ) that should probe the adoption of intentional stance toward a specific robot platform, a humanoid robot (Marchesi et al., 2019). The aim of their study was twofold: (1) Developing a tool that would allow for measuring whether humans would adopt, *in some contexts*, the intentional stance toward a robot; (2) Exploring *if* humans *would* sometimes adopt the intentional stance toward robots. Each item of ISQ consisted of pictorial scenarios, depicting the iCub robot interacting with objects and/or humans. Each item included two sentences, in addition to the scenario. One of the sentences was always explaining

[2] Part of this paragraph was published as 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, 450.

iCub's behavior referring to the design stance (i.e., mechanistic explanation), whereas the other was always describing iCub's behavior referring to mental states (i.e., mentalistic explanation). Participants were asked to move the slider towards the sentence that better described the scenario according to them. Results of this study indicated that participants showed a slight bias toward the mechanistic explanation when they were asked to evaluate robot actions, which is in line with the essence of the concept of the intentional stance (Dennett, 1987). The authors speculate that, to understand the behavior 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). However, and interestingly, the design stance descriptions were not always chosen by their participants to explain the iCub's actions. This indicates that participants have (at times) also chosen mentalistic explanations of the given scenarios. Therefore, in principle, it might be possible to induce 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 behavioral characteristics (e.g., contingency of its behavior on participant's behavior, 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. Indeed, humans possess many types of vocabulary for describing 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 behavior, 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 the ISQ questionnaire, the robot was always the same, but the action and the environment/context around it changed across conditions, modulating participants' ratings. However, the ISQ ratings 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 such results, the authors argue that the adoption of mentalistic or mechanistic models to explain a robot's action do not rely only on intrinsic properties of the agent, but also on contingent factors (Waytz 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). The study proposed by Marchesi et al. showed that it is possible to induce adoption of the intentional stance toward the iCub robot at times, perhaps due to its human-like appearance. However, further research needs to explore what are the

exact factors (individual difference, cultural context, specific characteristics of robot appearance or behavior) that influence the adoption of intentional stance.

1.3 Rationale of the project

In short, individuals spontaneously generate representations of other agents' goals, desires, and intentions. This natural tendency enables individuals to understand human behavior in terms of underlying mental states, making the behavior meaningful (Fonagy, 2018). However, while interacting with artifacts (i.e. computers, smartphones, etc.), humans usually tend to rely more on their functionality and design rather than on their intentional states (Dennett, 1971). This hypothesis might not hold in the case of complex artificial agents, such as robots, which create the illusion of human-likeness, due to their physical appearance and due to the behavior that they can display. Indeed, the design of artificial agents able to display human-like behaviors can be fundamental to increase the naturalness perceived by the human counterpart during the interaction and, consequently, facilitate social attunement. Several characteristics have been identified as crucial to enhance engagement during the interaction with artificial agents, and one of these characteristics is variability (Gielniak, Liu, & Thomaz, 2013). The advent of complex robotic systems allows researchers to implement highly variable human behaviors in artificial agents, to study more in detail on which information humans rely the most when evaluating biological motion. Understanding mechanisms of human perception of synthetic motion can facilitate, in the future, human-robot interaction (Heider, F., & Simmel, 1944). Embedding social robots with human-like behaviors may also lead to the adoption of mindreading strategies, and, as a cascade effect, to attune (on a social level) with the artificial agent. Studies on biological motion perception demonstrated that motion cues influence social attunement towards artificial agents, triggering even empathetic and mindreading processes (Miller & Saygin, 2013; Frith & Frith, 1999). However, it is still unclear to what extent it is possible to manipulate the perception that humans have toward an artificial agent by manipulating the behaviors displayed by the latter. Thus, we designed a series of human-robot interaction scenarios aimed at testing the tendency to attribute human-likeness and intentions towards artificial agents. We used the humanoid robot iCub (Metta et al., 2008) both in real-time interaction and in screen-based experiments. Through the adoption of a systematic approach, we exposed large sets of participants to subtle manipulations of the robot's behavior and tested their reactions using explicit and implicit measures, combined with questionnaires assessing their individual differences. Importantly, we based the implementation and the manipulation of the robot's behaviors on pre-recorded human behaviors.

Such a comprehensive approach to the topic allows for a deeper understanding of the perceptual and cognitive processes that are involved during the interaction between humans and artificial agents.

The first two studies, reported in Publication I and Publication II, examine individuals' sensitivity to hints of human-likeness displayed by an avatar of the iCub robot and by the same embodied robot, respectively. For these two studies, we adopted a self-report questionnaire to assess participants' tendency to attribute anthropomorphic traits and intentions towards artificial agents. In the subsequent studies, reported in Publication III and Publication IV, we examined whether hints of human-likeness affect individuals' attentional engagement during the visual processing of the iCub robot's behavior. To this end, we combined self-report questionnaires used in the first two studies with implicit measures (i.e. eye-tracking data).

Publication I

The first study of the PhD thesis, reported in Publication I, aimed at investigating whether implementing robots with behaviors reflecting attentional capture modulate individuals' perception of its human-likeness. Thus, we implemented pre-recorded humans' behaviors in a virtual version of the iCub robot. Such recordings were acquired through an inertial sensor mounted on the head of a group of participants (N = 20), while they were engaged in a solitaire card game task, and a series of distracting stimuli were presented. Behavioral reactions of participants were extracted and implemented in the iCub simulator. We examined whether parameters of the movement implemented in the robot (i.e. angle amplitude, overall time spent on a target) modulate participants' ratings of its human-likeness, and potential correlation with sociodemographic factors (i.e. gender, age). Our results suggested that the temporal dynamic characterizing the behaviors affected individuals' ratings more than spatial information. Thus, we concluded that it is fundamental to take into account the temporal profile of behaviors implemented in artificial agents if the aim is to make them appear human-like.

Publication II

The second study, reported in Publication II, aimed at understanding human processing of subtle hints of human-likeness displayed by an embodied artificial agent (i.e. the iCub robot). The paradigm consisted of an observation of the robot and was aimed to assess whether humans can perceive subtle differences in the robot's behavior when they have either no information regarding the behavior itself or explicit information regarding the process of implementation of the behavior. We exposed participants to a robot behaving as a human being (based on pre-recorded data, collected during the

study reported as Publication I) or in a machine-like way (this behavior was pre-programmed to be stereotypical and repetitive). Participants were asked to complete several self-report questionnaires after each session with the robot. Then, data collected after each session were compared, as well as differences between the two groups of participants that received different information. Our results highlighted a crucial role of individuals' knowledge on their sensitivity to human-based behavior displayed by an artificial agent, as well as on their attribution of anthropomorphic traits towards the same agent. We concluded that individuals' knowledge-related biases might override perceptual evidence of human-likeness when observing the behavior of a robot.

Publication III

The third study, reported in Publication III, examined behavioral correlates of perceptual processing of a humanoid face during a screen-based paradigm. In particular, the study focused on perceptual and attentional processing underpinning human-robot interaction. The hypothesis is that decomposing the robot behavior in its single components might lead to a better comprehension of the perceptual processes elicited in the human during the observation of robot behavior. The present study aimed to combine self-report and eye-tracking measures to understand human explicit and implicit processes associated with a systematic variation of the same robot's behavior. As a secondary aim, we assessed the discrepancy between explicit and implicit measures of engagement, to further investigate the processes associated with the interpretation of humanoid behavior. Our results pointed out that individuals display higher attentional engagement when the robot displays a human-like behavior than when the artificial agent is behaving mechanically. Additionally, we found that individuals tend to attribute higher ratings of human-likeness to slow, rather than fast, behaviors. We concluded that implementing human-like behaviors in an artificial agent might facilitate attentional processes required for communicating with it, although explicit attributions of human-likeness towards it might depend upon other factors (i.e. temporal dynamics of the behavior).

Publication IV

The final study reported in this thesis, Publication IV (Experiment 1 and Experiment 2), was aimed at investigating whether humans are facilitated in the visual processing of information conveyed by a human or by a robot. We manipulated both agents' behaviors in order to display either active, mentalistic behaviors, or passive, repetitive behaviors. In Experiment 1, we combined eye-tracking and performance measures with self-report data, to understand the relationship between visual processing, decision making, and explicit awareness with the nature of the agent and the behavior

displayed in a screen-based experiment. In Experiment 2, we used the same set of stimuli in an online study to evaluate individuals' tendency to attribute anthropomorphic traits based on the behavior displayed by the two agents (i.e. the human and the iCub). Our results showed that individuals' attention was more engaged when observing seemingly intentional behaviors than when observing mechanical ones. Furthermore, individuals recognized intentional behaviors more accurately than mechanical ones. We concluded that, among human-based behaviors, the ones showing a clear intent spontaneously engage individuals' attention, and modulate human-likeness attribution towards the agent displaying them.

Publication I constitutes the manuscript of the paper "Ghiglino, D., De Tommaso, D., & Wykowska, A. (2018, November). Attributing human-likeness to an avatar: the role of time and space in the perception of biological motion. In *International Conference on Social Robotics* (pp. 400-409). Springer, Cham."

Publication II constitutes the manuscript of the paper "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. In *Proceedings of the 42nd Annual Virtual Meeting of the Cognitive Science Society (Cogsci 2020)* (pp. 952-958)"

Publication III constitutes the manuscript of the paper "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."

Publication IV constitutes the manuscript of the paper "Ghiglino, D., Willemse, C., De Tommaso, D., & Wykowska, A. (under submission, 2020). Mind the eyes: artificial agents' eye movements modulate attentional engagement and anthropomorphic attribution. Currently under submission to Frontiers in Robotics and AI."

SECTION II -PUBLICATIONS

2.1 Publication I: Attributing human-likeness to an avatar: the role of time and space in the perception of biological motion

Ghiglino D.[1-2], De Tommaso D.[1] and Wykowska A.[1]

[1] Istituto Italiano di Tecnologia, Via Morego, 30, 16163, Genova, Italy

[2] DIBRIS, Università degli Studi di Genova, Via Opera Pia 13, 16145, Genova, Italy

2.1.1 Abstract

Despite well-developed cognitive control mechanisms in most adult healthy humans, attention can still be captured by irrelevant distracting stimuli occurring in the environment. However, when it comes to artificial agents, such as humanoid robots, one might assume that their attention is "programmed" to follow a task, thus, being distracted by attention-capturing stimuli would not be expected. We were interested in whether a behavior that reflects attentional capture in a humanoid robot would increase its perception as human-like. We implemented human behaviors in a virtual version of the iCub robot. Twenty participants' head movements were recorded, through an inertial sensor, during a solitaire card game, while a series of distracting videos were presented on a screen in their peripheral field of view. Eight participants were selected, and their behavioral reactions (i.e. inertial sensor coordinates, etc.) were extracted and implemented in the simulator. In Experiment 2, twenty-four new participants were asked to rate the human-likeness of the avatar movements. We examined whether movement parameters (i.e. angle amplitude, overall time spent on a distractor) influenced participants' ratings of human-likeness and if there was any correlation with sociodemographic factors (i.e. gender, age). Results showed a gender effect on human-likeness ratings ($t=2.425$, $p=.024$). Thus, we computed a GLM analysis including gender as a covariate. The main effect of the time of movement ($F=9.179$, $p=.006$) surviving Bonferroni correction ($p<.05$) was found. We conclude that humans rely more on temporal than on spatial information when evaluating properties (specifically, human-likeness) of the biological motion of humanoid-shaped avatars.

Keywords: Human-likeness of robot behavior, Biological Motion, Humanoid robots.

2.1.2 Introduction

In designing artificial agents that are to appear human-like to increase perceive naturalness and facilitate social attunement, many researchers address the issue of creating human-like behavior. Several characteristics have been identified, and one crucial characteristic is variability (Gielniak, Liu, Thomaz, 2013): behavioral observations demonstrate that humans never display the same behavior twice. For example, several studies demonstrated that subjects tend to adopt unique patterns of kinematic strategies to attend the very same target (Freedman and Sparks, 2000; Stergiou and Decker, 2011; Desmurget et al., 1995). The recent advent of complex humanoid systems, allow researchers to implement fragmentized human behaviors in artificial agents, to study more in detail on which information humans rely the most when evaluating biological motion. Furthermore, a deeper understanding of the human perception of synthetic motion will facilitate, in the future, human-robot interaction (Khatib et al., 2004). This stems in part from the fact that humans, when

interacting with other mammals (Fox, 2006), easily understand goals, motivation, and beliefs behind human-like behaviors (Blakemore and Decety, 2001), also relying on motion clarity. It is not clear whether artificial motion patterns of a robot would be as easily understood and predicted. Therefore, it is of high importance to examine what parameters of robot behavior make it well-understood by human users. Evidence from literature pointed out that motion cues might influence the social attunement perceived towards artificial agents, enhancing even empathetic and mentalizing processes (Miller and Saygin, 2013; Heider and Simmel, 1944, Frith and Frith, 1999). Starting with observing and recording human motion, several techniques can be used to transfer movement parameters in artificial agents (Pollard et al., 2002; Lee and Lee, 2006; Aggarwal and Cai, 1999). However, given the huge variability of humans' motion, it is still unclear which components of observed behaviors affect the most perception of human-likeness. The projection of human motion in a simulated environment might be a suitable method to study systematically these factors.

2.1.2.1 Aim of the study

The goal of the present study was to investigate how human participants perceive biological movement displayed in an artificial agent in terms of human-likeness. We selected an attention-capture scenario because attention capture seems to be a very human-like phenomenon. Humans (and several other animal species) have developed mechanisms to attend relevant events in the environment. The "decision" of the brain to attend to a given event in the environment is made through a combination of bottom-up characteristics of the stimulus (e.g., the salience of the stimulus) and internal top-down factors of the agent (e.g., bias towards emotional stimuli, or a particular sound of, for example, one's own child's voice). However, in many cases, the brain attends to stimuli that "capture" attention through their salience, although this disrupts a given task at hand. Think, for example, of driving. The driver should be focused on the road ahead of him/her and on keeping the car in the assumed lane. However, if there is a very loud distracting sound or bright light flashing in the peripheral vision, the driver might be attracted by this event, and in consequence lose focus on the task, potentially causing an accident. Therefore, although evolutionarily adaptive, the attentional capture phenomenon can be disruptive for a task. In this context, one might think that artificial intelligence should be better adapted to the successful completion of a given task, and not allow being distracted by peripheral events that might result in sub-optimal performance in a task. We reasoned, that "being distracted" – especially with variable ways of reacting to the distracting stimuli might be perceived as an essentially human-like feature. We, therefore, set out to test if equipping a humanoid

robot with behaviors reflecting attentional capture would make it be perceived as human-like, and which particular aspects of the behavior would be crucial for attributions of human-likeness. To this aim, we recorded human head and eye movement during an attentional capture paradigm. The recorded behaviors were filtered, and eight different movement profiles were implemented on an iCub (Metta et al., 2008) simulator. Then, a group of participants was asked to rate the human-likeness of the movements of the simulator.

2.1.3 Materials and Methods

2.1.3.1 Attentional capture with humans

Selection of Distracting Stimuli. Sixty HD quality videos were selected from YouTube accordingly to the following criteria: (1) presence of a single salient sound in the whole sequence (i.e. a phone ring, a woman laugh, a door slam, etc.); (2) absence of inappropriate contents (i.e. politic, racism, sexism, etc.); (3) more than 100 M views. Selected videos were edited using Apple Final Cut Pro, in order to make all of them last for the same amount of time (twenty seconds). Fifty-five anonymous Italian participants were asked to rate the emotional content of the videos through an online platform (soscisurvey.de, Leiner, 2018), using a ten-point Likert scale (0=not emotional at all; 10=strongly emotional). One subgroup (n=26) was asked to rate only the audio tracks of the videos. The other subgroup (N=30) was asked to rate both the audio and the visual component of the videos. After collecting the data, the ratings of the two groups were compared. Four of the sixty initial stimuli were excluded because of the inconsistency between the ratings of the two groups. The remaining fifty-six videos were categorized into "Affective" and "Non-affective" stimuli, using the median score of the raters as a cutoff value between the two categories. Eighteen videos were then extracted, according to the following criteria: the nine with the lowest score ("Non-affective" videos) and the nine with the highest score ("Affective" videos). By using Apple Final Cut Pro, audio tracks of the final eighteen videos were manipulated, in order to increase the salience of one single sound per video (i.e. the phone ring, the woman laugh, the door slam). Furthermore, we edited the videos in order to ensure that the physical properties of the sounds (i.e. volume and sampling rate, 44.1 kHz) were consistent. For each video, the volume of the single salient sound was increased, while all the other sounds were reduced. The final pool of videos was implemented in an attentional capture paradigm as distracting stimuli.

2.1.3.2 Recording of humans' behaviors

Participants**.** Twenty-two healthy young adults (9 females; 19-34 years of age) were recruited. All participants were native Italian language speakers with no history of psychiatric or neurological diagnosis, substance abuse, or psychoactive medication. All participants had normal or corrected-to-normal vision and reported no history of hearing impairment. Experimental protocols followed the ethical standards laid down in the Declaration of Helsinki and procedures were approved by the local Ethics Committee (Comitato Etico Regione Liguria). Each participant provided written informed consent to participate in the experiment. Participants were not informed regarding the purpose of the study before the experiment but were debriefed upon completion.

Experimental design**.** Participants were seated in a sound-attenuated experimental booth with dimmed light, in front of a notebook screen (HP Stream 14-ax011nl, 1366 x 768) (Fig. 1).

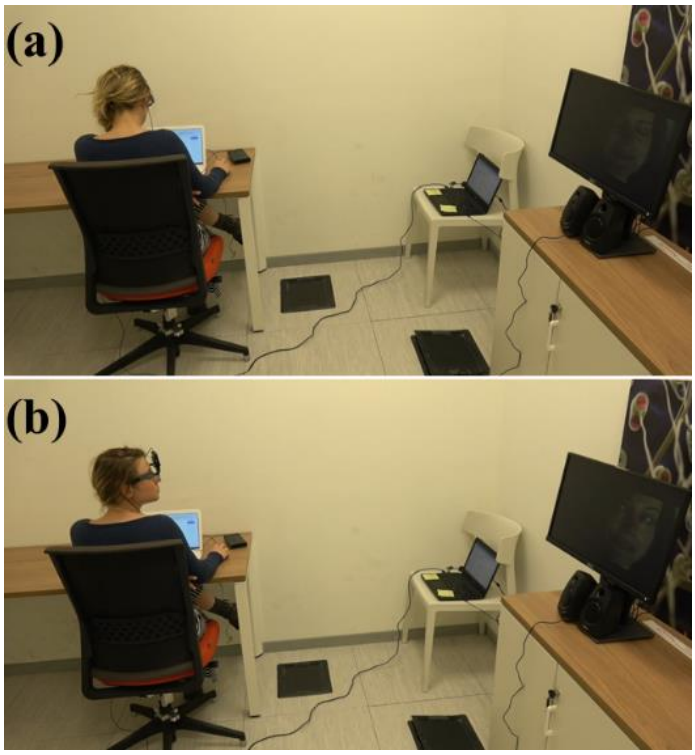

Fig. 1. Experimental setup: (a) participant is engaged in a solitary game on the laptop; (b) participant reacts to a distracting stimulus.

They were instructed to perform a solitary card game (spider one-suit) on the notebook and to pay attention to the game. While participants were engaged in the card game, distracting stimuli were presented in the far periphery of their field of view (100° on the right, 227 cm of distance), on a second computer screen (DELL S2716DG, 2560 x 1440 pixels). The audio tracks of the distracting stimuli were played through loudspeakers (Logitech LGT-Z130), located under the second screen. The experiment was programmed and run on OpenSesame (Mathot, Schreij and Theeuwes, 2012).

Participants' eye movements were recorded throughout the entire duration of the experiment with a mobile eye-tracking device (TobiiAB, 2015). Head movements were recorded using an inertial sensor (Bosch Sensortec BNO055 Intelligent 9-Axis Absolute Orientation Sensor (Sensortec, 2014) mounted on the eye-tracker and integrated into the OpenSesame experiment. We implemented a periodic task, running at 50Hz, that requests every 20ms the Euler angles to the inertial sensor. The absolute values of these angles, together with the sampling timestamp (Timestamp, Yaw, Pitch, and Roll) were saved in a .csv file, one for each distractor stimulus. Specifically, the periodic task was synchronized with the video stimuli, so that the duration of each inertial measure was aligned with the duration of the video. For each experiment, we collected 18 sessions for each participant, in total 360 .csv files.

Data Analyses**.** Participants' data were extracted from the eye-tracker and the inertial sensor through Tobii Pro Lab and OpenSesame, respectively. Two participants were excluded due to the poor quality of their data. Participants' reactions to distracting stimuli were defined as head rotations of at least 30° on the yaw axis (horizontal plane) of the inertial sensor. Reactions of participants were treated and analyzed as a count variable. For each subject, three final parameters were extracted: (1) total amount of distractions during the whole experiment, (2) total amount of distractions that occurred during "Affective" stimuli, and (3) total amount of distractions occurred during "Non-affective" stimuli. A Wilcoxon matched-pairs test was used to verify a potential difference between "Affective" and "Non-affective" conditions. Furthermore, to explore gender differences in distractibility among participants, a Fisher exact test was used to compare males and females, separately for "Affective" and "Non-affective" conditions. In order to apply the Fisher Exact Test, the number of reactions was converted into a relative percentage estimated on the single subject. Finally, a binomial test (n=20, p=50%, 1-α=.95) was used to identify the most distracting stimuli of our pool. Two sounds (a gun shot and a woman's orgasm) survived the .95 threshold, meaning that at least 70% of our sample reacted to sound with a distraction).

2.1.3.3 Implementation of humans' behaviors in an iCub simulator

Selection of behaviors**.** During the attentional capture paradigm, fifteen participants reacted to the sound "gun shot" and fourteen participants reacted to the sound "woman orgasm". Thus, we took into consideration the resulting twenty-nine reactions. For each reaction, we extracted two main parameters: (1) amplitude (°) of the movement; (2) time (s) spent on the distractor. The first parameter represented the angle of rotation of the head toward the distracting screen and was calculated as the difference between the average position assumed by the head of the participant during the whole

video and the maximum distance reached on the horizontal plane (yaw axis of the inertial sensor) during the same temporal window. The time spent on the distractor was estimated as the time spent by the subject on a point of the horizontal plane exceeding two standard deviations from the average position of the head. Setting this high threshold allowed us to extract thirteen reactions from the initial pool. Then, the median value of the amplitude (Mdn=51,108°) and the median value of the time spent on the distractor (Mdn=1,664 s) were calculated and used as a cutoff to classify the reactions. Specifically, reactions were divided into four categories, accordingly to the combination of the amplitude of the movement and the time spent on the distractor, namely:

Amplitude and time above the median; (2) Amplitude above the median and time below the median; (3) Amplitude and time below the median; (4) Amplitude below the median and time below the median.

For each condition, the two most representative reactions were extracted (one for the “gun shot” and one for the “woman orgasm”). Eight reactions from eight different participants were selected as the final pool.

Reproduction of the head movements on the iCub simulator. The iCub simulator (Fig. 1) has been designed to reproduce the physics and the dynamics of the robot.

Fig. 2. Example of the iCub simulator.

It has been implemented collecting data directly from the robot design specifications in order to achieve a replication as accurately as possible. Moreover, the software architecture is the same used to control the physical robot. Specifically, we decided to use the Direct Position Control algorithm (see http://wiki.icub.org/images/c/cf/ICub_Control_Modes_1_1.pdf) for sending the joint positions to the iCub head. According to the specifications available in the iCub Wiki (http://wiki.icub.org/wiki/ICub_joints#Head_2.0), the head joints are the ones with indexes 0, 1, and 2, respectively the neck Pitch, Roll, and Yaw. At first, we needed to normalize the Euler angles

recorded with the inertial sensor to get relative angles concerning the initial head pose at the onset of the stimulus. In such a way, we transferred on the robot the relative rotation due to the distractor, assuming always the same starting head pose. The experiment was designed to guarantee, with good approximation, this assumption. In fact, the participants were always looking straight at the screen whenever a video stimulus occurred. We excluded all the other recordings not satisfying this condition. This preprocessing of the data was enough to reproduce on the iCub simulator the head movements using the Direct Position Control algorithm. This control technique is used whenever joint positions are sent at a high frequency because no trajectory generation in between is needed.

2.1.3.4 Human-likeness survey

Participants. Twenty-four participants (13 females; 26-60 years of age) completed an online survey evaluating the human-likeness of the iCub simulator. 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 ethical committee (Comitato Etico Regione Liguria).

Experimental design. Eight videos of six seconds each were recorded from the simulator. Videos were then uploaded on an online platform (soscisurvey.de) and associated with the following question: "On a scale from 1 (extremely mechanistic) to 10 (extremely human-like), how would you rate iCub behaviors in terms of human-likeness?". Each video and the associated question was presented ten times during the survey, mixed with the other items in random order. Participants rated the human-likeness of the simulations, relying only on motion information. They were not informed that the behaviors were all based on previous recordings of humans' motions, but they were debriefed after the survey. To investigate whether the ratings were influenced by subjective factors, participants were also asked to complete the Empathy Quotient (EQ) questionnaire [Baron-Cohen] after the survey.

Data analyses. A two-sample T-Test was used to assess gender differences in our sample's ratings. Pearson's correlations were applied to evaluate possible correlations between participants' global ratings and subjective measures (EQ).

To explore how the components of biological motion (amplitude of the movement and time spent on the distractor) affect ratings of human-likeness, statistical analyses were applied. The amplitude of the movement and time spent on the distractor was entered as two-level within-categorical predictors

in the context of the General Linear Model (GLM). The gender of our participants was included in the model as a nuisance covariate. Post hoc effects were estimated by calculating the Bonferroni test.

2.1.4 Results

2.1.4.1 Recordings of human behaviors

Statistical analyses performed on the average number of reactions across participants revealed a significant difference between Non-affective and Affective stimuli.

Specifically, the Wilcoxon Matched Pairs Test detected a significant difference (N=20, T=3.5, Z=3.789, p<.001) between the average number of reactions that occurred during Non Affective stimuli (M=2.30, SD=2.00) and Affective stimuli (M=5.60, SD=2.46) (Fig. 2).

For both Non-affective and Affective conditions, Fisher Exact Test revealed no significant effect of gender (p>.05) on distractibility during the experiment (Fig. 3).

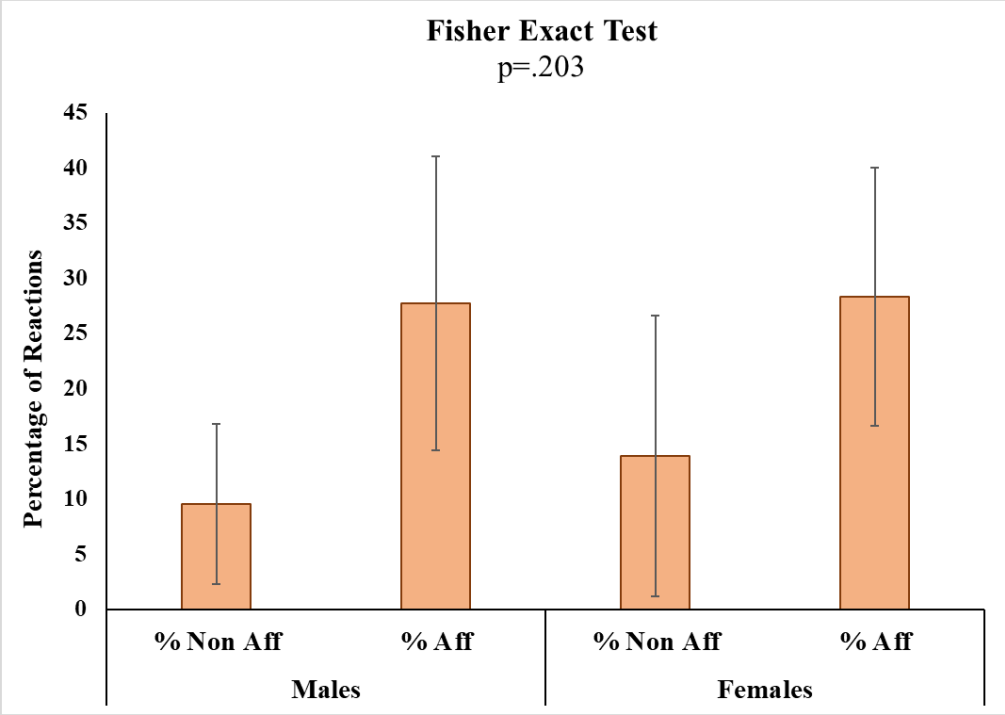

Fig. 3. Differences between males and females on the percentage of reaction displayed during Non-affective (Non-aff) and Affective (Aff) stimuli; vertical bars denote standard deviation of the data.

2.1.4.2 Human-likeness survey

The two-sample T-Test revealed a significant difference (T=2.425, p<.05) of gender on the human-likeness ratings, pointing out that females (M= 6.39, SD=1.38) usually rate higher human-likeness than males (M=4.90, SD=1.64) (Fig. 4). No correlation was found between ratings and subjects' Empathy Quotient scores.

The analyses modeled in the General Linear Model revealed no significant interaction between the amplitude of the movement and time spent on the distractor (F=0.48, p=.50). Furthermore, no main effect was found for the amplitude (F=0.18, p=.68), although results showed a significant main effect of the time spent on the distractor (F=9.18, p<.01) (Fig. 5) surviving Bonferroni correction (p<.01). Specifically, results suggest that longer time spent on the distractor might determine higher human-likeness ratings.

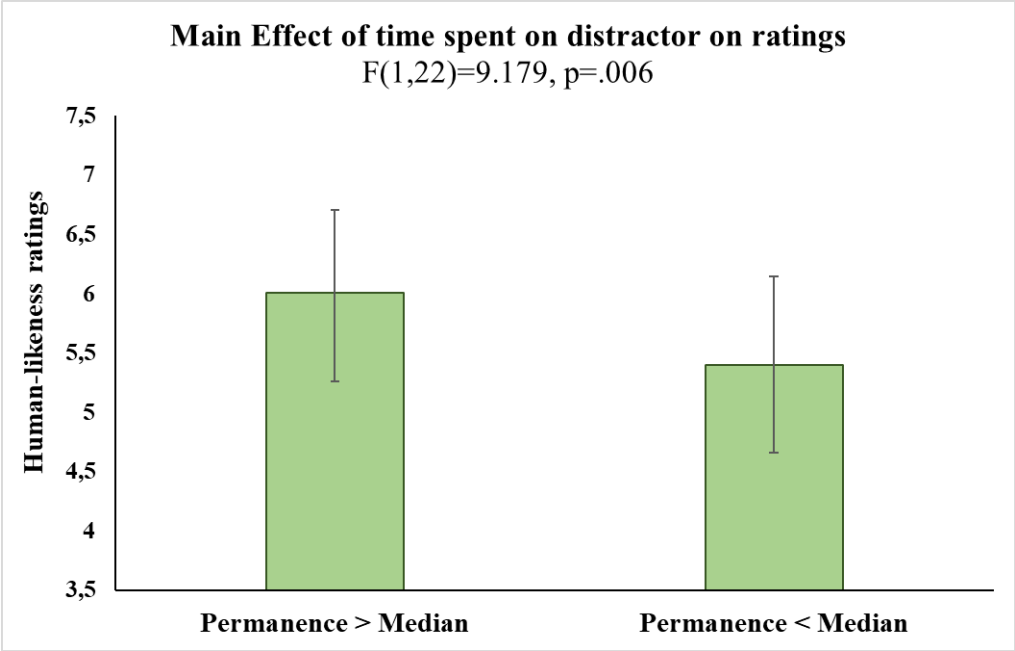

Fig. 5. The main effect of the "Time spent on distractor" on human-likeness ratings; vertical bars denote confidence intervals; Cousineau procedure was applied for correcting bars for within-participants comparisons.

2.1.5 General discussion

Our study aimed to examine parameters of biological motion implemented on a humanoid robot avatar that determine the perceived human-likeness of the motion.

In Experiment 1 we focused on the recording of human behaviors. We recorded participants' head and eye movements during an attentional paradigm. Before implementing the recorded data on an iCub simulator, a preliminary check of the data was required. Thus, we investigated whether

differences between participants (males vs females) or between conditions (Non-affective vs Affective) affected our results. No difference was found between males and females, suggesting that we could use all participants' recordings regardless of their gender for subsequent implementation. At the same time, we found a difference between our experimental conditions. Specifically, results showed that Affective stimuli (i.e. a laugh, a cry, a scream, etc.) elicited more frequent reactions compared to Non-affective ones (i.e. a phone ring, a metal drop, a door closing, etc.). We combined this result within the binomial test, to extract the most representative behaviors recorded during the attentional capture paradigm.

Then, eight behaviors of different participants were selected, extracted, and, subsequently, implemented on an iCub simulator. An independent sample of participants was asked to rate the human likeness of the robot in the simulator, relying only on motion information. Our results showed that females ratings of human-likeness were generally higher than males' ratings. This might suggest that females might be more prone to attribute human likeness than males to a robot simulator, regardless of the physical properties of the movement displayed. In line with previous research (Bisio et al., 2014), we also confirmed that humans, when asked to judge biological motion, rely more on temporal, than on spatial information. Interestingly, although all movements were copied from human behaviors, the average rating of participants was around 5.48. This might suggest that regardless of the naturalness of the movement, humans are still biased by additional visual information (the robot shape) when evaluating biological motion. Furthermore, a large variability was detected between participants' ratings. Despite the lack of correlation between the Empathy Quotient and the ratings, we hypothesize the existence of personality traits that might influence participants' ratings. Further studies should investigate which factors might explain this variability.

2.1.6 Conclusion

Our results showed that temporal features of a movement are crucial in the perceived human-likeness of a movement exhibited by an avatar of a humanoid robot. Thus, particular attention shall be paid to temporal trajectory when using avatars (or robots) to reproduce humans' behavior. Furthermore, large variability detected in participants' ratings of human-likeness and the gender difference suggests the necessity of investigating more in detail individual differences, especially when exploring attribution of human-likeness to an artificial agent.

2.1.7 Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant awarded to A. Wykowska, titled "InStance: Intentional Stance for Social Attunement. Grant agreement No: 715058)

References

Aggarwal, J. K., & Cai, Q. (1999). Human motion analysis: A review. Computer vision and image understanding, 73(3), 428-440.

Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., & Pozzo, T. (2014). Motor contagion during human-human and human-robot interaction. PloS one, 9(8), e106172.

Blakemore, S. J., & Decety, J. (2001). From the perception of action to the understanding of intention. Nature Reviews Neuroscience, 2(8), 561.

Desmurget, M., Rossetti, Y., Prablanc, C., Jeannerod, M., & Stelmach, G. E. (1995). Representation of hand position prior to movement and motor variability. Canadian journal of physiology and pharmacology, 73(2), 262-272.

Fox, R. (2006). Animal behaviours, post-human lives: Everyday negotiations of the animal–human divide in pet-keeping. Social & Cultural Geography, 7(4), 525-537.

Freedman, E. G., & Sparks, D. L. (2000). Coordination of the eyes and head: movement kinematics. Experimental brain research, 131(1), 22-32.

Frith, C. D., & Frith, U. (1999). Interacting minds--a biological basis. Science, 286(5445), 1692-1695.

Gielniak, M. J., Liu, C. K., & Thomaz, A. L. (2013). Generating human-like motion for robots. The International Journal of Robotics Research, 32(11), 1275-1301.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. The American journal of psychology, 57(2), 243-259.

http://wiki.icub.org/images/c/cf/ICub_Control_Modes_1_1.pdf

http://wiki.icub.org/wiki/ICub_joints#Head_2.0

http://wiki.icub.org/wiki/Simulator_README

Khatib, O., Warren, J., De Sapio, V., & Sentis, L. (2004). Human-like motion from physiologically-based potential energies. In On advances in robot kinematics (pp. 145-154). Springer, Dordrecht.

Lee, J., & Lee, K. H. (2006). Precomputing avatar behavior from human motion data. Graphical Models, 68(2), 158-174.

Leiner, D. J. (2018). SoSci Survey (Version 2.5.00-i1142) [Computer software]. Available at http://www.soscisurvey.com

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.

Metta, G., Sandini, G., Vernon, D., Natale, L., & Nori, F. (2008, August). The iCub humanoid robot: an open platform for research in embodied cognition. In Proceedings of the 8th workshop on performance metrics for intelligent systems (pp. 50-56). ACM.

Miller, L. E., & Saygin, A. P. (2013). Individual differences in the perception of biological motion: links to social cognition and motor imagery. Cognition, 128(2), 140-148.

Pollard, N. S., Hodgins, J. K., Riley, M. J., & Atkeson, C. G. (2002). Adapting human motion for the control of a humanoid robot. In Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on (Vol. 2, pp. 1390-1397). IEEE.

Sensortec, B. (2014). Intelligent 9-axis absolute orientation sensor. BNO055 datasheet, November.

Stergiou, N., & Decker, L. M. (2011). Human movement variability, nonlinear dynamics, and pathology: is there a connection?. Human movement science, 30(5), 869-888.

TobiiAB, Stockholm (2015) 'Tobii Pro Glasses 2 Product Description'.

2.2 Publicaction II: Can I get your (robot) attention? Human sensitivity to subtle hints of human-likeness in a humanoid robot's behavior

Ghiglino D. [1-2], De Tommaso D.[1], Willemse C.[1], Marchesi S.[1], Wykowska A.[1]

[1] Istituto Italiano di Tecnologia, Via Morego, 30, 16163, Genova, Italy

[2] DIBRIS, Università degli Studi di Genova, Via Opera Pia, 13, 16145, Genova, Italy

2.2.1 Abstract

Designing artificial agents that can closely imitate human behavior, might influence humans in perceiving them as intentional agents. Nonetheless, the factors that are crucial for an artificial agent to be perceived as an animated and anthropomorphic being still need to be addressed. In the current study, we investigated some of the factors that might affect the perception of a robot's behavior as human-like or intentional. To meet this aim, seventy-nine participants were exposed to two different behaviors of a humanoid robot under two different instructions. Before the experiment, participants' biases towards robotics as well as their personality traits were assessed. Our results suggest that participants' sensitivity to human-likeness relies more on their expectations rather than on perceptual cues.

Keywords: Human-robot interaction, humanoid robot, social cognition, intentional stance, mental states, instruction manipulation

2.2.2 Introduction

In everyday life, we are frequently exposed to different smart technologies. From our smartphones to avatars in computer games, and soon perhaps humanoid robots, we are surrounded by artificial agents created to interact with us. Already during the design phase of an artificial agent, engineers often endow it with functions aimed to promote interaction and engagement with it, ranging from its "communicative" abilities to the movements it produces. The idea that an artificial agent able to behave like a human being would boost the spontaneity and naturalness of interaction is well supported by the literature (Ficocelli, Terao, Nejat, 2015; Mirning et al., 2017; Wiese, Metta & Wykowska, 2017). Providing an artificial agent with human-like behaviors might increase social attunement toward it, and this aspect might be crucial for deploying artificial agents in environments where social interaction with them is desirable (e.g., robot-assisted training for individuals diagnosed with autism; Scassellati, Admondi, Matarić, 2012). In fact, several authors demonstrated the advantages of providing artificial agents with human-like behaviors on the quality of interaction with humans (Hancock et al., 2011; Thepsoonthorn, Ogawa & Miyake, 2018).

Perceiving human-likeness from an artificial agent's behavior appears to be modulated by its behavioral capabilities, ranging from the kinematics of the movement (Gielniak, Liu & Thomaz, 2013) to the agent's responsiveness to external stimuli (Willemse & Wykowska, 2019). Even during the interaction with conspecifics, humans rely partially on motion cues when they need to infer the mental states underpinning behavior. Similar processes might be activated during the interaction with embodied artificial agents, such as humanoid robots. At the same time, a humanoid robot that can

faithfully reproduce human-like behavior may undermine the interaction, causing a shift in attribution: from being endearing to being uncanny (Mori, 1970). Furthermore, it is still not clear whether individual biases and prior knowledge related to artificial agents can override perceptual evidence of human-like traits (Hinz, Ciardo & Wykowska, 2019). We hypothesize that human sensitivity to such characteristics varies depending on individual differences and available contextual information. The current study aims to investigate human sensitivity to anthropomorphic characteristics of robot's behavior, based on motion cues, under different conditions of prior knowledge. To meet this aim, we manipulated the human-likeness of the behavior displayed by the robot and the explicitness of instructions provided to the participants. As a secondary aim, we explored some of the individual differences that affect general attitudes towards robots, and the attribution of human-likeness consequently.

2.2.3 Methods

2.2.3.1 Participants

Seventy-nine participants took part in the experiment (mean age = 24.0, SD = 4.4, 50 females). All participants reported no history of psychiatric or neurological diagnosis, substance abuse, or psychiatric medication. Our experimental protocols followed the ethical standards laid down in the Declaration of Helsinki and were approved by the local Ethics Committee (Comitato Etico Regione Liguria). Each participant provided written informed consent to participate in the experiment. Participants were not informed regarding the purpose of the study before the experiment but were debriefed upon completion.

2.2.3.2 Stimuli and Apparatus

In the current study, we sat our participants in a dimly lit sound-attenuated room, in front of an iCub robot (Metta et al. 2008; Natale et al. 2017) that was "playing" a solitaire card game on a laptop located in front of it. We placed a screen connected to a loudspeaker on the right of the iCub robot, on which we played scenes of various movies that were aimed to "distract" the robot from the game. Neither of the screens' displays was visible from the participant's position, but the sound produced by the loudspeaker was audible to everyone in the room (Fig. 1). The setup of the current study was the replica of a previous attentional capture experiment, which involved human participants playing the same solitaire card game while being distracted by the same sequence of movie scenes (see Ghiglino, De Tommaso & Wykowska, 2018 for details).

Figure 1. Experimental setup.

Experimental design and procedure. Before the experiment, we asked all participants to complete a brief sociodemographic questionnaire along with the Autism Quotient test (AQ, Baron-Cohen, et al., 2001), the Big Five Inventory (BFI, John & Srivastava, 1999), and the Negative Attitude Towards Robots Scale (NARS, Syrdal, et al., 2009). We adopted these questionnaires as they are all freely available, easy to administer, and vastly used to broadly assess individual differences that might affect human-robot interaction (see, for example, Schweinberger, Pohl & Winkler, 2020; Muller & Richert, 2018).

All participants of the present experiment were exposed to two different conditions determined by the behavior displayed by the robot: human-like or machine-like. The order of these conditions was counterbalanced between participants. Each behavior consisted of an 8-minutes sequence of eye- and head-movements.

In the human-like condition, the robot's behavior was derived from the recordings of a human participant's eyes and head movement collected using an eye-tracker (Tobii Pro Glasses 2) and an inertial sensor (Bosch Sensortec BNO055 Intelligent 9-Axis Absolute Orientation Sensor) during the attentional capture experiment mentioned above. Human data recorded in our previous experiment were transferred to the iCub head and eyes using a minimum-jerk controlling algorithm. It is important to point out that the behavior observed in the recordings of the human participant was highly variable: each reaction to a distracting stimulus was different from the others in terms of temporal and spatial kinematics (ranging from minimal and fast to wide and slow movements). The behavior displayed by the robot in the "human-like" condition was aimed to embody the same variability and unpredictability of the behavior recorded from the human.

In contrast, for the machine-like condition, we programmed the robot to display repetitive, predictable, and constant behavior. Thus, the machine-like behavior consisted of only one pattern of neck and eye movements, based on the average temporal and spatial movement dynamics extracted from the human recording of the aforementioned experiment. To maximize the difference between the two conditions, during the machine-like behavior, the robot was programmed to move its eyes from left to right repetitively while "playing" the solitaire card game and to react to each distracting stimulus with exactly the same head turn.

We asked the first forty participants (mean age = 24.1±3.73; mean education = 15.8±2.3; 24 females) to carefully observe the robot's behavior during both conditions without adding any further instruction or information. The remaining thirty-nine participants (mean age =24.3 ± 5.07; mean education = 15.2 ± 2.0; 26 females) were told explicitly, from the beginning of the experiment, that the robot would display two different behaviors, and that their task would be to identify which one was based on a human's recordings.

After each condition, all seventy-nine participants filled out the GodSpeed questionnaire (Bartneck et al., 2009) to assess the tendency to attribute anthropomorphic, animated, and likable traits to a robot, and they took part in the InStance test (Marchesi et al., 2019) to investigate the tendency of humans to explain the behavior of a robot using either a mentalistic or a mechanistic vocabulary.

After the completion of both experimental sessions and the questionnaires, all participants were asked if they noticed any difference between the two behaviors displayed by the robot. In case of a positive answer, participants were asked to elaborate on their answer, explicating which one of the two behaviors they thought was more similar to human behavior and why. We expected that participants who noticed the difference between the two conditions would be unanimous on the "correct" attribution of human-likeness. However, we received unexpected human-likeness attributions toward the machine-like condition that we kept into consideration during the data analysis. Eventually, this final explicit question allowed us to differentiate people in terms of sensitivity to the behavioral manipulation and terms of correctly attributed/misattributed human-likeness.

2.2.4 Data Analysis

To explore the effects of our experimental manipulation, several mixed effect general linear models (GLM) were applied in R studio. In each model, we considered the responses in the GodSpeed questionnaire and the InStance test as separate dependent variables and each participant's intercept as a random factor. We included instruction manipulation (Explicit vs No Instructions) and the behavior

displayed by the robot (HumanLike vs MachineLike) as fixed factors. This family of models allowed us to explore the main effects of the single factors and the interaction between the two.

Additionally, we aimed at exploring the effect of participants' attribution of human-likeness on the InStance and the GodSpeed ratings. Thus, we further grouped our participants based on their sensitivity to the subtle differences between the robot's behaviors and on the explicit attribution of human-likeness (provided at the end of the experiment). To avoid confounding effects and/or overfitting of the data, we analyzed participants that received explicit instructions separately from participants that received no instructions. This decision was made also taking into consideration the way participants distributed themselves in the three response groups across the two instructions conditions (under no instructions: 14 correctly attributed human-likeness, 9 misattributed human-likeness, 17 no attribution; under explicit instructions: 31 correctly attributed human-likeness, 8 misattributed human-likeness, 0 no attribution). This between-group difference was tested using a chi-squared test. For all the mixed models, pairwise posthoc comparisons were estimated using the Tukey method. Due to the way linear mixed models partition variance, and the lack of consensus on the calculation of effect sizes for individual model terms (Rights and Sterba, 2019), we estimated standardized effect sizes only in posthoc analyses.

To investigate individual differences that affect human sensitivity to subtle hints of human-likeness in a humanoid robot's behavior, we calculated Pearson's correlation coefficients between the AQ, BFI, NARS, sociodemographic information, GodSpeed questionnaire, and Instance tests. Since we were interested in assessing individual differences that might play a role in the general attitude towards robots, for each participant we used the averages of the GodSpeed subscales and InStance scores as input variables of the correlation matrix.

2.2.5 Results

2.2.5.1 Instruction Manipulation and Robot Behavior

Instance ratings**.** We did not find any significant effects on the InStance scores due to the instructions manipulation ($F(1, 77)=0.41$, $p=.522$), of the behavior displayed by the robot ($F(1, 77)=2.16$, $p=.146$) or of the interaction between the two ($F(1, 77)=0.57$, $p=.455$) (Fig. 2).

GodSpeed ratings**.** We found a significant interaction effect on the Anthropomorphism scores between instructions manipulation and behavior displayed by the iCub ($F(1, 77)=5.64$, $p=.020$), paralleled by a main effect of the behavior ($F(1, 77)=11.11$, $p=.001$). A null effect of instructions

manipulation emerged from the data on this subscale ($F(1, 77)=0.05$, $p=.82$). Under explicit instructions, planned comparisons revealed a significant difference in Anthropomorphism scores: participants tended to attribute higher anthropomorphism to the human-like behavior than to the machine-like behavior ($t(77)=4.01$, $p<.001$). The same pattern was found on the Animacy subscale scores, highlighting the interaction between the instructions and the behavior ($F(1, 77)= 9.33$, $p=.003$), a main effect of the behavior ($F(1, 77)= 9.08$, $p=.004$) and a non-significant effect of the instructions ($F(1, 77)=0.20$, $p=.654$). Planned comparisons pointed out a significant difference in the Animacy scores between the human-like and the machine-like behaviors in the group that received explicit instructions ($t(77)=4.26$, $p<.001$). Interestingly, for the Likeability subscale scores, we found a single main effect of the instruction manipulation ($F(1, 77)=12.14$, $p<.001$), but neither a significant effect of behavior ($F(1, 77)=3.50$, $p=.065$) nor of interaction ($F(1, 77)=2.03$, $p=.158$). Post-hoc comparisons revealed a significant difference between the two instructions provided to the participant on the perceived likeability of the robot both after the human-like ($t(77)=-2.71$, $p=.038$) and after the machine-like ($t(77)=-3.93$, $p<.001$) behaviors (see Fig. 2 for details).

2.2.5.2 Robot's Behavior and Participants' attribution

The frequencies of participants' human-likeness attribution were different between the two instructions we provided them with ($\chi^2(2)= 23.47$, $p<.001$). Thus, we ran the subsequent analyses separately for the two instruction groups.

No instructions group**.** No main effect of the robot's behavior was found on the InStance ratings ($F(1, 37)=0.06$, $p=.805$), nor on the Anthropomorphism ($F(1, 37)= 0.35$, $p=.558$), Animacy ($F(1, 37)= 0.59$, $p=.448$) and Likeability ($F(1, 37)=0.23$, $p=.638$) subscales of the GodSpeed.

Similarly, participant's human-likeness attribution did not affect neither InStance ratings ($F(2, 37)=1.04$, $p=.363$), nor Anthropomorphism ($F(2, 37)=1.95$, $p=.157$), Animacy ($F(2, 37)= 0.52$, $p=.602$) or Likeability scores ($F(2, 37)=0.42$, $p=.662$).

No interaction between robot's behavior and participants' attribution was found on the InStance ($F(2, 37)=1.76$, $p=.186$) and Anthropomorphism ($F(2, 37)=2.21$, $p=.124$). An interaction between behavior and attribution was found on Animacy ($F(2, 37)=6.16$, $p=.004$) and Likeability ($F(2, 37)=7.65$, $p=.002$) scores. The effect on the Animacy scores did not survive posthoc comparisons. Post-hoc analysis revealed that participants misattributing human-likeness tended to attribute higher likeability to the robot displaying the machine-like behavior, compared to the human-like behavior ($t(37)=3.77$, $p=.035$, $d=1.24$).

Explicit instructions group**.** No main effect of the robot's behavior was found on the InStance ratings (F(1, 37)=0.01, p=.940), or on the Anthropomorphism (F(1, 37)=0.51, p=.477), Animacy (F(1, 37)=0.19, p=.669) and Likeability (F(1, 37)=0.22, p=.639) subscales of the GodSpeed.

No main effect of participant's attribution was found on the InStance ratings (F(1, 37)=0.40, p=.529), or on the Anthropomorphism (F(1, 37)= 1.45, p= .235), Animacy (F(1, 37)=0.00, p=1.00) and Likeability (F(1, 37)=0.45, p=.504) subscales of the GodSpeed.

A significant interaction between the robot's behavior and participant's human-likeness attribution was found on the InStance ratings (F(1, 37)=4.31, p=.045), and on the Anthropomorphism (F(1, 37)= 16.89, p<.001), Animacy (F(1, 37)= 25.40, p<.001) and Likeability (F(1, 37)=4.36, p=.044) scores (Fig. 3). The interaction effect on the InStance scores did not survive posthoc comparisons. Planned comparisons revealed that participants making the correct human-likeness attribution provided higher ratings on Anthropomorphism (t(37)=5.33, p<.001, d=-1.75), Animacy (t(37)=6.04, p<.001, d=-1.98) and Likeability (t(37)=2.82, p=.036, d=-0.92) subscales after seeing the human-like behavior. Additionally, participants providing the unexpected attribution rated higher than participants providing the expected attribution after seeing the machine-like behavior on the Anthropomorphism (t(37)=-3.23, p=.010, d=-0.30) and the Animacy (t(37)=-2.97 p =.021, d=-0.72) subscales.

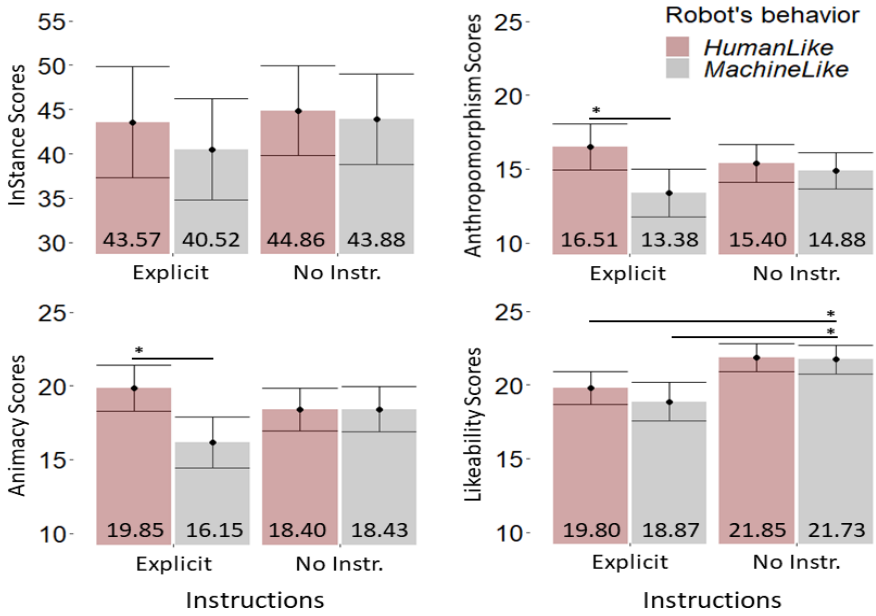

Figure 2: Bar charts showing the fixed effect on questionnaire scores due to the interaction between the instructions provided to the participants and the behavior displayed by the robot (GLM). Error bars: +/- 1. SE. Asterisks denote significant comparisons. Numbers indicate the mean value of each cell.

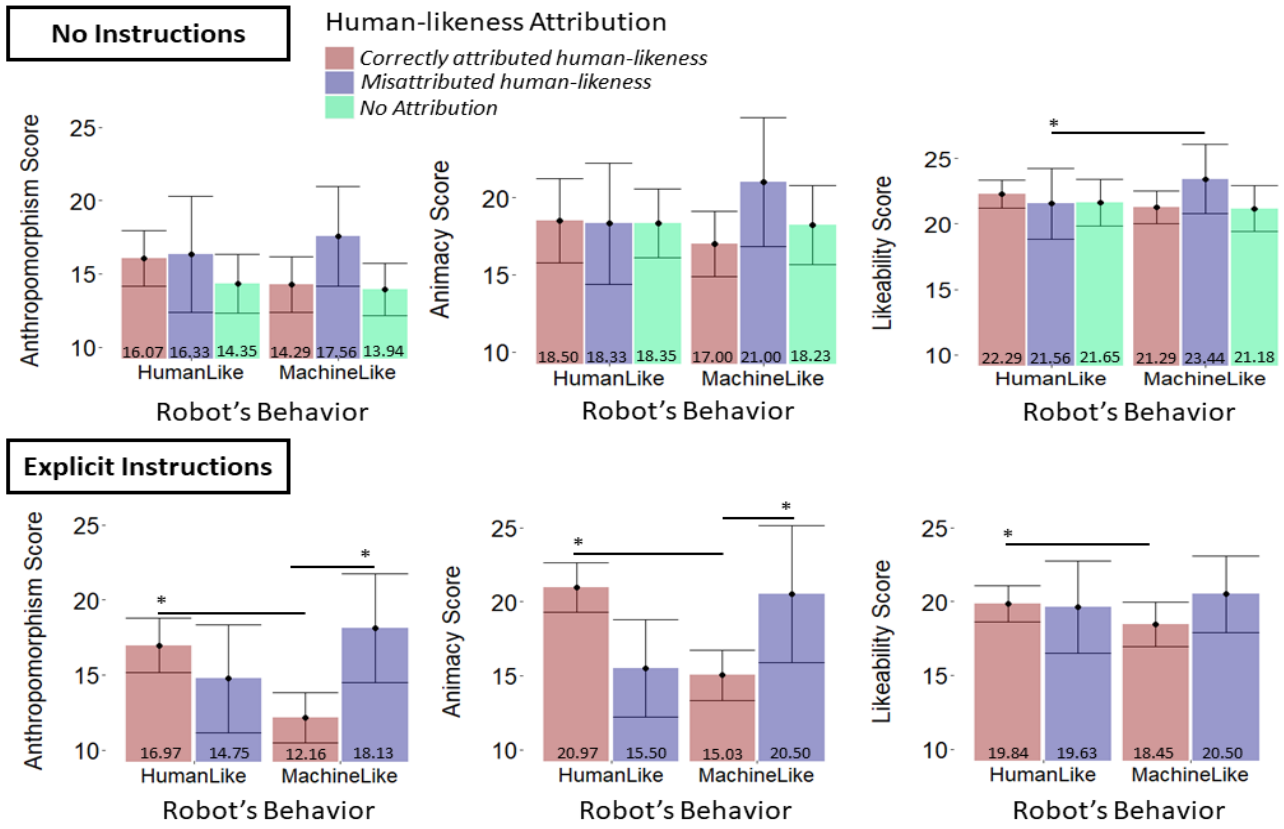

Figure 3: Bar charts showing the fixed effect on the GodSpeed scores due to the interaction between the instructions provided to the participants and the behavior displayed by the robot (GLM). Error bars: +/- 1. SE. Asterisks denote significant comparisons. Numbers indicate the mean value of each cell.

2.2.5.3 Individual differences

Our analyses showed small negative correlations between the InStance score and years of education ($r(77)=-.257$, $p=.022$) and AQ scores ($r(77)=-.246$, $p=.029$). We also found small negative correlations between the Likeability subscale of the GodSpeed questionnaire and the three subscales of the NARS ("Situations of Interaction": $r(77)=-.27$, $p=.018$; "Social Influence": $r(77)=-.23$, $p=.040$; "Emotions in Interaction" $r(77)=-.29$, $p=.008$). Additionally, our results pointed out a systematic, although small, set of correlations between BFI and NARS subscales. Specifically, the Conscientiousness subscale of the BFI correlates negatively with the subscales "Situations of Interaction" ($r(77)=-.32$, $p=.004$) and "Social Influence" ($r(77)=-.28$, $p=.012$) of the NARS. Besides, the Neuroticism subscale of the BFI correlates positively with the subscales "Situations of Interaction" ($r(77)=.24$, $p=.036$) and "Emotions in Interaction" ($r(77)=.28$, $p=.011$) of the NARS.

2.2.6 Discussion

The main aim of the current study was to assess whether the information available before the interaction with an artificial agent modulates human sensitivity to subtle hints of an agent's human-likeness. Our data showed that prior knowledge related to the behaviors that we implemented in the robot affected the sensitivity to behavioral manipulation. When we provided no *a-priori* information related to the nature of the behaviors implemented in the robot, participants overlooked the details of the behaviors. Consequently, nearly half of the sample provided with no instructions was not able to recognize any difference between the human-like and the machine-like behaviors. Furthermore, even those participants who spotted the differences between the behaviors often misattributed human-likeness. In addition, we could not find any significant differences in their InStance and GodSpeed scores between conditions. In contrast, all the participants who received the explicit instructions detected a difference between the two behaviors, and this was reflected in the anthropomorphism, animacy, and likeability attributed to the robot.

When we prompted our participants' attention to notice hints of human-likeness in the behaviors of the robot, they tended to differentiate more their answers in the GodSpeed questionnaire between conditions, as if their perception of anthropomorphism, animacy, and likeability depended mainly on their belief of what a human-like movement should look like. This suggests that subtle evidence of behavioral human-likeness might be too weak of a signal during tasks merely involving the observation of artificial agents' behavior. This might be related to the fact that in natural interactions with humans, we are usually not monitoring (or not being asked to monitor) the human-likeness of the counterpart's behavior. Thus, human-likeness might be an implicit feature of human behavior, which we derive only if needed to explain the behavior of a non-human agent. Therefore, during everyday life, our sensitivity to such subtle hints might be low, as we more likely perceive "gestalt" relations between behavioral and contextual elements rather than pure and distinct behavioral features (Spelke, 1990; Hamlyn, 2017).

Our results suggest that the concept of human-likeness itself varies across individuals, overriding perceptual evidence – our participants tended to confirm their own biases and modulated their responses on the GodSpeed questionnaire based on their own perception of human-like behavior, rather than the actual human-like behavior of the robot. This casts a shadow on the idea that having artificial agents able to behave exactly like human beings would, improve social interaction with them, as people appear to have very different priors related to the concept of human-likeness. Indeed, participants perceived the robot as more likable when they received no information related to its

behaviors, regardless of its human-likeness. In other words, perceived, but not actual, human-likeness influenced the likeability of the robot. Thus, the attractiveness of interacting with a humanoid robot might be independent of the subtle behaviors it displays, but might rather depend on the users' attitudes toward it. This further suggests that the less an individual knows about the process of implementation of behavior in a robot, the more they enjoy the interaction with it and perceives it as more engaging. Taken together, these results suggest that the differences in knowledge between participants override perceptual evidence and tweak individual sensitivity to behavioral cues.

The presence of individual differences that affect the way humans interact with artificial agents is further supported by the correlation between BFI and NARS subscales. Our results showed that certain personality traits, such as neuroticism and conscientiousness, influenced participants' attitudes towards robots. High neuroticism scores are often associated with the tendency to experience negative emotions during social interaction (Kaplan et al., 2015). The positive correlation between neurotic traits and NARS scores supports previous literature, suggesting that neurotic people might experience discomfort during the interaction with artificial agents, similarly to how they feel in interactions with other humans (Müller & Richert, 2018). On the other hand, high conscientiousness often relates to better self-regulation and emotional stability, which positively affect social interaction (Smith, Barstead, Rubin, 2017), and might as well ease the interaction with artificial agents. The negative correlation between InStance and AQ scores further supports the idea that social abilities affect humans' general attitude towards artificial agents. Indeed, people with higher autistic traits appeared to have difficulties with explaining the behavior of a robot in terms of the underpinning mental states, relying more on mechanistic terms rather than on mentalistic vocabulary. This might be due to the familiarity that a person has regarding a certain vocabulary when interpreting behaviors in general. Besides, we also found a negative correlation between the Instance test score and the participants' education. We speculate that participants with a higher level of education might be more familiar with the design and functionality of technology in general. This prior knowledge might bias them to explain our robot's behavior relying more on its mechanical apparatus rather than its "desires" and "intentions". We postulate that personality traits and attitudes that play a role in the interaction between humans translate into different approaches towards artificial agents as well. This hypothesis is further supported by the negative correlation between NARS subscales and the perceived Likeability of the robot, indicating that participants' attitudes towards robots affect their engagement during the interaction. Future studies should further explore individual differences that affect participants' behavior and attitudes toward robots to understand whether they play a similar role during human-human and human-robot interactions.

In conclusion, our study suggests that individual knowledge, beliefs, and biases play a major role in modulating human perception of an artificial agent's behavior. These influences seem to be even stronger than perceptual evidence during observational scenarios and need to be taken into consideration in future studies.

2.2.7 Acknowledgments

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

References

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of autism and developmental disorders*.

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*.

Ficocelli, M., Terao, J., & Nejat, G. (2015). Promoting interactions between humans and robots using robotic emotional behavior. *In Proceedings of IEEE transactions on cybernetics*.

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. *In Proceedings of the 10th International Conference on Social Robotics*. Springer, Cham.

Gielniak, M. J., Liu, C. K., & Thomaz, A. L. (2013). Generating human-like motion for robots. *The International Journal of Robotics Research*.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*.

Hinz, N. A., Ciardo, F., & Wykowska, A. (2019). Individual differences in attitude toward robots predict behavior in human-robot interaction. *In International Conference on Social Robotics*. Springer, Cham.

John, O. P., and Srivastava, S. (1999). "The Big Five trait taxonomy: History, measurement and theoretical perspectives". *In Handbook of personality: Theory and research*. New York: Guilford.

Kaplan, S. C., Levinson, C. A., Rodebaugh, T. L., Menatti, A., & Weeks, J. W. (2015). Social anxiety and the Big Five personality traits: The interactive relationship of trust and openness. *Cognitive behaviour therapy*.

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the Intentional Stance towards humanoid robots? *Frontiers in psychology*.

Metta, G., Sandini, G., Vernon, D., Natale, L., & Nori, F. (2008). The iCub humanoid robot: an open platform for research in embodied cognition. *In Proceedings of the 8th workshop on performance metrics for intelligent systems*. ACM.

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI*.

Mori, M. (1970). The uncanny valley. *Energy*.

Müller, S. L., & Richert, A. (2018). The Big-Five Personality Dimensions and Attitudes to-wards Robots: A Cross-Sectional Study. *In Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference*.

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

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological methods*.

RStudio Team (2015), RStudio: Integrated Development for R. RStudio, Inc., Boston, MA, URL: http://www.rstudio.com/.

Scassellati, B., Admoni, H., & Matarić, M. (2012). Robots for use in autism research. *Annual review of biomedical engineering*.

Schweinberger, S. R., Pohl, M., & Winkler, P. (2020). Autistic traits, personality, and evaluations of humanoid robots by young and older adults. *Computers in Human Behavior.*

Smith, K. A., Barstead, M. G., & Rubin, K. H. (2017). Neuroticism and conscientiousness as moderators of the relation between social withdrawal and internalizing problems in adolescence. *Journal of youth and adolescence*.

Syrdal, D. S., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2009). The negative attitudes towards robots scale and reactions to robot behavior in a live human-robot interaction study. *Adaptive and emergent behaviour and complex systems*.

Thepsoonthorn, C., Ogawa, K. I., & Miyake, Y. (2018). The relationship between robot's nonverbal behaviour and human's likability based on human's personality. *Scientific reports*.

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: using neuroscientific methods to make robots appear more social. *Frontiers in psychology*.

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

2.3 Publication III: At first sight: robots' subtle eye movement parameters affect human attentional engagement, spontaneous attunement, and perceived human-likeness

Ghiglino D.[1-2] , Willemse C. [1], De Tommaso D.[1], Bossi F. [1], Wykowska A.[1]

[1] Istituto Italiano di Tecnologia, Via Morego, 30, 16163, Genova, Italy

[2] DIBRIS, Università degli Studi di Genova, Via Opera Pia, 13, 16145, Genova, Italy

2.3.1 Abstract

Human-robot interaction research could benefit from knowing how various parameters of robotic eye movement control affect specific cognitive mechanisms of the user, such as attention or perception. In the present study, we systematically teased apart control parameters of Trajectory Time of robot eye movements (rTT) between two joint positions and Fixation Duration (rFD) on each of these positions of the iCub robot. We showed recordings of these behaviors to participants and asked them to rate each video on how human-like the robot's behavior appeared. Additionally, we recorded participants' eye movements to examine whether the different control parameters evoked different effects on cognition and attention. We found that slow but variable robot eye movements yielded relatively higher human-likeness ratings. On the other hand, the eye-tracking data suggest that the human range of rTT is most engaging and evoked spontaneous involvement in joint attention. The pattern observed in subjective ratings was paralleled only by one measure in the implicit objective metrics, namely the frequency of spontaneous attentional following. These findings provide significant clues for controller design to improve the interaction between humans and artificial agents.

Index Terms— Behavioral sciences, Biological control systems, Robot control, Humanoid robots, Eye movements

2.3.2 Introduction

The human eyes play a special role in daily interactions with others. With gaze, we efficiently communicate to others our internal mental states and our interest in the external environment (Kobayashi & Kohshima, 2001; Emery, 2000). Sometimes, just by looking at a person's, we are able to infer the other's intentions, emotions, or action plans. During interaction with another human, the eye region typically is the most attended of all facial features and the most used source of information, independent of the specific characteristics of the interaction (Itier and Batty, 2009).

In order to improve the quality of the interaction, Pelachaud and Bilvi (2003) proposed a communicative model of gaze for embodied conversational agents. The authors claimed that simple variations in the temporal components of gaze behavior might induce different attributions toward the artificial agent. Other authors tested the implementation of gaze behavior in artificial agents during different activities, such as storytelling, conversation and reading (for a review, see Ruhland et al., 2015) and concluded overall that this implementation requires integration of knowledge from a large number of disciplines, from neuroscience to computer graphics. In order to design such behaviors, a better understanding of human gaze is needed.

From a practical point of view, robot designers developed several tools to improve the naturalness of their artificial agents' behaviors. Biologically plausible controllers might facilitate communication during human-robot interaction. Indeed, the implementation of human-like behavior in an artificial agent might encourage the adoption of the same mental models humans spontaneously adopt towards their conspecifics. However, additional studies are needed in order to test this hypothesis and provide designers additional guidelines defining the best design of such controllers. In this context, we present a systematic approach of exploring not only the reliability of theoretical models but also participants' perception of an artificial agent displaying various types of gaze behaviors. We use objective (eye-tracking) and subjective measures (ratings) in order to obtain a comprehensive picture of how various gaze behaviors are received by the user.

2.3.2.1 Aims

In this work, we investigate how different configurations of the same robot controller may affect cognition and attentional engagement of the user, as well as subjective impression of the robot's human-likeness while maintaining the same task for all the conditions. We aim to provide roboticists with novel methods, grounded in cognitive psychology, for developing customizable controllers and for using effective strategies to configure the existing ones. To address the aims of our study, we filmed an iCub robot (see Metta et al., 2008; Natale et al., 2017) that was systematically manipulated to display two specific parameters of eye movements in the iCub controller: Trajectory Time (rTT) and Fixation Duration (rFD). rTT refers to the time required for the robot's pupil to shift from one fixed position to one other fixed position in space. rFD refers to the amount of time that the robot's pupil spends on a given target before moving again. We administered a rating scale to examine how these manipulations affected subjective attributions of human- likeness. Furthermore, to tease apart how human cognitive and attentional mechanisms are affected by these manipulations, we tracked participants' eye movements, as eye movement patterns are closely related to attention and cognition (Deubel and Schneider, 1996). We expected variations in rTT and rFD to affect subjective attributions of human-likeness as well as characteristic features of attentional engagement and cognitive resources related to observing the robot behavior.

2.3.3 Methods

2.3.3.1 Participants

Thirty-four participants were recruited for this experiment (mean age = 25; S.D. = 3.9 years; 21 females). All participants reported normal or corrected-to-normal vision and reported no history of

psychiatric or neurological diagnosis, substance abuse or psychiatric medication. Our experimental protocols followed the ethical standards laid down in the Declaration of Helsinki and were approved by the local Ethics Committee (Comitato Etico Regione Liguria). Each participant provided written informed consent to participate in the experiment. Participants were not informed regarding the purpose of the study before the experiment but were debriefed upon completion.

2.3.3.2 Stimuli

In the current study, we estimated an average human-like TT by applying the model proposed by Baloh et al. (1975), using the following formula: $TT = 37\ ms + 2.7 * |\alpha|$, where α represents the visual angle (in degrees) between two targets. For small angles, the calculated trajectory times would be beyond the iCub physical constraints, thus we selected a visual angle of 60 degrees between the two joint positions. Based on this angle, an average TT of 200 ms was estimated. In the present study, we manipulated the velocity profile and the periodic state of TT. While the first refers to the speed of the eye movement, the latter refers to the variability displayed during the trial.

Overall, we considered the implementation of the following behavior in the iCub robot:

(1) A fixed, behavior, showing no variability, calculated as an "average human behavior".

(2) A "human-range variable" behavior based on literature and human's eye models, showing variability.

(3) A "slow-range variable" behavior, designed to be considerably slower than the human-range behavior.

This design allowed us to define the effect of periodic state manipulation by comparing behavior (1) and (2), and the effect of velocity profile by comparing behavior (2) and (3).

Consequently, we defined three conditions for rTT: a fixed behavior (F_{TT}) during which rTT was constant, a human-range variable (HRV_{TT}) behavior and a slow-range variable behavior (SRV_{TT}). With regard to rFD, previous studies suggested that for humans, the typical pause time between two subsequent eye movements (i.e. fixation duration) is approximately 200 ms (Salthouse and Ellis, 1980). We decided to refer to this lower bound and to explore the same variability range adopted for rTT. Therefore, we adopted the same approach used for rTT, defining three conditions for rFD: a fixed behavior (F_{FD}) during which rFD was constant, a human-range (HRV_{FD}) behavior during which rFD was variable (Andrews and Coppola, 1999), and a slow-range variable behavior (SRV_{FD}). Then, we combined these two factors in a 3 X 3 repeated-measures design, generating 9 final conditions (Table 1).

		Robot Fixation Duration (rFD)		
		300 ms	200-400 ms	400-600 ms
Robot Trajectory Time (rTT)	200 ms	F_{rTT} - F_{rFD}	F_{rTT} -HRV_{rFD}	F_{rTT} - SRV_{rFD}
	100-300 ms	HRV_{rTT} - F_{rFD}	HRV_{rTT} -HRV_{rFD}	HRV_{rTT} - SRV_{rFD}
	300 - 500 ms	SRV_{rTT} - F_{rFD}	SRV_{rTT} -HRV_{rFD}	HRV_{rTT} - SRV_{rFD}

Table 1. Summary of the experimental conditions for rTT and rFD.

Based on our experimental conditions, we implemented the nine different behaviors in the iCub robot. The behaviors were filmed using a 4K Handycam FDR-AX53 by Sony (Minato, Tokyo, Japan). Each video started with the robot looking straight-ahead (2 sec). Then, the eyes started moving from the initial position (*I*) to either position *A* (right) or *B* (left). Then, the eyes moved from one position to the other, for another ten seconds. Immediately after, the robot eyes returned to the *I* position. Subsequently, the head of the robot turned either to the right or to the left with a 70 degrees amplitude. Eventually, the head and the eyes returned to the initial position. The manipulation of occasional head movement was introduced in order to test whether participants would engage in a spontaneous attentional following (measured by fixations that would land laterally with respect to the face, in the same direction as the head movement) and whether their spontaneous attentional following would depend on the robot behavior. Importantly, for the measures of the spontaneous attentional following would depend on the robot behavior. Importantly, for the measures of the spontaneous attentional following, the stimuli remained identical across conditions(i.e., the head movement was always the same). Therefore, any differential effects would be due to some sort of "priming" by preceding robot behavior (fixed, variable, human-range or slow-range). Each behavior was filmed twice (one starting from I to A and one starting from I to B). Consequently, we filmed 18 videos to be used for the experiment.

2.3.3.3 The iCub's gaze controller

In this study, we used the Cartesian 6-DoF gaze controller developed for the iCub robot (Roncone et al., 2016). We implemented this to control the trajectories of the neck (TN) and of the eyes (TE) of

the robot for looking at 3D Cartesian fixation points in space. This controller fits well with our requirements since it allows specifying the point-to-point execution time for the neck (TN) and the eyes (TE).

Therefore, we implemented the robot's eye movements simply by tuning the TE parameter, considering the nine different conditions shown in Table 1.

We implemented a Python script interfacing with the IGazeController Yarp class (Paul et al., 2014). The functions for controlling both the eyes and the neck are shown in Listing 1. In the moveEyes method, we controlled the robot to move the eyes between two pre-defined fixation points for a total duration of 10 seconds.

The fixation points are provided in relative angles (azimuth, elevation, and vergence) according to the controller's specifications. Once 10 seconds had elapsed, the moveNeck method was called for shifting the gaze in a pre-defined location (to the left or right). In this case, we released the block on the neck to let the robot rotate the head along the yaw angles.

```
def moveEyes(FD_MIN, FD_MAX, TT_MIN, TT_MAX):
    IGazeControl.blockNeckYaw()
    IGazeControl.blockNeckPitch()
    IGazeControl.blockNeckRoll()
    while True:
        fd = random.uniform(FD_MIN, FD_MAX)
        tt = random.uniform(TT_MIN, TT_MAX)
        fp = gaze_positions.next()
        IGazeControl.setEyesTrajTime(tt)
        IGazeControl.lookAtRelAnglesSync(fp)
        time.sleep(fd)

def moveNeck():
    IGazeControl.blockNeckYaw()
    IGazeControl.lookAtRelAnglesSync(YAW_POS)
    time.sleep(DISTRACTOR_TIME)
    IGazeControl.lookAtAbsAnglesSync(INIT_GP)
```

Listing 1: Python functions for controlling the eyes and the neck of the robot in the different condition of the task.

2.3.3.4 Apparatus

The experimental session took place in a dimly-lit room. Stimuli were presented on a 22" LCD screen (resolution: 1366 x 768). A chinrest was mounted on the edge of a table, in order to maintain a distance of 63 cm between participants' eyes and the screen for the entire duration of the experiment. Consequently, the forward-looking robot's face subtended 5.5º by 7.1º of visual angle. We used a screen-mounted SMI RED500 eye-tracker by iMotion (Boston, Massachusetts, USA) with a sampling

rate of 500 Hz and spatial accuracy of 0.4º to record binocular gaze data. The experiment was programmed in and presented with OpenSesame 3.1.8 (Mathot, Schreij and Theeuwes, 2012) using the Legacy backend and the PyGaze library.

2.3.3.5 Procedure

We instructed participants to carefully watch the videos and to evaluate, on a 6-point scale how much the behavior displayed by the robot was human-like (0=not at all, 5=extremely). Each video was repeated six-times across the whole experiment (108 trials in total) in a random order of presentation. The experiment consisted of four blocks of 27 trials to allow for self-paced breaks. During each trial, log messages were sent to the eye-tracker at the video onset, the onset of the robot head movement, and the video offset. For details on the trial structure, see Fig. 1.

Prior to the task and before starting the second half of the experiment, a 9-point calibration and 4-point validation thereof were carried out (mean accuracy = 0.89º; S.D. = 0.70). Additionally, participants were recalibrated when deemed necessary (e.g. when a participant moved their head from the chinrest).

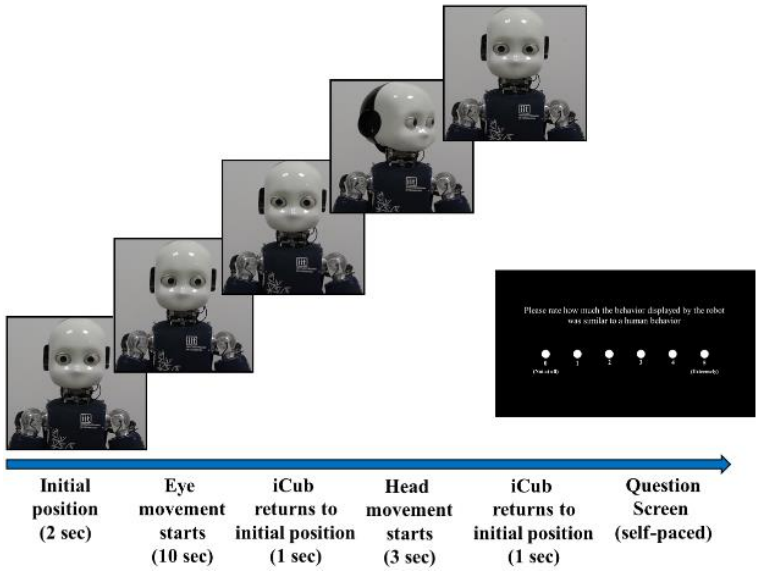

Fig. 1. A planar view of the iCub robot used in this study. The picture depicts the typical trial sequence, illustrating the iCub's behavior (staggered panels) and the subsequent human-likeness rating screen (bottom-right). In this example, the robot moves from *I* to *A* and turns the head to the right.

2.3.3.6 Analyses

In order to explore the potential effects of rTT and rFD on participants' ratings, a mixed model was applied in R Studio. We considered participants' responses as the dependent variable and intercept as a random factor. Then, we computed rTT and rFD as fixed factors of the model. This procedure allowed us to explore the main effects of the single factors and the effect of interaction between the two. Pairwise post hoc comparisons were estimated using the Tukey method.

For the purpose of investigating whether the eye movements implemented in the iCub controller affected participants' eye movements during the videos, analyses on the eye-tracking data were performed. Three participants (mean age = 24.67; S.D. = 1.53; 1 female) were excluded from these analyses due to technical errors in the eye-tracker recordings. In order to facilitate processing of the eye-tracking data, we defined an Area of Interest (AoI): the eye region of the robot. (see Fig. 2 for details).

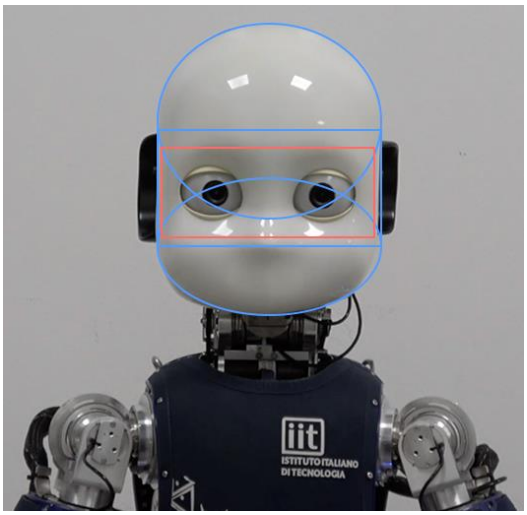

Fig. 2. A frame of the trial sequence depicting the iCub robot. The figure indicates the Area of Interest (AoI)used for the analyses, defined as the red rectangle.

For each participant, we calculated proportional dwell time (the amount of recorded gaze samples within the AoI regardless of eye movement type) and total fixation count to investigate *where* our participants attended. We examined the average fixation duration to underpin the *temporal* characteristics of these mechanisms on the AoI per condition between the onset of the video and the iCub's head movement. Finally, we investigated whether a fixation lateral to the face occurred within 2,500 ms after the robot head movement in each trial as a measure of spontaneous gaze following.

We then processed the landing position -the horizontal vector- of the first lateral fixation in the same direction as the head movement.

Considering the skewed distribution of our data, we computed these metrics as the dependent variables of different Mixed Models. Each model's output provided us with predicted values of all the metrics input as dependent variables. Such predicted values derive from raw data and were corrected based on the effects taken into account in the Mixed Models. In order to maximize comprehensibility and graphical rendering of the effects, we plotted predictive values instead of the raw data. Subsequent pairwise post hoc comparisons were estimated using the Tukey method.

2.3.4 Results

2.3.4.1 Subjective reports

For the human-likeness ratings, we found that the slow-range variable (SRV) condition was evaluated as most human-like. We observed a main effect of condition both on rTT (F=7.67, p<0.001, Fig. 3, left) and on rFD (F=34.61, p<0.001, Fig. 3, right).

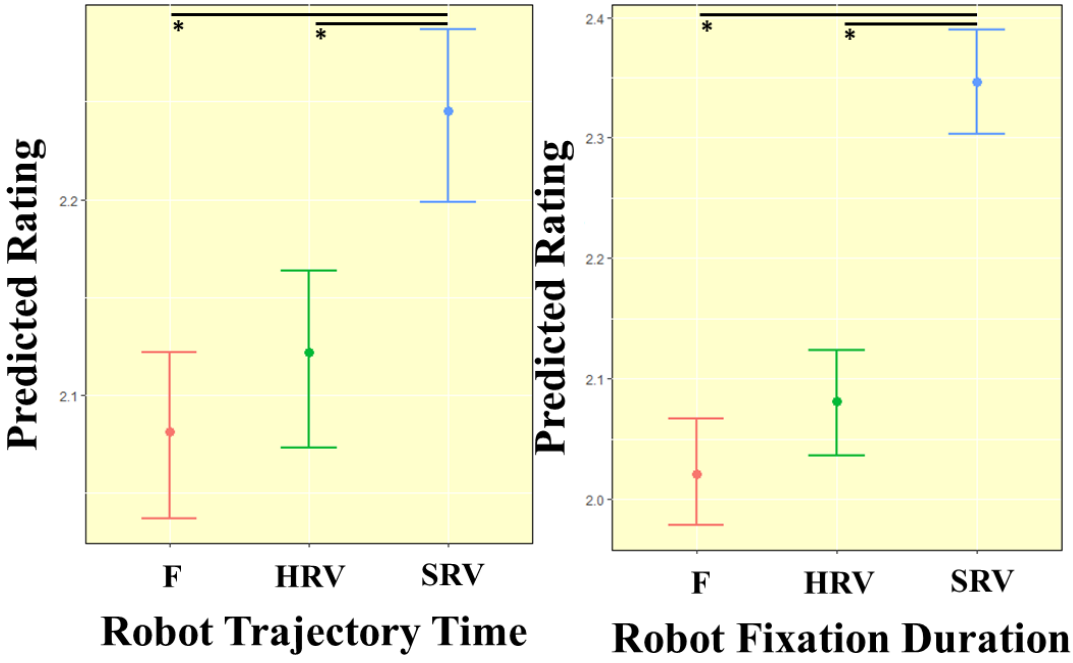

Fig. 3. Human-likeness ratings across the study. On the left: robot's Trajectory Time (rTT) main effect on participants' ratings (F: Mean=2.08, S.D.=1.30; HRV: Mean=2.12., S.D.=1.29; SRV: Mean=2.25, S.D.=1.32). On the right: robot's Fixation Duration (rFD) main effects on participants' ratings (F: Mean=2.02, S.D.=1.12; HRV: Mean=2.08., S.D.=1.35; SRV: Mean=2.35, S.D.=1.33).

Predicted values are plotted on the y-axes in order to facilitate the interpretation of results. Vertical bars denote +/- 1 standard error. Horizontal bars denote differences surviving post hoc comparison.

Post hoc comparisons revealed that our participants evaluated the SRV condition significantly more human-like than the other two conditions. No significant differences were found between the fixed and human-range conditions (see Table 2 for detailed comparisons).

	Contrast	z.ratio	p.value
Fixation Duration	F_{rFD} vs HRV_{rFD}	-1.377	0.3528
	F_{rFD} vs SRV_{rFD}	-7.803	<0.0001*
	HRV_{rFD} vs SRV_{rFD}	-6.419	<0.0001*
Trajectory Time	F_{rTT} vs HRV_{rTT}	-0.614	0.8123
	F_{rTT} vs SRV_{rTT}	-3.660	0.0007*
	HRV_{rTT} vs SRV_{rTT}	-3.046	0.0066*

Table 2. Z ratio and p values resulting from the post hoc comparisons for the human-likeness ratings; asterisks denote the significant results.

2.3.4.2 Objective measures: eye-tracking data

The eye-tracking data is visualized in Fig 4. Heat maps suggested the presence of differences in fixation patterns due to TT manipulation.

Our results in the objective implicit eye-tracking measures showed a somewhat different pattern than the subjective explicit reports.

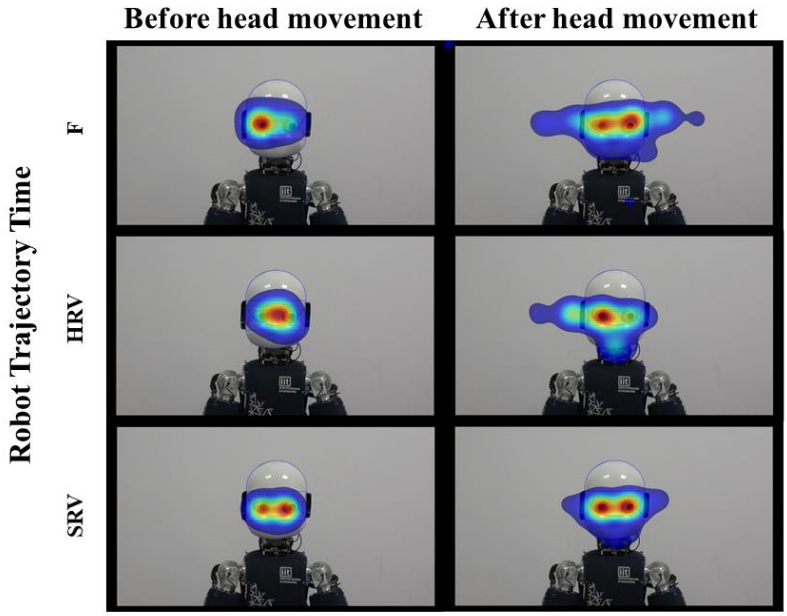

Fig. 4 Fixation heat maps for each rTT condition across the experiment. On the left: trial onset until head movement; on the right: head movement until the end of the trial).

Human-range variability condition (HRV). Importantly, the eye-tracking data showed that at the implicit level of processing, human-range variability (HRV) engaged participants' attention more than the slow-range variability (SRV) – a differential effect on fixation durations, and evoked higher degree of spontaneous joint attention than the other two conditions – an effect on the range of lateral fixations and speed of attentional following. In more detail, our results showed that participants fixated on the eye region longer for the HRV condition, as compared to SRV, as evidenced by the significant main effect of rTT for the eye region ($F=4.84$, $p=0.01$) in the average fixation duration, and significant difference between HRV and SRV, ($z=3.03$, $p=0.01$), planned comparison. No significant differences were found between F and HRV ($z=-2.13$, $p=0.08$) or between SRV and F ($z=0.93$, $p=0.63$) (Fig. 5. left). Furthermore, participants showed a higher degree of spontaneous attentional following in the HRV condition, relative to the other conditions: analyses on lateral fixations recorded after the robot head movement revealed a significant main effect of rFD on the horizontal vector of the first fixation location ($F=4.41$, $p=0.01$). Planned comparisons revealed a significant difference between F and HRV condition ($t=-2.81$, $p=0.01$). A marginal difference was found between HRV and SRV condition ($t=2.24$, $p=0.07$), while no differences were found between F and SRV ($t=-0.58$, $p=0.83$) (Fig 5, right).

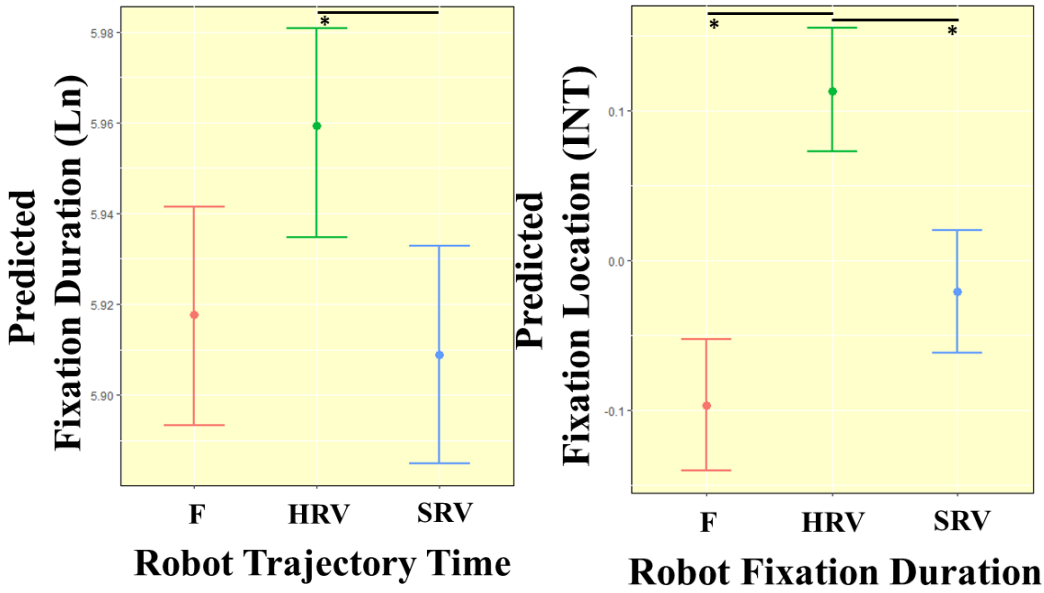

Fig. 5. Fixation patterns across the study. On the left: robot's Trajectory Time (rTT) main effect on participants' fixation duration (F: Mean=439ms, S.D.=309ms; HRV: Mean=460ms., S.D.=321ms; SRV: Mean=428ms, S.D.=275ms). Log-transformed (Ln) predicted values are plotted on the y-axis in order to facilitate the interpretation of results. On the right: robot's Fixation Duration (rFD) main effects on the absolute horizontal vector of the first fixation. (F: Mean=2.6°, S.D.=1.8°; HRV: Mean=4.2°., S.D.=3.0°; SRV: Mean=3.8°, S.D.=2.4°). Location was estimated as the distance from the center of the iCub face. Predicted Blom inverse normal transformed values (INT) are plotted on the y-axis in order to facilitate the interpretation of results. Vertical bars denote +/- 1 standard error. Horizontal bars denote differences surviving post hoc comparison.

Fixed-behavior condition (F). Interestingly, also the fixed behavior condition elicited distinctive gaze patterns, compared to other conditions, which was also – similarly to the HRV condition - not in line with the subjective ratings.

Specifically, we found that participants' gaze was directed toward the eye region significantly more (in terms of dwell times) during the Fixed-behavior condition, than during the other two conditions, as evidenced by a main effect of rTT on participants' proportion of dwell times on the eye region of the iCub (F=4.01, p=0.02), and the significant differences between SRV (z=2.48, p=0.04) and HRV (z=2.43, p=0.04) conditions (Fig. 6, left). Analyses did not reveal any effects of rFD on eye region dwell time or of either TT or FD on the dwell time on the whole face.

Furthermore, a significant main effect of rTT was found on the number of fixations that occurred in the eye region (F=4.46, p=0.01). In the Fixed-behavior condition, the number of fixations was larger

than in the HRV condition (z=2.96, p=0.01). No significant differences were found between SRV and F (z=1.12, p=0.50) or HRV and SRV (z=-1.82, p=0.16) (Fig. 6, right).

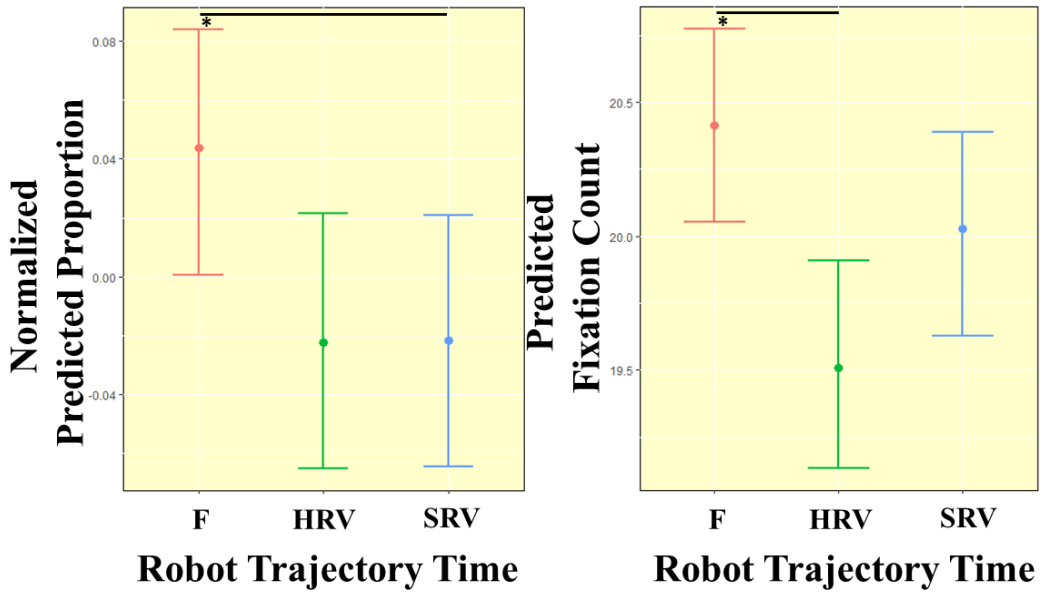

Fig. 6. Fixation patterns across the study. On the left: robot's Trajectory Time (rTT) main effect on participants' proportional dwell time (eye region) (F: Mean=0.82, S.D.=0.27; HRV: Mean=0.80., S.D.=0.29; SRV: Mean=0.80, S.D.=0.28). Normalized predicted values of dwell time are plotted on the y-axis in order to facilitate the interpretation of results. On the right: robot's Trajectory Time (rTT) main effects on participants' amount of fixations (F: Mean=20.41, S.D.=9.89; HRV: Mean=19.51, S.D.=10.11; SRV: Mean=20.03, S.D.=9.83). Predicted values of fixation count are plotted on the y-axis in order to facilitate the interpretation of results. Vertical bars denote +/- 1 standard error. Horizontal bars denote differences surviving post hoc comparison.

Slow-range variable condition (SRV). The slow-range variable condition elicited a distinctive pattern only in the frequency of instances of attentional following, and this is the only result from the implicit measures that follow the explicit subjective reports. The generalized linear model revealed a significant main effect of rTT on the occurrence of spontaneous gaze following ($X^2(2)$=6.82, p=0.03). Specifically, planned comparisons revealed a significant difference between SRV and HRV (z=-2.46, p=0.04). No difference were found between F and either HRV (z=0.48, p=0.88) and SRV (z=-1.97, p=0.11) (Fig. 7).

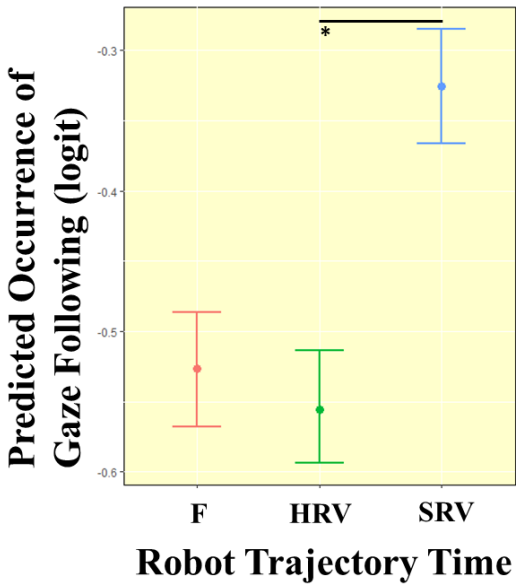

Fig. 7. Spontaneous gaze following of participants during the study. Robot Trajectory Time (rTT) main effect on participants' on the proportion of participants' occurrence of gaze following behavior (F: 35.7% of total F events, HRV: 35.7% of total HRV events, SRV: 39.6% of total SRV events). Logit transformed predicted proportion is plotted on the y-axis. Vertical bars denote +/- 1 standard error. Horizontal bars denote differences surviving post hoc comparison

2.3.5 Discussion

The aim of our study was to examine how various parameters of humanoid eye movements affect the subjective impression of human-likeness and attentional engagement, measured with implicit objective measures (eye-tracking). We manipulated the behavior of the iCub humanoid to display either fixed patterns (fixed trajectory times and fixed fixation durations) or variable trajectory times and fixation durations with a human-range variability (HRV) or a slow-range variability (SRV).

Our results showed that the SRV elicited the highest degree of human-like impression, as reported in subjective ratings. Interestingly, when asked to elaborate on their choices, 59% of our sample reported that the "slower" behavior showing variability seemed to be more natural than the others were. Some of them reported that this specific behavior seemed to be fluid, while the "faster" behavior seemed "glitchy". We speculate that when humans approach a robot, they automatically adopt most available strategies to interpret and predict robot behavior (see Marchesi et al., 2019 for a more elaborate argumentation along these lines). These strategies might be influenced by prior assumptions,

knowledge[3], expectations (all not necessarily realistic) that participants have regarding how a human-like behavior looks like.

Importantly for the purposes of our study, the implicit objective measures showed a different, more informative, pattern. The eye-tracking data indicated that HRV attracted more attention and evoked more attentional engagement, as evidenced by longer fixation durations on the eye region in this condition, compared to the SRV condition. Furthermore, the human-range variability affected joint attention, as participants showed a larger degree of following iCub's directional cues (further location of a lateral fixation elicited by the iCub's head movement), as compared to the other conditions. These results show that participants' implicit (perhaps more automatic) attentional mechanisms became (socially) attuned with the robot behavior when it displayed human-range variability and that this kind of behavior elicits more attentional engagement.

On the other hand, the fixed, repetitive, "mechanical" behavior of the robot, although also showed a divergent pattern of results than the explicit subjective measures, affected the cognitive mechanisms of participants in a different way than the HRV condition. Specifically, it induced a larger number of fixations and visits (proportional dwell times), as compared to the other conditions. This might indicate that participants "scanned" iCub's face more (showed a higher number but shorter fixations) in the mechanistic condition, perhaps because the brain perceived it as unnatural and unfamiliar behavior. This is in line with literature investigating immediateness of biological motion recognition, and suggest the existence of low-level processes that we use to discriminate biological and non-biological motion (Dittrich, 1999). We speculate that humans require a higher amount of fixations to scan an agent displaying unnatural behaviors, while fewer fixations are needed when the behavior is biologically plausible.

Finally, the pattern of results observed in the subjective ratings was paralleled by only one implicit measure, namely the proportion of instances of attentional following (Fig. 7). This indicates that perhaps the general frequency of attentional following was detected at the higher-level of cognitive processes, while other, more implicit and subtle cognitive mechanisms were not. This speculation is based on the following reasoning: explicit measures pinpoint cognitive processes that are accessible to conscious awareness, hence they are higher-level than those that can be captured by implicit measures. In our study, participants reported that the SRV condition appeared most human-like. This might have been a consequence of detecting that at a lower level of processing, they followed the

[3] Our participants' education, for example, negatively correlated with the ratings, suggesting that the more a person might be informed about technology, science or research, the more s/he avoids attributing high human-likeness toward an artificial agent (r=-0.35, p=0.04).

head movements of the robot more when it displayed a slower range of eye movements, relative to faster ranges. In contrast, the other measures (i.e. fixation duration, predicted fixation location, Fig. 5) – although clear markers of attentional engagement – were too low-level to reach the conscious (and thereby reportable) level of processing.

An alternative explanation might be that at the higher-level of processing, participants' responses were prone to various biases, such as assumptions regarding what constitutes a "human-like" behavior or expectations related to robot behavior. Those biases might have affected conscious reports. As a consequence, the frequency of following the head movements of the robot (fig. 7) was influenced by those higher-level biases, which would be in line with previous literature on top-down biases in attentional following (Ozdem et al., 2017; Wiese et al., 2012; Wykowska et al., 2014). Interestingly, the more other mechanisms of attentional engagement (reflected by fixation durations and range of following) were not prone to top-down biases, as they were presumably at a much lower-level of processing.

Overall, our results show that explicit subjective reports alone do not provide a comprehensive picture of cognitive mechanisms evoked by observation of (or interaction with) a robot. Objective measures are necessary to complement subjective reports by addressing specific, and often low-level implicit cognitive processes, an argument put forward in previous literature, due to a dissociation that has been observed between explicit and implicit measures (Kompatsiari et al., 2018).

Related to robot implementations directly, the differences we found in human attentional engagement, as well as a subjective impressions evoked by superficially similarly looking conditions hint that users' interaction with a robot can be qualitatively affected by subtle differences in its behavioral design. This will be investigated further in our future work. Our findings suggest different strategies to use for the iCub' gaze controller depending on the type of interaction the scenario requires to establish with the user. T_E values between 300-500 ms may evoke the impression of more 'naturalness' in the robot's movements. Faster eye movements may appear less smooth (T_E values below 300 ms), but if they involve human-range variability (100-300 ms), they should evoke higher attentional engagement. In this context, the T_E default value (T_E = 250 ms) could be reconsidered to be increased. Importantly, the variability of T_E values among different fixations is a parameter which should certainly be considered in robot behavioral design. Our results showed that added variability induces a higher impression of human-likeness, and is more attentionally engaging. A human-like range of trajectory time elicits most attentional engagement, and attunement in the form of spontaneous joint attention.

2.3.6 Conclusions

In summary, our results confirmed that both implicit and explicit measures need to be taken into account when evaluating the user's reception of a robot behavioral design. Our data show that at the level of conscious subjective impressions, the variability of behavior (trajectory times in the case of our experiment) create the most human-like impression. Fixed-time mechanistic trajectories do not only appear as least human-like, but they also induce fragmented, scattered and short "glimpses", which might have a distracting effect on the user and impair smoothness of interaction. Finally, variable robot behavior with human-range of trajectory times attracts attentional focus most, and thereby is most engaging, even though this might not reflect in subjective conscious impressions.

Throughout the present study, we proposed an approach that uses research methods from cognitive psychology to test engineering parameters. Combining such approaches is beneficial for the future of both disciplines, by facilitating the interaction between humans and artificial agents and by improving our knowledge about ourselves.

2.3.7 Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant awarded to A. Wykowska, titled "InStance: Intentional Stance for Social Attunement. Grant agreement No: 715058).

References

Andrews, T. J., & Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. Vision research, 39(17), 2947-2953.

Baloh, R. W., Sills, A. W., Kumley, W. E., & Honrubia, V. (1975). Quantitative measurement of saccade amplitude, duration, and velocity. Neurology, 25(11), 1065-1065.

Dittrich, W. H. (1999, March). Seeing biological motion-Is there a role for cognitive strategies?. In International Gesture Workshop (pp. 3-22). Springer, Berlin, Heidelberg.

Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience & Biobehavioral Reviews, 24(6), 581-604.

H. Deubel, W.X. Schneider, "Saccade target selection and object recognition: Evidence for a common attentional mechanism", *Vision research,* 1996, 36(12):1827-37.

Itier, R. J., & Batty, M. (2009). Neural bases of eye and gaze processing: the core of social cognition. Neuroscience & Biobehavioral Reviews, 33(6), 843-863.

K. Kompatsiari, F. Ciardo, V. Tikhanoff, G. Metta, A. Wykowska, "On the role of eye contact in gaze cueing", *Scientific reports,* 2018, 8(1):17842.

Kobayashi, H., & Kohshima, S. (2001). Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye. Journal of human evolution, 40(5), 419-435.

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

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the Intentional Stance towards humanoid robots?. Frontiers in Psychology.

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.

Metta, G., Sandini, G., Vernon, D., Natale, L., & Nori, F. (2008, August). The iCub humanoid robot: an open platform for research in embodied cognition. In Proceedings of the 8th workshop on performance metrics for intelligent systems (pp. 50-56). ACM.

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.

Paul, F., Elena, C., Daniele, D., Ali, P., Giorgio, M., & Lorenzo, N. (2014). A middle way for robotics middleware. JOURNAL OF SOFTWARE ENGINEERING IN ROBOTICS, 5(2), 42-49.

Pelachaud, C., & Bilvi, M. (2003, September). Modelling gaze behavior for conversational agents. In International Workshop on Intelligent Virtual Agents (pp. 93-100). Springer, Berlin, Heidelberg.

PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eye tracking experiments. Behavior research methods, 46(4), 913-921.

Roncone, A., Pattacini, U., Metta, G., Natale L.: A cartesian 6-DoF gaze controller for humanoid robots. In: Proceedings of Robotics: Science and Systems, Ann Arbor, MI, 18–22 June 2016.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Ruhland, K., Peters, C. E., Andrist, S., Badler, J. B., Badler, N. I., Gleicher, M., ... & McDonnell, R. (2015, September). A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. In Computer Graphics Forum (Vol. 34, No. 6, pp. 299-326).

Salthouse, T. A., & Ellis, C. L. (1980). Determinants of eye-fixation duration. The American journal of psychology, 207-234.

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. PLoS ONE 7(9): e45391

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), e94339

2.4 Publication IV: Mind the eyes: artificial agents' eye movements modulate attentional engagement and anthropomorphic attribution

Ghiglino D.[1-2] , Willemse C. [1], De Tommaso D.[1], Wykowska A.[1]

[1] Istituto Italiano di Tecnologia, Via Morego, 30, 16163, Genova, Italy

[2] DIBRIS, Università degli Studi di Genova, Via Opera Pia, 13, 16145, Genova, Italy

2.4.1 Abstract

Artificial agents are on their way to interact with us daily. Thus, the design of embodied artificial agents that can easily cooperate with humans is crucial for their deployment in social scenarios. Endowing artificial agents with human-like behavior may boost individuals' engagement during the interaction. We tested this hypothesis in two screen-based experiments. In the first one, we compared attentional engagement displayed by participants while they observed the same set of behaviors displayed by an avatar of a humanoid robot and a human. In the second experiment, we assessed the individuals' tendency to attribute anthropomorphic traits towards the same agents displaying the same behaviors. The results of both experiments suggest that individuals need less effort to process and interpret an artificial agent's behavior when it closely resembles one of a human being. Our results support the idea that including subtle hints of human-likeness in artificial agents' behaviors would ease the communication between them and the human counterpart during interactive scenarios.

Keywords: Humanoid robot, Attentional engagement, Intentional Stance, Mindreading, Eye movements

2.4.2 General Introduction

"Deep", "sparkling", "expressive", "curious", "sad": these are only a few of the adjectives that we can use to describe someone's eyes. Some writers even referred to this sense as the window to the soul, as it can provide information related to others' mental states, emotions, and intentions (Vaidya, Jin & Fellows, 2014). Indeed, every one of us has experienced the feeling to resonate with someone else just at first glance, by making eye contact. If we think about our everyday life, for example, it may happen that we meet a stranger and we immediately understand whether he is sad or happy, just by the look in his eyes (Lee & Anderson, 2017). Neurotypical individuals are usually quite sensitive to

the information conveyed by the eyes and are relatively proficient in inferring other agents' mental states using such a limited source of information (Baron-Cohen et al., 2001). For example, When engaged in a joint action with another person, like moving a heavy object, people are spontaneously inclined to monitor the partner's eyes to infer his/her mental states (Huang et al., 2015).

The relevance of the ability to "read" mental states through the eyes has been widely studied in the literature. A number of studies demonstrated that understanding another agent's gaze direction and pattern could be crucial to accomplish a joint task. For example, gaze can cue attention towards an intended object (Sebanz, Bekkering & Knoblich), it can signal interests in an event happening in the environment (Meyer, Sleiderink & Levelt, 1998), and even anticipate motor actions (Johansson et al., 2001). Indeed the ability to understand such cues is fundamental in social environments (Butterworth, 1991).

Thousands of years of interaction contributed to the development of this ability (Hauser, 1996; Scott-Phillips, 2010), to the point that people appear to notice gaze cues even when the agent that is displaying them is artificial (Fiore et al., 2013). This may be due to the spontaneous adoption of cognitive strategies that are similar to those involved in interpreting human-like behaviors displayed by non-human agents (Chaminade et al., 2012). Previous research showed, for example, that biologically plausible eye-movements displayed by an artificial agent engage an individual's attention more than mechanistic movements (Ghiglino et al., 2020a).

We speculate that endowing subtle hints of human-likeness in the behaviors displayed by an artificial agent would promote the implicit association between that agent's behavior and the behaviors individuals experience during every-day interactions (Banks, 2019). Indeed, even the tendency to attribute a mind towards an artificial agent appears to increase linearly with its perceived human-likeness (Krach et al., 2008). Therefore, equipping artificial agents with a behavioral repertoire that is typical of human beings may create the impression that the behavior they display is motivated by mental states and intentions and, consequently, facilitate social attunement (Wiese, Metta & Wykowska, 2017). As a cascade effect, this impression would facilitate the understanding of the behaviors that the artificial agent displays and would increase the chance of attributing anthropomorphic traits to the agent. Understanding how these spontaneous associations work would provide useful insights for artificial agents' developers, and smoothen the inclusion of artificial agents in contexts where the interaction between technology and human is required (Dautenhahn, 2007).

To investigate this role of human-like eye-movements, we designed two screen-based experiments. Specifically, we explored attentional engagement towards a humanoid and towards a human avatar displaying the same behaviors. In the first experiment, we focused on implicit engagement (i.e.

attentional focus, decision times) displayed by individuals while observing the behaviors of the two agents. In the second experiment, we explored individuals' explicit attribution of anthropomorphism towards the robot and the human displaying the same behavior (self-report scales). Finally, we compared the results of both experiments, to understand whether subtle hints of human-likeness affect only implicit processing as well as explicit attribution of anthropomorphic traits.

2.4.3 Experiment 1

Our first experiment investigated whether the appearance of the agent (natural vs artificial), the behavior displayed by the agent (seemingly intentional vs mechanistic), and the context in which the agent is acting (congruent or incongruent with the behavior) modulate spontaneous attentional engagement, during the observation of other agents involved in a task. As a secondary aim, we explored whether these factors affected the ability to recognize an agent's behavior during a decision-making task.

2.4.3.1 Methods

Participants. Fifty-three participants were recruited for this experiment (mean age = 25.2 years SD = 5.0, 37 females). All participants reported normal or corrected-to-normal vision and no history of psychiatric or neurological diagnosis, substance abuse, or psychiatric medication. Our experimental protocols followed the ethical standards laid down in the Declaration of Helsinki and were approved by the local Ethics Committee (Comitato Etico Regione Liguria). All participants provided written informed consent to participate in the experiment.
Due to a technical problem with the eye-tracker, we excluded twenty-one participants from data analyses (more than 30% of their data were corrupted). Excluded subjects were all individuals with corrected-to-normal vision wearing glasses or corrective lenses. Despite passing the calibration procedure successfully, a large portion of their eye-tracking data was not recorded. Therefore, our final sample consisted of thirty-two participants (mean age = 24.5 years ± 3.63, 22 females).

Stimuli. To address the aims of our first experiment, we filmed the face of a human actor while he was either actively Reading a text on a monitor located in front of him ("intentional", highly variable behavior in terms of temporal and spatial dynamics) or passively following a dot that was moving across the same monitor ("mechanistic", repetitive behavior). This latter behavior closely resembled the procedure for Calibrating an eye-tracker, requiring the subject to fixate on a dot that appears on

the screen in several locations. While the actor was filmed, we recorded his eye-movements using a Tobii Pro Spectrum eye-tracker (TobiiAB Stockholm, 2015). The eye-tracker recorded the cartesian coordinates of the gaze point relative to the screen during both actions, at a sampling rate of 600 Hz. We implemented the eye movement kinematics recorded from the human in a humanoid agent, the iCub robot (Metta et al., 2010), which was filmed while emulating the human's behavior. Then, based on the recordings of the human and the iCub, for each agent, we generated two videos where the agents were "Reading a text" and two videos where they were "Calibrating". The duration of each video was fourteen seconds. The videos of the robot were coupled with the videos of the human so that both agents displayed the very same eye-movements (either "Reading" or "Calibrating") at the same time-frequency.

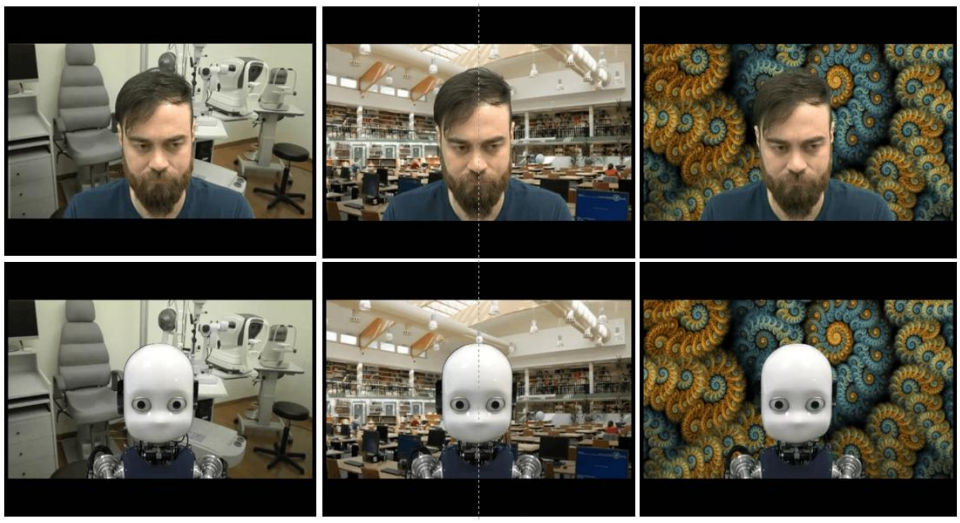

Figure 1 – Examples of videos used in the experiment

Procedure and Apparatus. Before starting the experiment, we informed participants about the content of the videos we generated, showing example videos of both agents displaying the "Reading" and the "Calibrating" behaviors. During this familiarization phase, we informed them that the two displayed behaviors corresponded either to the "Reading" or to the "Calibrating".

During the experiment, videos were presented on a 23.8″ LCD screen (resolution: 1920 × 1080). Participants' head position was limited by a chinrest that was mounted at the edge of the table, at a horizontal distance of 60 cm from the screen. We recorded the participants' binocular gaze data with a screen-mounted Tobii Pro Spectrum eye-tracker with a sampling rate of 600 Hz. The illumination of the room was kept constant throughout the experimental sessions. Videos and questions were displayed with OpenSesame 3.2.8 (Mathôt, Schreij & Theeuwes, 2011).

We instructed participants to carefully watch the videos to detect, as quickly as possible, whether the behavior displayed by the agent was either “Reading” or “Calibrating". Participants provided their responses by pressing the buttons of a keyboard corresponding to the letters M and Z, counterbalanced across the blocks. After providing their response, participants were asked to rate the confidence in their decision. When this last rating was provided, or in case of a timeout, a new trial started, with a fixation cross presented for five seconds.
Participants’ decision times (DTs), the accuracy of the detections, and confidence ratings were saved along with the eye-tracker data (fixation duration and fixation count).

Analyses. To explore the effects of Agent, Behavior, and Context on the participants’ attentional engagement during the task, we adopted various mixed models on our eye-tracking data, using R Studio (RStudio Team, 2015). We defined three main areas of interest (AOI) a priori: (1) the area corresponding to the eye region of the agents; (2) the area corresponding to the face region of the agents (excluding the eyes); and (3) the area corresponding to the background behind the agents (excluding the face). 79.07% of total fixations were recorded within the first AOI (eye region), 5.84% within the second (face region), and 15.09% within the third (background region). Considering the insufficient amount of data points in the non-eye AOIs, we focused our analyses mainly on the eye region. We excluded trials in which the participants provided the incorrect attribution (less than 1% of the total trials) from the analysis.
Fixation duration was the dependent variable of a linear mixed model. Agent, behavior, and context were treated as fixed factors and the subjects’ intercept as a random factor. Then, we converted each participants’ fixation count relative to each AOI into fixation proportions (i.e. the ratio of fixations directed towards each AOI compared to the total number of fixations). Considering the negatively skewed distribution of fixation proportion on the eye-region, data were arcsine transformed before the analyses. Then, the arcsine transformed fixation proportion on the eye region was included as the dependent variable of another mixed model, where agent, behavior, and context were treated as fixed factors and the subjects' intercept as a random factor.
Finally, we analyzed participants’ DTs with an additional linear model. We adopted a minimal a priori data trimming (Harald Baayen, R., & Milin, 2010). Given the positively skewed distribution of DTs, we applied a logarithmic transformation to the data. Then, log-transformed DTs was included as the dependent variable of a final mixed model, where agent, behavior, and context were treated as fixed factors and the subjects' intercept as a random factor.

To compensate for the lack of consensus on the calculation of standardized effect sizes for individual model terms (Rights & Sterba, 2019), for each model we calculated parameters estimated (β) and their associated t-tests (t, p-value) using the Satterthwaite approximation method for degrees of freedom. Furthermore, for each parameter estimated we reported the corresponding bootstrapped 95% confidence intervals. We reported mean values of each dependent variable divided by conditions in the Supplementary Materials to ease the reading of the results. To avoid redundancy, in the main text we reported only statistics relative to significant results. Non-significant results can be found in the Supplementary Materials, along with the original script used for data analysis.

2.4.3.2 Results

Fixation duration. To assess the effect of the Agent, its Behavior and the surrounding Context on attentional processing, we first analyzed the inter-trial differences in fixation duration. No interaction effects were found between Agent, Behavior and Context on fixation duration (all p-values > .05). We found a main effect of the Agent [$\beta = -11.80$, $t(340) = -11.80$, $p = .028$, 95% CI = (-22.14, -1.45); Figure 2A] and a Main effect of the Context [$\beta = 19.01$, $t(340) = 3.52$, $p < .001$, 95% CI = (8.57, 29.45); Figure 2B]. Planned comparisons revealed that longer fixations occurred when the Agent was Human compared with the Robot ($t(340) = 4.73$, $p < .001$), and when the Context was Non-Informative compared with both Congruent and Incongruent contexts(Congruent vs Non-Informative: $t(340) = -7.74$ $p < .001$; Incongruent vs Non-Informative: $t(340) = -7.83$, $p < .001$).

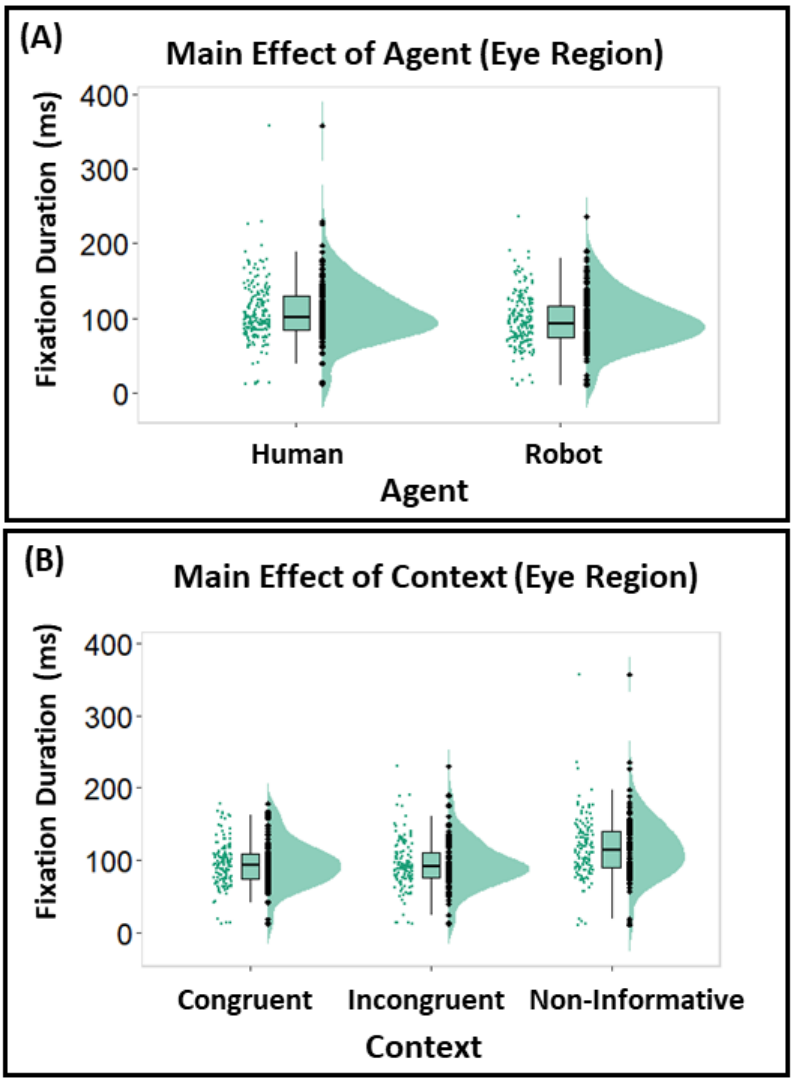

Figure 2 – Raincloud plots showing the fixed effect on Fixation Duration due to the main effects of Agent (A), and Context (B) (GLM). Boxplots associated with the raincloud plots depict median values (black horizontal lines), interquartile ranges (black boxes), and upper-lower quartile intervals (black whiskers).

Fixation proportion. We also analyzed the effects of the Agent, its Behavior and the Context on fixation proportion. Here, our analysis indicated a significant three-way interaction between Agent, Behavior and Context [$\beta = -0.08$, $t(341) = -2.01$, $p = .045$, 95% CI = (-0.16, -0.01); Figure 3A] and a significant two-way interaction between Agent and Behavior [$\beta = 0.10$, $t(341) = 3.65$, $p < .001$, 95% CI = (0.05, 0.16); Figure 3B]. Planned comparisons showed that participants tended to fixate more often on the eye region of the human during the Calibrating behavior rather than during the Reading behavior, when the context was congruent ($t(341) = 5.85$, $p < .001$). A similar difference between the behaviors was found when the context was non-informative ($t(341) = 3.33$, $p = .045$). Similarly,

when the context was congruent, participants tended to fixate more often on the robot than on the human when these agents were displaying the Reading behavior ($t(341) = 5.42$, $p < .001$). Likewise, we found a difference between the agents when the context was incongruent ($t(341) = -3.58$, $p = .020$). These results were confirmed by planned comparisons performed on the two-way interaction, highlighting that the behavior that required fewer fixations was the human's Reading compared to the human's Calibrating ($t(341) = -7.86$, $p < .001$) and to the robot's Reading ($t(341) = -7.01$, $p < .001$). Overall, the Reading behavior required a lower amount of fixations than the Calibrating behavior, as highlighted by the main effect of the Behavior [$\beta = -0.12$, $t(341) = -5.85$, $p < .001$, 95% CI = (-0.16, -0.08)] and subsequent planned comparisons ($t(340) = -7.30$, $p < .001$). The interaction was paralleled by a main effect of the Agent [$\beta = -0.18$, $t(32, 341) = -6.32$, $p < .001$, 95% CI = (-0.24, -0.13)], indicating that participants were faster to identify the behavior when displayed by the robot ($t(341) = 7.87$, $p < .001$). Finally, we found a main effect of the Behavior too [$\beta = -0.22$, $t(32, 341) = -7.71$, $p < .001$, 95% CI = (-0.28, -0.17)], indicating that the Reading behavior was faster to identify than the Calibrating behavior ($t(341) = 12.34$, $p < .001$).

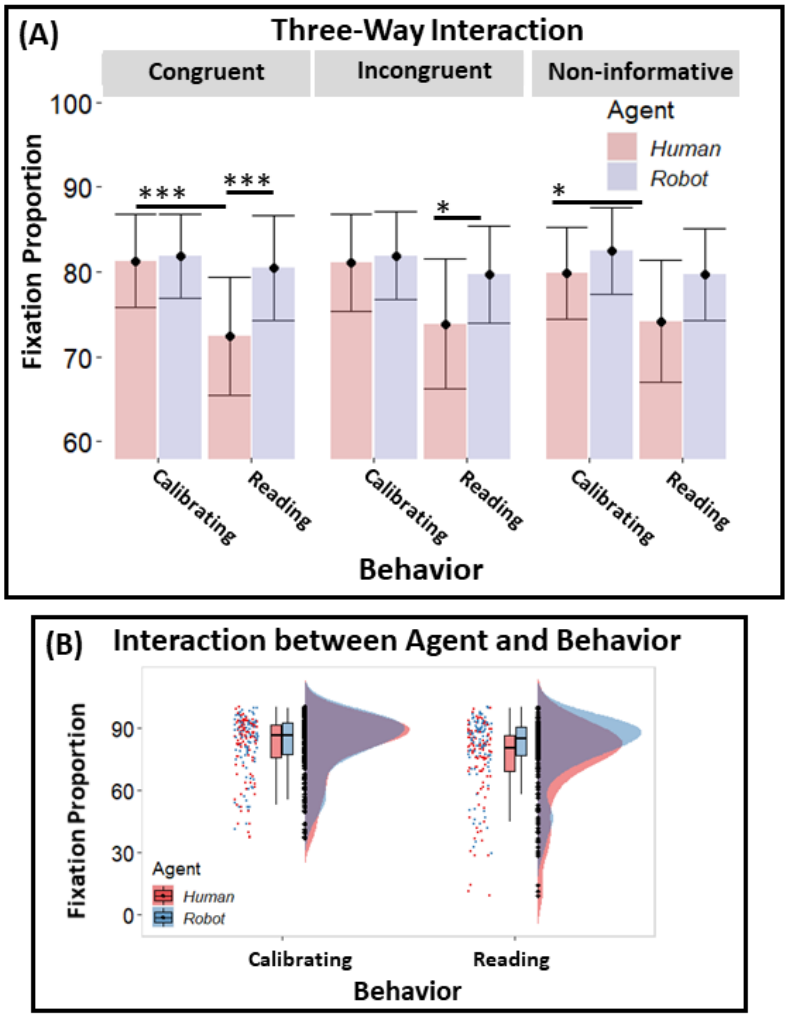

Figure 3 – Histograms and raincloud plots showing respectively the thee-way interaction between Agent, Behavior, and Context (A) and the two-way interaction between Agent and Behavior (B) (GLM). Vertical bars of the histograms denote +/- 1 standard error, dots denote mean values, horizontal bars denote differences surviving post hoc comparison. Asterisks define the level of significance of the comparison (*p<.05, **p<.01, p<.001). Boxplots associated with the raincloud plots depict median values (black horizontal lines), interquartile ranges (black boxes), and upper-lower quartile intervals (black whiskers).

Decision times. To investigate the effect of Agent, Behavior, and Context on the ability to recognize behaviors during the task, we also analyzed our participants' decision times. The analysis pointed out a two-way interaction between the Agent and the Behavior [$\beta = 0.17$, $t(341) = 4.25$, $p < .001$, 95% CI = (0.09, 0.25); Figure 4]. Planned comparisons revealed that the behavior that took longer to identify was the Calibrating behavior displayed by the human when compared to the robot Calibrating ($t(341) = 10.19$, $p < .001$) or to the human Reading ($t(341) = 13.36$, $p < .001$). Furthermore, the

Calibrating behavior displayed by the robot took longer to be identified than the Reading behavior displayed by the same agent (t(341) = 4.10, p < .001).

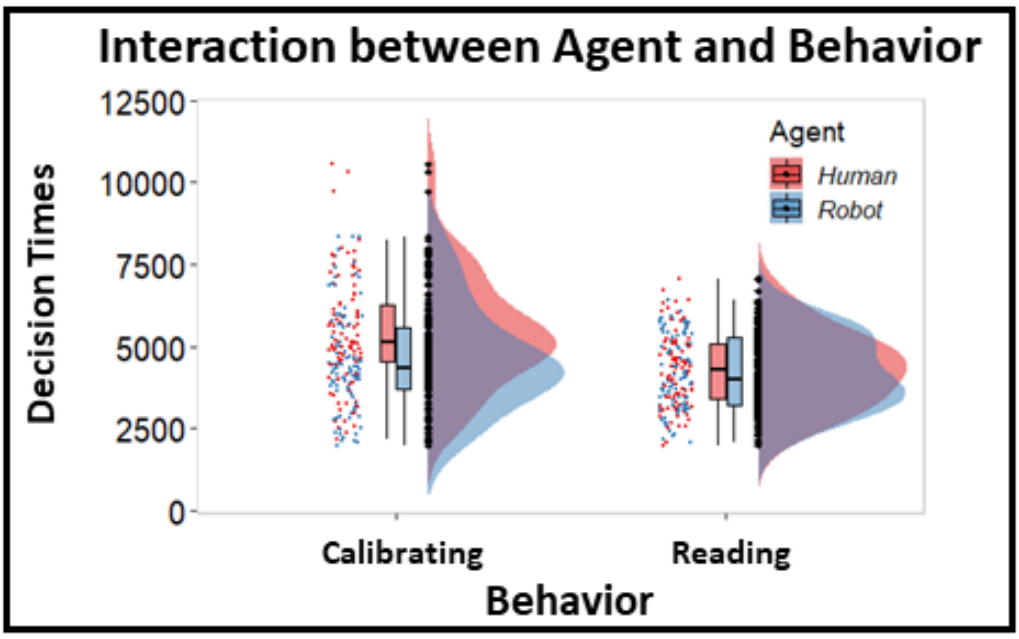

Figure 4 - Histograms and raincloud plots showing the two-way interaction between Agent and Behavior (GLM). Boxplots associated with the raincloud plots depict median values (black horizontal lines), interquartile ranges (black boxes), and upper-lower quartile intervals (black whiskers).

The interaction was paralleled by a main effect of the Agent [β = -0.18, $t_{(32, 341)}$ = -6.32, p < .001, 95% CI = (-0.24, -0.13)], indicating that participants were faster to identify the behavior when displayed by the robot ($t_{(341)}$ = 7.87, p < .001). Also a main effect of the Behavior was observed [β = -0.22, $t_{(32, 341)}$ = -7.71, p < .001, 95% CI = (-0.28, -0.17)], indicating that the *Reading* behavior was faster to identify than the *Calibrating* behavior ($t_{(341)}$ = 12.34, p < .001).

2.4.3.3 Discussion

With this experiment, we investigated attentional engagement during a novel task that required the observation of a human and a robot displaying the same set of behaviors. The results indicated that participants displayed longer fixations towards the eye region of the human compared with the same region of the robot. Fixation duration is often used as an implicit measure of attentional engagement (Ghiglino et al., 2020a; Nummenmaa, Hyönä & Calvo, 2006). Longer fixations are thought to indicate higher interest than shorter ones (Geisen & Bergstrom, 2017). Indeed, a human agent might engage

individuals' spontaneous attention more than an artificial agent, due to the natural acquaintance people have with their conspecifics (Byrne, 1991).

The interactions we found on fixation proportion are in line with this hypothesis. Indeed, participants explored the area surrounding the eyes mainly when the agent was the human, and when he was displaying the Reading behavior. Conversely, individuals explored the face and the background regions less when the agent was the robot relative to the human, and when the behavior was Calibrating compared with Reading. This suggests that, during the task, participants' attentional resources were focused almost solely on the eye movements of the agent when the agent was artificial, and when the behavior was "mechanistic". The ratio of on-target versus all-targets fixations (i.e. the proportion of fixations on a specific area) is often associated with the processing of critical visual information (Holmqvist et al., 2011). We, therefore, conclude that participants required less attentional efforts to interpret the behavior that they were able to relate to the most (i.e. the Reading), especially when the human face, to whom we are more accustomed, displayed it. Indeed, understanding intentional behaviors should be easier than attempting to identify mechanistic ones (Mele & William, 1992).

This is in line with the results we found on participants' decision times. Specifically, we found that Reading behavior was relatively fast to identify, while the Calibrating behavior required more time to be recognized. Importantly for the aim of the study, the condition that costs the longest decision time corresponded to the stimuli where the human was displaying the Calibrating behavior as if observing an "intentional" agent that displays a mechanistic behavior requires higher processing effort. Interestingly, participants were faster in recognizing both behaviors when the robot displayed them than when the human was. This peculiar effect can be explained by taking into account the expectations that individuals might have towards the two agents. From a purely anecdotal point of view, during the debriefing, a small group of participants reported that they were surprised seeing the human behaving "like a robot" (i.e. during the Calibrating behavior). We claim that humans approach artificial agents and their conspecifics with different attitudes that could modulate the way they interpret behaviors (Hinz, Ciardo, Wykowska, 2019). Based on our results, we can also speculate that participants were expecting the robot to display a variety of behaviors (i.e. to behave like a human), but they were not expecting the human to behave in a repetitive, mechanistic way (i.e. to behave like a robot).

Along with the effects of Agent and Behavior, we also found the effect of Context on attentional processing. In particular, when the Context was non-informative, participants' fixations on the eye region were longer. This may indicate that our participants were more engaged by both agents'

behaviors when they were not distracted by the semantic content of the context (i.e. when the context was congruent and incongruent). This is in line with past research investigating the relation between local con global features of visual information (De Cesarei & Loftus, 2011). Indeed, the presence of a "realistic" context might have distracted our participants from the behavior and the agent, attracting their attention towards the background. Thus, the cognitive cost associated with the processing of Congruent and Incongruent backgrounds could explain the presence of shorter fixation on the eye region of both agents. The three-way interaction we found on fixation proportions is in line with this hypothesis; indicating that the interaction between the Agent and the Behavior is particularly strong when the Context is congruent with the Behavior. Thus, we can claim that context could prime the attention towards local cues.

Taken together, these findings highlight the complex interplay between visual information and attentional engagement, suggesting that intentional agents and seemingly intentional behaviors spontaneously attract individuals' attention. However, it might also be the case that the effects we discussed could be biased by familiarity. Perhaps both the Human-agent and the Reading behavior were simply more familiar to the participants than the Robot who was Calibrating, respectively. Indeed, we had to provide examples of the Calibrating behavior to participants before the experiment, as it is not common behavior for a human being. In a natural environment, this kind of behavior is displayed only during medical visits (eye-exam). On the contrary, Reading is an action commonly used in everyday life, and this might have facilitated individuals in the early detection of such behavior. Therefore, the results we found with Experiment 1 might have been biased due to the disparity of the behaviors we selected in terms of prior exposure.

Therefore, after Experiment 1, we needed to clarify whether the effects we found could be explained with reference to the familiarity participants had with the two behaviors, rather than with reference to the degree of intentionality displayed in the behaviors. For this reason, we designed a second experiment, in which we focused more on the self-report impressions that the second group of participants had towards the behaviors used in Experiment 1. Thus, we tested the familiarity of the participants with the behaviors along with their attribution of anthropomorphic traits towards the human and the robot.

2.4.4 Experiment 2

Our second experiment investigated how individuals explicitly interpret the behaviors displayed by two different agents, namely the iCub robot and a human. We exposed our participants to several videos depicting the humanoid and the human engaged in certain activities on a computer, and we

asked them to infer what the agent was doing. We explored our participants' spontaneous attributions as well as their tendency to attribute anthropomorphic traits towards the two agents. This allowed us for a deeper comprehension of the results we found in Experiment 1.

2.4.4.1 Methods

Participants. Fifty participants took part in this experiment and were tested via Prolific (Prolific, 2015), an online recruiting platform (mean age = 26.1 ± 6.0, 20 females). All participants reported normal or corrected-to-normal vision and no history of psychiatric or neurological diagnosis, substance abuse, or psychiatric medication. All participants declared that their first language was English. Each participant provided a simplified informed consent (adapted for online studies) before the beginning of the experiment. Our experimental protocols followed the ethical standards laid down in the Declaration of Helsinki and were approved by the local Ethics Committee (Comitato Etico Regione Liguria).

Stimuli and Apparatus. To address the aim of our second experiment, we used the same pool of stimuli used in Experiment 1, with a few modifications. Since here we were interested exclusively in the interpretation of the behavior as a function of the agent displaying it, we removed the background information from the videos (i.e. we used the original green-screen background). We also included a third behavior that we filmed at the same time as the Calibrating and the Reading behaviors, which corresponded to the agents Watching movies. We excluded this behavior from Experiment 1, as we wanted to have a clear distinction between the active, more "mentalistic" behavior (i.e. Reading) and the passive, more "mechanistic" behavior (i.e. Calibrating). Human eye-movement while watching movies is a visually-guided behavior, but it is not purely stimulus-driven and might constitute a fuzzy category between "intentional" and "non-intentional" behaviors (Peters & Itti, 2007). Furthermore, in Experiment 2 we also wanted to clarify whether the differences between the Calibrating and Reading (found in Experiment 1) were due to the familiarity with the behaviors (i.e. Calibrating being unfamiliar to most of the participants) or to the proprieties of the behavior (i.e. mentalistic vs mechanistic). Thus, by adding Watching, we included an additional behavior that was qualitatively different from the Reading yet with similar familiarity. Consequently, we generated a pool of 12 videos fitting a 2 (agent; human, robot) x 3 (Behavior; Reading, Calibrating, Watching) design.

Procedure. We ran the experiment online, using Prolific to recruit participants and SoSci Survey (Leiner, 2016) to present the stimuli and collect individuals' responses. We instructed participants to

carefully watch the videos depicting the human and the iCub robot engaged in multiple activities on a computer screen. Before the beginning of the experiment, we asked participants to think about all the activities that a person can do with a computer (i.e. playing videogames, browsing, taking part in a meeting, etc.) and that their task would be to infer what the agents depicted in the videos were doing when we filmed them. Participants were allowed to type their answers without a word limit. After providing their attributions, participants were asked to report whether the behavior displayed by the agent looked familiar to them (two-alternative forced-choice: yes/no), and to rate, on a 10-point Likert scale, how much the agent was aware, focused, and interested, as well as the naturalness of the displayed behavior.

Analyses. We extracted the verbs used by the participants to describe the actions depicted in the videos. Then, we converted each verb into its non-personal form (gerund). Thus, for each video, we excerpted fifty verbs describing the behavior enacted by the agent, according to participants' answers. We then performed a text mining analysis on the verbs to determine the frequency of their use across the entire experiment. Then, we compared the frequencies of the most common verbs across conditions, using a series of generalized linear mixed models (GLMM) in R Studio. Agent and Behavior were treated as fixed factors of the model, and the subjects' intercept was treated as a random factor. Given the nature of our dependent variable (frequency of use), Poisson's frequency distribution was used as a reference function for the models.

Separately, we analyzed the familiarity reported by participants with each video. We used a GLMM to compare conditions. Agent and Behavior were treated as fixed factors and the subjects' intercept was treated as a random factor. Since the dependent variable was binary (familiarity), we used the binomial distribution as the reference function of the model.

Finally, we analyzed participants' ratings on their perceived naturalness of the behavior as well as their ratings on perceived awareness, focus, and interest displayed by the agent. Considering the negatively skewed distribution of ratings, data were arcsine transformed before the analyses. Then, we applied a series of linear mixed models (GLM) to investigate the effects of the Agent and Behavior, treated as fixed factors, on the ratings, given the subjects' intercept as a random factor.

To clarify whether the effects found on the ratings could be better explained by participants' familiarity with the behaviors, rather than by our experimental design, we estimated four final alternative linear models that comprised familiarity as the only fixed factor and each rating as a dependent variable. Then, we evaluated the adequacy of each model fit based on a Chi-square difference test and the Akaike's Information Criterion (AIC) associated with each model.

2.4.4.2 Results

The ten most used verbs to describe the agents' behaviors were: reading (count = 207), looking (count = 94), watching (count = 86), playing (count = 36), browsing (count = 25), following (count = 24), staring (count = 17), moving (count = 13), meeting (count = 12), working (count = 12) (Figure 5). Only the first three verbs led to converging models, therefore we excluded all the other verbs from data analysis to avoid overfitting of data (Finnoff, Hergert & Zimmermann, 1993).

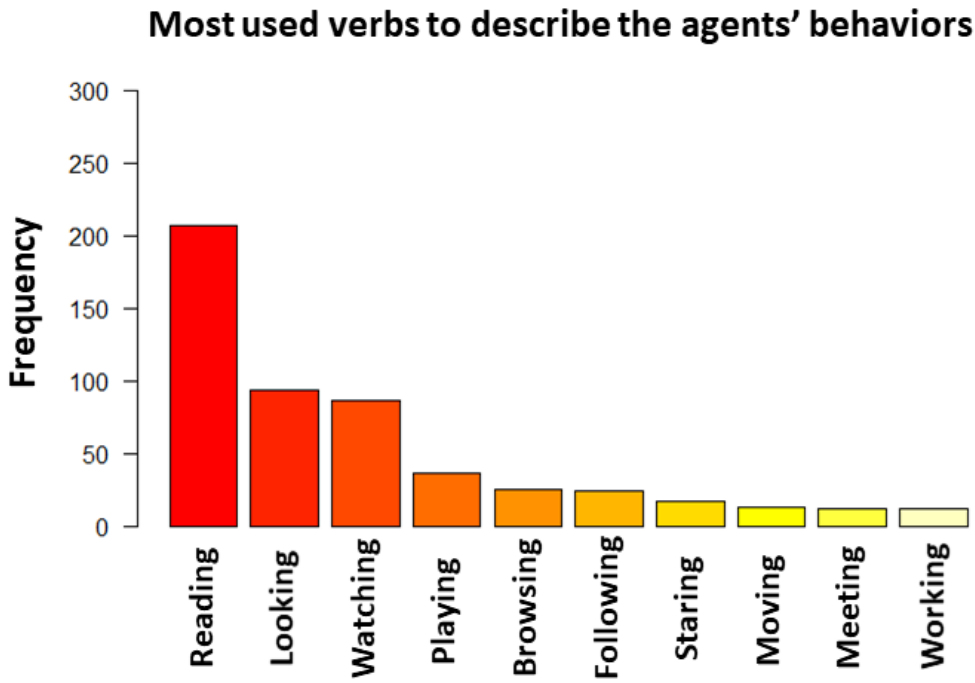

Figure 5 – Frequency plot of the ten most used verbs used by participants to describe the agents' behaviors.

Regarding the frequency of use of the verb "reading", we found a significant main effect of the Behavior [$\beta = 2.69$, $t(293) = 6.37$, $p < .001$, 95% CI = (1.86, 3.51)], indicating that this verb was used significantly more to describe the Reading behavior rather than for the Calibrating ($z(297) = 8.97$, $p < .001$) and Watching ($z(297) = 9.09$, $p < .001$) behaviors. We found a complementary main effect of the Behavior on the frequency of use of the verb "looking" [$\beta = -1.99$, $t(293) = -3.24$, $p = .001$, 95% CI = (-3.20, -0.79)], indicating that this latter verb was used less frequently to describe the Reading behavior than to describe the Calibrating ($z(297) = -4.48$, $p < .001$) or the Watching ($z(297) = -3.84$, $p < .001$) behaviors. We also found a trend of the Behavior on the frequency of use of the verb "watching" [$\beta = 0.58$, $t(293) = 1.74$, $p = .082$, 95% CI = (-0.07, 1.23)] that did not reach significance, but suggested that such verb was used to describe the Watching behavior more often

than for the Reading behavior ($z(297) = 4.33$, $p < .001$). In addition, participants used the verb "watching" slightly more often after the Watching behavior than after the Calibrating one ($z(297) = 2.16$, $p = .078$), and more often after the Calibrating behavior than after the Reading behavior ($z(297) = 2.63$, $p = .023$).

When we analyzed the evaluation of familiarity attributed to the videos, we observed a main effect of both the Agent [$\beta = -1.70$, $t(293) = -3.23$, $p = .001$, 95% CI = (-2.72, -0.67); Figure 6A] and the Behavior [$\beta = 2.74$, $t(293) = 3.15$, $p = .002$, 95% CI = (1.04, 4.45); Figure 6B]. The main effect of the Agent indicated that videos depicting the human agent were rated as more familiar than videos depicting the iCub ($z(297) = 3.94$, $p < .001$). The main effect of the Behavior indicated that the Reading behavior was perceived as more familiar than both the Calibrating ($z(297) = 4.82$, $p < .001$) and Watching ($z(297) = 6.10$, $p < .001$) behaviors. Surprisingly, the Calibrating behavior was evaluated as slightly more familiar than the Watching behavior ($z(297) = 2.38$, $p = .046$).

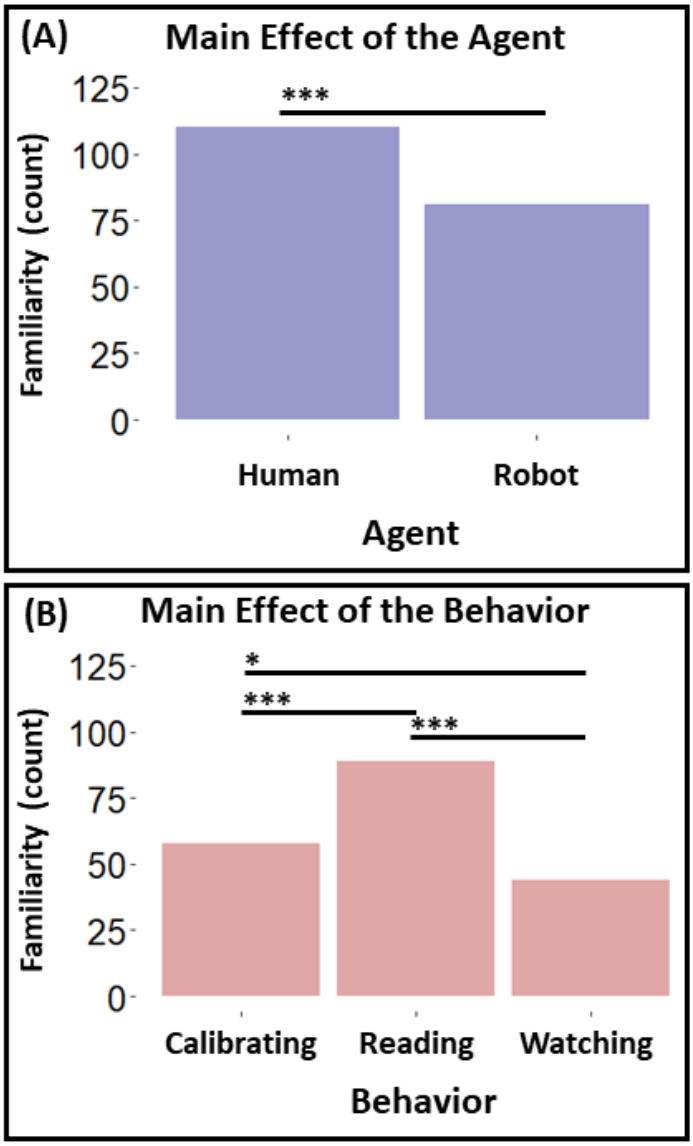

Figure 6 – Histograms representing participants' familiarity with the Agent (a) and with the Behavior (b). Horizontal bars denote differences surviving post hoc comparison, asterisks define the level of significance of the comparison (*p<.05, **p<.01, p<.001).

Our analyses on participants' ratings of anthropomorphic traits pointed out a systematic main effect of the Agent on all the attributes (i.e. "Naturalness", "Awareness", "Focus", "Interest"). Specifically, the human always received higher ratings than the iCub (see Supplementary Materials for detailed comparisons). Furthermore, we found a systematic main effect of the Behavior, indicating that the Reading behavior received higher ratings than both the Calibrating and the Watching behaviors (see Supplementary Materials for detailed comparisons). There was no interaction effect between Agent and Behavior.

Finally, we compared whether the effects on participants' ratings could be better explained by their Familiarity with the behaviors, rather than by the intrinsic characteristics of the Agent and the Behavior. For all comparisons, the most predictive models were the ones including the Agent and the Behavior as fixed factors, instead of the Familiarity (see Table 1 for detailed comparisons).

Table 1 - Detailed Akaike's Information Criterion (AIC) of models for each comparison

Measure	**Fixed Factor(s)**	**AIC**	***χ2***	***p***
Naturalness	*Familiarity*	-1040		
	*Agent*Behavior*	-1085.5	53.52	< .001
Awareness	*Familiarity*	-1056.4		
	*Agent*Behavior*	-1107.3	58.95	< .001
Focus	*Familiarity*	-1108.2		
	*Agent*Behavior*	-1114	13.86	0.008
Interest	*Familiarity*	-959.93		
	*Agent*Behavior*	-984.17	32.248	< .001

2.4.4.3 Discussion

With our second experiment, we tested individuals' familiarity with the behaviors and agents used in Experiment 1. When asked to infer the Agents' actions, participants were highly accurate in identifying the Reading behavior, which we designed to be the "intentional" behavior of our stimuli. Indeed, the eye-movements recorded during the Watching and the Calibrating behaviors were dependent on the occurrence of visual stimuli, or, in other words, to a bottom-up oculomotor capture

(Troscianko & Hinde, 2011). On the other hand, the eye-movements performed during the Reading behavior were actively controlled by the agent himself, who was indeed displaying a top-down modulated action (Radach, Huestegge & Reilly, 2008). Observing an "intentional" behavior (i.e. the Reading behavior in our experiment) may elicit social cognitive mechanisms related to mindreading, which would, consequently, facilitate its identification. This facilitation may sound trivial when applied to a natural human-human interaction, as we usually assume human behavior to be driven by underpinning mental states and intentions (Dennett, 1971). However, it may be less intuitive when applied to artificial agents, as the same facilitation may not apply during observation of robot behavior. Indeed, robots do not possess a proper mind to read, but eliciting the ascription of a mind towards them could foster human-robot interaction, potentially smoothing the communication between natural and artificial systems (Wiese, Metta & Wykowska, 2017). Indeed, our result suggests that for our participants it was easier to identify the correct behavior when the action displayed was seemingly intentional. We claim that embedding intentional behaviors into embodied, artificial agents could boost social engagement by smoothing communication.

In line with this hypothesis, participants rated the Reading behavior as more natural than the other behaviors. Furthermore, when either the human or the robot was displaying it, participants tended to rate the agent as more focused, interested, and aware. This suggests that behavioral cues of intentionality may affect individuals' tendency to attribute anthropomorphic traits towards an artificial agent.

It is important to point out that participants perceived the Reading behavior as the most familiar of the set, regardless of the agent that was displaying it. Additionally, the nature of the Agent affected the attribution of naturalness towards the Behavior, along with the perceived focus, interest, and awareness of the Agent (i.e. participants reported high familiarity with the videos that were depicting the human agent). However, the model comparisons revealed that the nature of the Agent and the Behavior explain our data better than the familiarity ratings alone. This supports the idea that familiarity alone cannot fully explain the differences we found in participants' attributions. At the same time, we recognize that intentionality alone might not be the only factor affecting individuals' attribution of anthropomorphic traits towards natural and artificial agents.

2.4.5 General Conclusions

In the current study, we presented two experiments aimed at investigating how individuals perceive and attribute human-likeness traits towards natural and artificial agents depending upon the level of

"intentionality" displayed by their behaviors. Taken together, the results of both experiments suggest that observing a human and a humanoid displaying the same set of behaviors evokes different implicit attentional processes and, consequently, different explicit attributions.

Our first experiment highlighted the differences in spontaneous attentional engagement during the visual processing of the behavior displayed by the two agents. Processing behaviors that appeared as "intentional" (i.e. controlled by the agent itself) required less attentional effort than the "mechanistic" ones (i.e. purely stimulus-driven). Based on the results of our second experiment, we associate attentional engagement with the attribution of human-like traits towards the agent that displays the behavior. Indeed, in our second experiment, participants evaluated the seemingly intentional behavior as the most anthropomorphic of the set. Additionally, the word-choice participant made to describe the behaviors was extremely accurate for the "intentional" ones, suggesting that it is easier for the observer to recognize the behavior of an artificial agent when the intent behind it is clear. It is important to point out that such facilitation does not depend solely upon the familiarity that participants perceived with the behaviors, but mostly to the degree of perceived intentionality and anthropomorphism. In other words, the degree of intentionality displayed by an artificial agent may affect attentional engagement, which, in turn, affects perceived familiarity and anthropomorphism. Thus, facilitating attentional engagement may be desirable to improve communication with artificial agents.

In this sense, endowing artificial agents with human-like behaviors may boost communication and attunement towards them, a crucial aspect for deploying robots in environments where social interaction is inevitable (e.g., assistive robotics) (Leite, Martinho & Paiva, 2013). Our results bring further clarity to these hypotheses, highlighting the complex interplay between explicit attribution of anthropomorphic traits and attentional engagement. We claim that the attribution of anthropomorphic traits towards an artificial agent is the consequence of the perceived difficulty in processing the information related to its behavior. In turn, such perceived complexity may be modulated by the ease to ascribe intentions towards the artificial agent. However, it is important to point out that clarifying the causal relationship between attentional processing and attribution of anthropomorphism goes beyond the scope of the current work, and should be investigated in future research.

The acceptance of robots as social agents might depend upon their ability to elicit adequately the same social cognitive processes that are required during human-human interaction, even at an implicit level (Dennett, 1971). At the same time, their behavior needs to be easy to predict and to understand from the user perspective (Leite, Martinho & Paiva, 2013). In the last decade, we have been exposed to seemingly smart devices daily. Technological progress made the interaction with technology

increasingly smooth and dynamic due to the implementation of human-like characteristics in the way artificial agents behave and communicate (González, Ramírez, & Viadel, 2012). The implementation of human-based and human-inspired behaviors in artificial agents may positively affect both implicit attentional processing and explicit attributions, and the spontaneity and naturalness of interaction (Dautenhahn, 2007). Furthermore, providing artificial agents with human-like behavior affects positively the quality of the interaction (Ghiglino et al., 2020b). In particular, when the physical aspect and the behavioral repertoire of artificial agents resemble one of the human beings, individuals tend to attribute spontaneously to the agent anthropomorphic traits, including mental states, intentional agency, and anthropomorphic traits (Ghiglino et al., 2020b). Subtle hints of human-likeness displayed by a humanoid robot seem to affect attentional engagement and attribution of anthropomorphic traits (see, for example, Thepsoonthorn, Ogawa & Miyake, 2018; Martini, Buzzell & Wiese, 2015). However, we demonstrated that such claims could not be generalized to all possible behaviors that artificial agents might display during spontaneous interaction with the users.

In conclusion, the current study supports the hypothesis that embedding robots with human-inspired behaviors may facilitate the interaction between them and humans. However, our results suggest that it is not sufficient to generate human-like behavior to ease the interaction. Besides, it may be crucial that the behavior exhibited by the agent displays traits that can be interpreted as intentional.

2.4.6 Acknowledgments

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

References

Banks, J. (2019). Theory of Mind in Social Robots: Replication of Five Established Human Tests. International Journal of Social Robotics, 12(2), 403–414. doi:10.1007/s12369-019-00588-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. Journal of Child Psychology and Psychiatry, 42(2), 241–251. doi:10.1111/1469-7610.00715

Butterworth, G. (1991). The ontogeny and phylogeny of joint visual attention. In A. Whiten (Ed.), Natural theories of mind: Evolution, development and simulation of everyday mindreading (p. 223–232). Basil Blackwell.

Byrne, R. W. (1991). Natural theories of mind: Evolution, development and simulation of everyday mindreading. A. Whiten (Ed.). Oxford: Basil Blackwell.

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. doi:10.3389/fnhum.2012.00103

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human–robot interaction. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1480), 679–704. doi:10.1098/rstb.2006.2004

De Cesarei, A., & Loftus, G. R. (2011). Global and local vision in natural scene identification. Psychonomic Bulletin & Review, 18(5), 840–847. doi:10.3758/s13423-011-0133-6

Dennett, D. C. (1971). Intentional Systems. Journal of Philosophy, 68(4), 87–106. doi:10.2307/2025382

Finnoff, W., Hergert, F., & Zimmermann, H. G. (1993). Improving model selection by nonconvergent methods. Neural Networks, 6(6), 771–783. doi:10.1016/s0893-6080(05)80122-4

Fiore, S. M., Wiltshire, T. J., Lobato, E. J. C., Jentsch, F. G., Huang, W. H., & Axelrod, B. (2013). Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior. Frontiers in Psychology, 4. doi:10.3389/fpsyg.2013.00859

Geisen, E., & Bergstrom, J. R. (2017). Usability testing for survey research. Morgan Kaufmann.

Ghiglino, D., De Tommaso, D., Willemse, C., Marchesi, S., & Wykowska, A. (2020b). Can I get your (robot) attention? Human sensitivity to subtle hints of human-likeness in a humanoid robot's behavior. In Cogsci 2020

Ghiglino, D., Willemse, C., Tommaso, D. D., Bossi, F., & Wykowska, A. (2020a). 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. doi:10.1515/pjbr-2020-0004

González, A., Ramírez, M. P., & Viadel, V. (2012). Attitudes of the Elderly Toward Information and Communications Technologies. Educational Gerontology, 38(9), 585–594. doi:10.1080/03601277.2011.595314

Harald Baayen, R., & Milin, P. (2010). Analyzing reaction times. International Journal of Psychological Research, 3(2), 12–28. doi:10.21500/20112084.807

Hauser, M. D. (1996). The evolution of communication. The MIT Press.

Hinz, N.-A., Ciardo, F., & Wykowska, A. (2019). Individual Differences in Attitude Toward Robots Predict Behavior in Human-Robot Interaction. Lecture Notes in Computer Science, 64–73. doi:10.1007/978-3-030-35888-4_7

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). Eye tracking: A comprehensive guide to methods and measures. OUP Oxford.

Huang, C.M., Andrist, S., Sauppé, A., & Mutlu, B. (2015). Using gaze patterns to predict task intent in collaboration. Frontiers in Psychology, 6. doi:10.3389/fpsyg.2015.01049

Johansson, R. S., Westling, G., Bäckström, A., & Flanagan, J. R. (2001). Eye–Hand Coordination in Object Manipulation. The Journal of Neuroscience, 21(17), 6917–6932. doi:10.1523/jneurosci.21-17-06917.2001

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. doi:10.1371/journal.pone.0002597

Lee, D. H., & Anderson, A. K. (2017). Reading What the Mind Thinks From How the Eye Sees. Psychological Science, 28(4), 494–503. doi:10.1177/0956797616687364

Leiner, D. J. (2016). SoSci Survey. Available at: https://www.soscisurvey.de

Leite, I., Martinho, C., & Paiva, A. (2013). Social Robots for Long-Term Interaction: A Survey. International Journal of Social Robotics, 5(2), 291–308. doi:10.1007/s12369-013-0178-y

Martini, M. C., Buzzell, G. A., & Wiese, E. (2015). Agent Appearance Modulates Mind Attribution and Social Attention in Human-Robot Interaction. Lecture Notes in Computer Science, 431–439. doi:10.1007/978-3-319-25554-5_43

Mathôt, S., Schreij, D., & Theeuwes, J. (2011). OpenSesame: An open-source, graphical experiment builder for the social sciences. Behavior Research Methods, 44(2), 314–324. doi:10.3758/s13428-011-0168-7

Mele, A. R., & William, H. (1992). Springs of Action. Understanding Intentional Behavior. Philosophical Books, 34(2), 116–120. doi:10.1111/j.1468-0149.1993.tb02853.x

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., … Montesano, L. (2010). The iCub humanoid robot: An open-systems platform for research in cognitive development. Neural Networks, 23(8-9), 1125–1134. doi:10.1016/j.neunet.2010.08.010

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: eye movements during noun phrase production. Cognition, 66(2), B25–B33. doi:10.1016/s0010-0277(98)00009-2

Nummenmaa, L., Hyönä, J., & Calvo, M. G. (2006). Eye movement assessment of selective attentional capture by emotional pictures. Emotion, 6(2), 257–268. doi:10.1037/1528-3542.6.2.257

Peters, R. J., & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. doi:10.1109/cvpr.2007.383337

Prolific, Oxford, UK (2015). Available at: https://www.prolific.co

Radach, R., Huestegge, L., & Reilly, R. (2008). The role of global top-down factors in local eye-movement control in reading. Psychological Research, 72(6), 675–688. doi:10.1007/s00426-008-0173-3

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. Psychological Methods, 24(3), 309–338. doi:10.1037/met0000184

RStudio Team, RStudio: Integrated Development for R. RStudio, 25 Inc., Boston, MA (2015). Available at: http://www.rstudio.com/

Scott-Phillips, T. C. (2010). The evolution of communication: Humans may be exceptional. Interaction Studies, Social Behaviour and Communication in Biological and Artificial Systems, 11(1), 78–99. doi:10.1075/is.11.1.07sco

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

Thepsoonthorn, C., Ogawa, K., & Miyake, Y. (2018). The Relationship between Robot's Nonverbal Behaviour and Human's Likability Based on Human's Personality. Scientific Reports, 8(1). doi:10.1038/s41598-018-25314-x

TobiiAB, Stockholm (2015) ‘Tobii Pro Spectrum Product Description’. Available at: https://nbtltd.com/wp-content/uploads/2018/05/tobii-pro-spectrum-product-description.pdf

Troscianko, T., & Hinde, S. (2011). Presence While Watching Movies. i-Perception, 2(4), 216–216. doi:10.1068/ic216

Vaidya, A. R., Jin, C., & Fellows, L. K. (2014). Eye spy: The predictive value of fixation patterns in detecting subtle and extreme emotions from faces. Cognition, 133(2), 443–456. doi:10.1016/j.cognition.2014.07.004

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots As Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social. Frontiers in Psychology, 8. doi:10.3389/fpsyg.2017.01663

Supplementary material

Table 2 – Average Fixation Duration (FD) divided per condition and AOI

Agent	Behavior	Context	Average FD (Eye Region)	Average FD (Face Region)	Average FD (Background Region)
Human	Calibrating	Congruent	101.25	69.04	68.08
		Incongruent	98.91	78.65	66.26
		Neutral	126.87	78.27	91.75
	Reading	Congruent	98.07	61.36	95.31
		Incongruent	99.16	70.39	93.68
		Neutral	120.16	69.85	105.15
Robot	Calibrating	Congruent	89.46	70.54	62.85
		Incongruent	85.76	63.74	63.96
		Neutral	116.28	77.26	84.89
	Reading	Congruent	88.44	75.71	76.36
		Incongruent	92.33	73.45	75.53
		Neutral	103.44	72.36	93.89

Table 3 – Detailed interactions and main effects on Fixation Duration (FD) in the AOI corresponding to eye region

Effect on FD	t - Value	p value	β - Values	C.I. 2.5%	C. I. 97.5%
Agent	-2.21	.028	-11.795	-22.141	-1.449
Behavior	-0.60	.432	-3.18	-13.526	7.166
Context	3.52	<.001	19.012	8.575	29.448
Agent x Behavior	0.29	.746	2.16	-12.472	16.791
Agent x Context	1.03	.991	7.81	-6.879	22.512
Behavior x Context	0.41	.183	3.422	-11.618	17.773
Agent x Behavior x Context	-1.39	.176	-14.905	-35.643	5.832

Table 4 – Average Fixation Proportion (FP) divided per condition and AOI

Agent	Behavior	Context	Average FP (Eye Region)	Average FP (Face Region)	Average FP (Background Region)
Human	Calibrating	Congruent	81.30	4.06	14.64
		Incongruent	81.01	5.12	13.87
		Neutral	79.83	5.71	14.46
	Reading	Congruent	72.47	7.80	19.74
		Incongruent	73.85	6.14	20.01
		Neutral	74.16	6.50	19.34

Robot	Calibrating	Congruent	81.88	5.28	12.84
		Incongruent	81.91	5.73	12.36
		Neutral	82.46	6.02	11.52
	Reading	Congruent	80.46	5.56	13.98
		Incongruent	79.75	5.76	14.50
		Neutral	79.71	6.43	13.86

Table 5 – Detailed interactions and main effects on Fixation Proportion (FP) in the AOI corresponding to eye region

Effect on FP	**t - Value**	**p value**	**β - Values**	**C.I. 2.5%**	**C. I. 97.5%**
Agent	0.26	.799	0.005	-0.033	0.044
Behavior	-5.85	<.001	-0.117	-0.156	-0.079
Context	-1.29	.891	-0.026	-0.065	0.013
Agent x Behavior	3.65	<.001	0.103	0.049	0.158
Agent x Context	1.23	.735	0.034	-0.02	0.09
Behavior x Context	1.79	.075	0.051	-0.004	0.106
Agent x Behavior x Context	-2.01	.045	-0.081	-0.158	-0.003

Table 6 – Average Decision Times (DTs) divided per condition

Agent	**Behavior**	**Context**	**Average DTs**
Human	Calibrating	Congruent	5372.65
		Incongruent	5353.54
		Neutral	5399.91
	Reading	Congruent	4235.94
		Incongruent	4204.50
		Neutral	4316.05
Robot	Calibrating	Congruent	4502.33
		Incongruent	4654.01
		Neutral	4570.06
	Reading	Congruent	4199.00
		Incongruent	4181.08
		Neutral	4167.55

Table 7 – Detailed interactions and main effects on Decision Times (DTs) in the AOI corresponding to eye region

Effect on DTs	**t - Value**	**p value**	**β - Values**	**C.I. 2.5%**	**C. I. 97.5%**
Agent	-6.32	<.001	-0.181	-0.237	-0.126

Behavior	-7.71	<.001	-0.221	-0.276	-0.165
Context	0.06	.958	0.001	-0.054	0.057
Agent x Behavior	4.25	<.001	0.172	0.094	0.251
Agent x Context	0.23	.604	0.028	-0.068	0.088
Behavior x Context	0.35	.679	-0.015	-0.064	0.092
Agent x Behavior x Context	-0.67	.800	-0.038	-0.149	0.072

Table 8 – Detailed effects of Agent and Behavior on participants' ratings of Naturalness, Awareness, Focus, and Interest

Measure	**Effect**	**Contrast**	**β**	$t_{(293)}$	***p***	**2.5 % CI**	**97.5 % CI**
Naturalness	Agent		-0.05	-6.82	< .001	-0.06	-0.03
		Human - Robot	-	11.76	< .001	-	-
	Behav.		0.03	4.74	< .001	0.02	0.05
		Calibrating - Reading	-	-7.14	< .001	-	-
		Calibrating - Watching	-	-1.52	.028	-	-
		Reading - Watching	-	5.62	< .001	-	-
Awareness	Agent		-0.04	-5.41	< .001	-0.05	-0.02
		Human - Robot	-	8.59	< .001	-	-
	Behav.		0.02	3.04	.003	0.01	0.03
		Calibrating – Reading	-	-5.27	< .001	-	-
		Calibrating - Watching	-	1.54	.272	-	-
		Reading - Watching	-	6.81	< .001	-	-
Focus	Agent		-0.03	-5.13	< .001	-0.04	-0.02
		Human - Robot	-	6.75	< .001	-	-
	Behav.		0.02	2.63	.009	0.00	0.03
		Calibrating - Reading	-	-4.77	< .001	-	-

		Calibrating - Watching	-	0.35	.936	-	-
		Reading - Watching	-	5.12	< .001	-	-
Interest	Agent		-0.04	-4.37	< .001	-0.05	-0.02
		Human - Robot	-	7.03	< .001	-	-
	Behav.		0.02	2.73	.007	0.01	0.04
		Calibrating – Reading	-	-3.81	.001	-	-
		Calibrating - Watching	-	1.77	.183	-	-
		Reading - Watching	-	5.57	< .001	-	-

SECTION III - GENERAL DISCUSSION

3 Synopsis of Results

The work described in this thesis aimed at identifying some of the behavioral parameters that might make robots appear more human-like. This, in turn, may facilitate interaction and communication between artificial agents and their users. Specifically, studies reported in Section II could provide useful insights for robot developers and designers, as we highlighted that human-based behaviors boost attentional engagement and the attribution of anthropomorphic traits. As a cascade effect, we speculate that this might maximize attunement with the artificial agent and smooth communication between the agent and its user. Furthermore, we showed the importance of combining both explicit and implicit measures to assess individual differences and how humans behave during HRI scenarios.

Publication I aimed at pinpointing the effects on human-likeness attribution towards a virtual agent (i.e. an avatar of the iCub robot) due to spatial and temporal information of the behavior it displays. Results of the study showed that temporal features of the movement displayed by the agent modulate the attribution of anthropomorphic traits more than spatial features. In detail, considering the set of behaviors implemented in the avatar, slower movements were rated by participants as more human-like. This result suggests that human-likeness attribution does not rely only on pure kinematic patterns, but also on temporal dynamics. We found also a gender difference in our sample (i.e. females tended to rate higher the human-likeness of the avatar than males, regardless of the features of the behavior), supporting the idea that individual differences could bias the attribution of anthropomorphic traits.

Following the results of Publication I, Publication II aimed at investigating more in detail individuals' sensitivity towards human-based behaviors implemented in an embodied artificial agent (i.e. the iCub robot). The results of the study suggest that individuals' a priori knowledge of the behavior displayed by the agent strongly affect their sensitivity in HRI scenarios. Specifically, the information that we provided to our participants before interacting with the robot overrode empirical evidence accumulated during the interaction and affected the attribution of anthropomorphic traits towards the iCub robot. In this study, we also further explored individual differences that could play a role during the interaction with artificial agents. Our results suggest that sociodemographic characteristics (i.e. years of education) and psychological traits (i.e., Conscientiousness and Neuroticism) that play a role during the interaction with other humans may affect also the interaction with humanoid robots or, at least, their attitudes towards them.

In Publication III we tested the use of implicit measures (i.e. eye-tracking measures) to assess individuals' reactions in HRI scenarios, along with self-report questionnaires. The study aimed at identifying potential differences in terms of visual processing and attentional engagement due to

subtle hints of human-likeness displayed by an artificial agent in a screen-based experiment. Results of the experiment showed that participants tended to attribute higher human-likeness to movements that were markedly slow and variable, rather than movements that were within the human temporal range. This result is partially in line with Publication I, as it seems that movements displaying a slower temporal dynamic are perceived as more anthropomorphic than faster ones. Interestingly, our analyses of eye-tracking measures pointed out that attentional engagement was higher when participants observed human-range behaviors rather than when they were observing slow or non-variable behaviors. Furthermore, considerably mechanistic behaviors (i.e. non-variable) recruited the use of a different scan path than human-range behaviors (i.e. variable), characterized by a greater number of fixations. Taken together, our results suggest that the explicit attribution of anthropomorphic traits does not necessarily reflect the easiness of processing behavioral information conveyed by an artificial agent. Indeed, explicit attributions may be biased by individuals' priors towards robotics, which might be too strong to be modulated during short interactions with a robot during experimental protocols in the lab. This claim is in line with what was reported in Publication II.

Publication IV complements results of Publication III, and combines, once again, implicit measures (i.e. eye-tracking data, decision times) with more explicit measures (i.e. self-report questionnaires). The aim of the last study presented in this thesis was to define differences in the processing of behaviors displayed by avatars either of a human being or a humanoid robot (i.e. the iCub robot). In the first experiment of this study, we focused on behaviors that can be defined as intentional (i.e. reading a text on a screen) and semi-intentional (i.e. following a dot moving on a screen. Results of this experiment pointed out a higher attentional engagement towards the human agent rather than the humanoid robot. Additionally, our results suggest that lower processing efforts\ is required to identify the "intentional" behavior. This result is in line with Publication III, as it suggests that, at the attentional level, highly variable human-based behaviors seem to be easier to process than repetitive, mechanistic behaviors (as in the case of following a dot on the screen). This supports the hypothesis that implementing human-based behavior into artificial agents can smoothen communication and, possibly, interaction. Furthermore, Experiment II of Publication IV highlighted that individuals are more accurate in the identification of intentional behaviors displayed by a robot than when it displayed more passive/mechanical behaviors. This is in line with Publication II and III, adding an important finding: implementing pure human-based kinematic patterns in an artificial agent might not be sufficient to elicit anthropomorphic attribution and to facilitate HRI if the intentions behind the behavior are not clear.

3.1 Implications for the investigation of social cognition in human-robot interaction

Robots embedded with social abilities are designed to assist humans in daily routines. Social robots are supposed to interact and support humans in several contexts, from the domestic to the healthcare environments. In order to smoothen the interaction and maximize the support that such technologies can provide, robots should be embedded with enough "social intelligence" to understand the needs of the user and provide an appropriate behavior. Tailoring the behavior of an artificial agent to the requests and needs of the human interacting with it is a huge challenge for social robotics. Aside from the development of appropriate technical solutions, it is crucial to understand, from a cognitive point of view, how humans interact with artificial agents. Indeed, the success of human-robot interaction depends not only on the technical capabilities and behavioral repertoire of the artificial agent but also on the ease with which humans attune and coordinate with it. Previous literature suggested that performance during human-robot interaction (HRI) may be boosted when the human socially attunes with the artificial agent, treating it as if it possessed mental states and anthropomorphic traits (see, for example, Wiese, Metta, & Wykowska, 2017). However, social cognitive mechanisms displayed by humans during HRI scenarios could be influenced also by personality traits, demographic characteristics, cultural belonging, and previous experience with artificial agents. Indeed, the same interaction with the same robot could be perceived as smooth or stressful, depending on the user's biases and prior beliefs. Several individual differences might concur in the attribution of mental states and intentional agency towards artificial agents. Still, the complex interplay between individual differences and artificial agents' behaviors requires further systematic investigation. In particular, it is still unclear whether priors and biases towards social robots could be modulated by exposing individuals to direct HRI. Our results and, in particular, results of Publication II, suggest that exposing individuals to direct HRI might not be sufficient to modulate priors and biases. Conversely, some individuals' dispositions towards artificial agents, including knowledge and experience with robotics, might override empirical evidence provided by the robot behavior and might modulate the attitude displayed during the interaction (Ciardo et al., 2020). Results reported in this thesis suggest the need of assessing such differences more in-depth, to better understand which personality traits and individual factors affect HRI. This includes the adoption of assessment tools that are typical of psychological research, to define whether the factors that affect human-human interaction influence

human-robot interaction as well. Exploring such a research topic is fundamental, as new technologies have become inexorably an integral part of our everyday life. However, it will be not sufficient to investigate HRI solely relying on the assessment of individuals' dispositions and explicit measures of human-likeness attribution. As suggested by our results, HRI research should focus also on the investigation of individuals' automatic (and, often, implicit) reactions during interactive scenarios. The adoption of implicit measures able to discriminate whether an individual engages and attunes with an artificial agent is crucial for maximizing the communication between the two agents.

Indeed, we are exposed to seemingly smart devices on a daily basis, and this makes the interaction with technology increasingly smooth and dynamic (Gonzàles et al., 2012). Recent literature well supports the idea that artificial agents able to behave like human beings facilitate the spontaneity and naturalness of interaction (Wiese, Metta & Wykowska, 2017). Providing artificial agents with human-like behavior affects positively the quality of the interaction (Thepsoonthorn, Ogawa & Miyake, 2018). In particular, when the physical aspect and the behavioral repertoire of artificial agents resemble one of the human beings, individuals tend to spontaneously attribute to the agent anthropomorphic traits, including mental states, intentional agency, (Martini, Buzzell, & Wiese, 2015). The studies included in this thesis support this hypothesis and showed consistently that not only the attribution of human-like traits towards an artificial agent but also spontaneous attentional engagement can be modulated by the behavior it displays. We hypothesize that facilitating attentional engagement would cause, as a cascade effect, the improvement of the communication with artificial agents. Indeed, we consistently showed that intentional, human-like behaviors tend to be easier to process for the participants of our experiments. Thus, embedding artificial agents with human-like behaviors is supposed to boost social attunement towards them, and this aspect could be particularly crucial for deploying social robots in environments where social interaction is desirable (e.g., assistive robotics) (Leite, Martinho, & Paiva, 2013). Indeed, treating artificial agents as if they were anthropomorphic entities with a synthetic "mind" may help individuals during natural interactions, as these processes allow us to predict and explain the behaviors displayed by other individuals in every-day interactions with our conspecifics (Barresi & Moore, 1996). Previous research showed that similar mechanisms might be activated in interaction with artificial agents (Krach et al., 2008), and vary depending upon individual differences and available contextual information (Hinz, Ciardo, Wykowska, 2019). Our results bring further clarity to these hypotheses, highlighting the complex interplay between individual differences, explicit attribution of anthropomorphic traits, mind-ascription, and attentional engagement. In particular, individual differences do not affect the implicit processing of visual information, which might rely on more low-level components of social-

cognition. Indeed, attentional engagement seems to be mainly driven by the behavior of the artificial agent, rather than by the individual's dispositions towards the agent itself. This provides useful insight that might be used for the development of artificial agents that will interact, in the future, with human users on a daily basis. The acceptance of robots as social companions might actually depend on their ability to adequately elicit such social cognitive processes during the interaction, even at an implicit level (Ghiglino et al., 2020). At the same time, social robots need to be able to understand the users' needs, feelings, and intentions in order to adapt their skills and behaviors accordingly (Leite, Martinho, & Paiva, 2013).

Previous studies investigated how attitudes and biases towards robotics could affect the interaction with artificial agents (Perez-Osorio et al., 2019), along with personality traits and cultural differences (Weiss, Evers, 2011). However, such systematic investigation of such individual differences has been scarce so far. Most of the past research relies solely on the administration of questionnaires and online surveys. Nevertheless, biases and attitudes towards social robots could be modulated during HRI scenarios, through the accumulation of empirical shreds of evidence (Cross, Hortensius, Wykowska, 2019). In the current thesis, we explored the hypothesis that prior beliefs and attitudes towards artificial agents may affect social interaction with social robots, which, in turn, could modulate priors and attitudes, in circular logic. Indeed, it is still unclear whether eliciting mind ascription and human-like traits towards social robots during the interaction could reduce negative attitudes and biases towards robots. Considerable progress has been made in recent years in terms of the technical realization of artificial agents. However, further interdisciplinary research is required to complement technical solutions with the understanding of social cognitive processes that are uniquely adopted during the interaction with social robots (Shellen, Wykowska, 2019). Indeed, the ability of robots to socially and intuitively interact with the users is still limited, and future research on the psychological underpinning of HRI could improve both the design of artificial agents and our understanding of social cognitive mechanisms in humans. The work presented in this thesis supports the idea that the investigation of social cognition mechanisms elicited in HRI scenarios needs to comprise the collection of explicit and implicit measures, assessing individual's differences/attitudes towards robots as well as attentional engagement during interaction with artificial agents. In particular, the combination of surveys and questionnaires with eye-tracker and behavioral measures allows for a better understanding of spontaneous social mechanisms elicited by HRI.

3.2 Limitations and Future Directions

Results reported in Section II highlight the importance of adopting a neuroscientific approach to HRI research. The integration of explicit and implicit measures of attunement and anthropomorphic attribution is desirable to explore how individuals interact with artificial agents. However, some limitations emerged during the studies that are reported in this thesis. First, we identified the lack of proper explicit measures that allow for a comprehensive understanding of individuals' propensity to treat artificial agents as mentalistic, human-like entities. Furthermore, despite the existence of validated tools to evaluate participants' attitudes towards artificial agents, we identified a lack of tools that can assess properly individuals' familiarity with robotics, which might affect attunement with a robotic agent. To compensate for these potential limitations, we developed two ad-hoc solutions: (1) the InStance questionnaire (Marchesi et al., 2019), which is aimed to evaluate individuals' tendency to describe the behavior of an artificial agent with reference to mental states and intentions or with reference to its functioning; (2) the RobEx questionnaire (Perez-Osorio et al., 2019), which is aimed to assess individuals' mastery with robotics. These measures provide useful pieces of information that should be taken into consideration for future research in HRI. However, future research should investigate more in-depth the relationship between such measures and other individuals' characteristics that might affect HRI (i.e., personality traits, general attitudes towards robots), to provide the HRI community with a comprehensive battery of tests that can properly assess individuals' dispositions.

Another potential limitation of the studies proposed in this thesis is the nature of the presented experiments, which were either observational or screen-based. Indeed, natural and spontaneous interactions with artificial agents might lead to different social-cognitive processes than those reported in Section II. In particular, we speculate that the exposure to artificial agents that display hints of human-likeness during more interactive scenarios may boost social attunement, thus overriding individual priors and biases. Indeed, future research should investigate scenarios in which the "intentions" of the agents become crucial for the execution of a task. However, it is important to point out that the work reported was meant to lay grounds for further research by providing well-controlled experiments addressing specific and "dissected" parameters of artificial agent behavior. Further research can extend these paradigms into more ecologically valid scenarios.

3.3 Conclusions

In the near future, social robots could assume roles of great relevance in the assistive care for vulnerable people, such as elderly or hospitalized patients. Therefore, the development of tailor-made

solutions that can ease the interaction between social robots and their users is crucial. In the current work, we proposed an innovative approach to the investigation of social cognition mechanisms that use the combination of technological solutions (robotics) and neuroscientific methods (eye-tracking). Indeed, before implementing robots with the ability to adapt to their users, it is fundamental to identify (1) individual differences that play a role in social attunement toward artificial agents and (2) specific robot behaviors that might facilitate social attunement, attentional engagement, and communication. From one side, a deep understanding of individuals' priors and biases towards robotics would optimize technical solutions aimed to bolster social interaction. In particular, it is crucial to understand the user's behavioral and personality indicators of reduced comfort, stress, and poor mental state attribution during HRI scenarios, as such factors could undermine the social acceptability of artificial agents (Whelan et al., 2018). At the same time, it is fundamental to investigate what are the characteristics of artificial agents' behavior that play a role in the ascription of mental states and intentions, as those might facilitate predictability and explainability of robot behavior, thereby boosting social attunement.

References

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, 287(1925), 20200244.

Aggarwal, J. K., & Cai, Q. (1999). Human motion analysis: A review. Computer vision and image understanding, 73(3), 428-440.

Andrews, T. J., & Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. Vision research, 39(17), 2947-2953.

Bahadori, S., Cesta, A., Grisetti, G., Iocchi, L., Leone, R., Nardi, D., ... & Rasconi, R. (2003, September). Robocare: an integrated robotic system for the domestic care of the elderly. In Proceedings of Workshop on Ambient Intelligence AI* IA-03, Pisa, Italy.

Baloh, R. W., Sills, A. W., Kumley, W. E., & Honrubia, V. (1975). Quantitative measurement of saccade amplitude, duration, and velocity. Neurology, 25(11), 1065-1065.

Banks, J. (2019). Theory of Mind in Social Robots: Replication of Five Established Human Tests. International Journal of Social Robotics, 12(2), 403–414. doi:10.1007/s12369-019-00588-x

Baron-Cohen, S., Campbell, R., Karmiloff-Smith, A., Grant, J., & Walker, J. (1995). Are children with autism blind to the mentalistic significance of the eyes?. British Journal of Developmental Psychology, 13(4), 379-398.

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. Journal of Child Psychology and Psychiatry, 42(2), 241–251. doi:10.1111/1469-7610.00715

Bartneck, C. (2008). The Godspeed Questionnaire Series.

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.

Bemelmans, R., Gelderblom, G. J., Jonker, P., & De Witte, L. (2012). Socially assistive robots in elderly care: A systematic review into effects and effectiveness. Journal of the American Medical Directors Association, 13(2), 114-120.

Bérubé, M. (2013). A Theory of Theory of Mind. American Scientist, 101(2), 148.

Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., & Pozzo, T. (2014). Motor contagion during human-human and human-robot interaction. PloS one, 9(8), e106172.

Blakemore, S. J., & Decety, J. (2001). From the perception of action to the understanding of intention. Nature Reviews Neuroscience, 2(8), 561.

Brentano, F. (1874/1911/1973). Psychology from an Empirical Standpoint, London: Routledge and Kegan Paul.

Broekens, J., Heerink, M., & Rosendal, H. (2009). Assistive social robots in elderly care: a review. Gerontechnology, 8(2), 94-103.

Burgos, J. E. (2007). About aboutness: Thoughts on intentional behaviorism. Behavior and Philosophy, 65-76.

Butterworth, G. (1991). The ontogeny and phylogeny of joint visual attention. In A. Whiten (Ed.), Natural theories of mind: Evolution, development and simulation of everyday mindreading (p. 223–232). Basil Blackwell.

Byrne, R. W. (1991). Natural theories of mind: Evolution, development and simulation of everyday mindreading. A. Whiten (Ed.). Oxford: Basil Blackwell.

Carlson, K., Wong, A. H. Y., Dung, T. A., Wong, A. C. Y., Tan, Y. K., & Wykowska, A. (2018, November). Training Autistic Children on Joint Attention Skills with a Robot. In International Conference on Social Robotics (pp. 86-92). Springer, Cham.

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. doi:10.3389/fnhum.2012.00103

Chevalier, G., & Deniau, J. M. (1990). Disinhibition as a basic process in the expression of striatal functions. Trends in neurosciences, 13(7), 277-280.

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

Ciardo, F., Ghiglino, D., Roselli, C., & Wykowska, A. (2020, November). The Effect of Individual Differences and Repetitive Interactions on Explicit and Implicit Attitudes Towards Robots. In International Conference on Social Robotics (pp. 466-477). Springer, Cham.

Cozolino, L. (2014). The neuroscience of human relationships: Attachment and the developing social brain. WW Norton & Company.

Csibra, G., & Gergely, G. (1998). The teleological origins of mentalistic action explanations: A developmental hypothesis. Developmental Science, 1(2), 255-259.

Dautenhahn, K. (2007). Methodology & themes of human-robot interaction: A growing research field. International Journal of Advanced Robotic Systems, 4(1), 15.

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human–robot interaction. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1480), 679–704. doi:10.1098/rstb.2006.2004

Dautenhahn, K., & Werry, I. (2004). Towards interactive robots in autism therapy: Background, motivation and challenges. Pragmatics & Cognition, 12(1), 1-35.

De Cesarei, A., & Loftus, G. R. (2011). Global and local vision in natural scene identification. Psychonomic Bulletin & Review, 18(5), 840–847. doi:10.3758/s13423-011-0133-6

Deniau, J. M., & Chevalier, G. (1985). Disinhibition as a basic process in the expression of striatal functions. II. The striato-nigral influence on thalamocortical cells of the ventromedial thalamic nucleus. Brain research, 334(2), 227-233.

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.

Denning, T., Matuszek, C., Koscher, K., Smith, J. R., & Kohno, T. (2009, September). A spotlight on security and privacy risks with future household robots: attacks and lessons. In Proceedings of the 11th international conference on Ubiquitous computing (pp. 105-114). ACM.

Desmurget, M., Rossetti, Y., Prablanc, C., Jeannerod, M., & Stelmach, G. E. (1995). Representation of hand position prior to movement and motor variability. Canadian journal of physiology and pharmacology, 73(2), 262-272.

Deubel, H, Schneider, W.X. (1996) "Saccade target selection and object recognition: Evidence for a common attentional mechanism", Vision research, 36(12):1827-37.

Diehl, J. J., Schmitt, L. M., Villano, M., & Crowell, C. R. (2012). The clinical use of robots for individuals with autism spectrum disorders: A critical review. Research in autism spectrum disorders, 6(1), 249-262.

Dittrich, W. H. (1999, March). Seeing biological motion-Is there a role for cognitive strategies?. In International Gesture Workshop (pp. 3-22). Springer, Berlin, Heidelberg.

Driver IV, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. Visual cognition, 6(5), 509-540.

Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience & Biobehavioral Reviews, 24(6), 581-604.

Erber, R., Wegner, D. M., & Therriault, N. (1996). On being cool and collected: Mood regulation in anticipation of social interaction. Journal of personality and social psychology, 70(4), 757.

Feil-Seifer, D., & Mataric, M. J. (2010). Dry your eyes: examining the roles of robots for childcare applications. Interaction Studies, 11(2), 208.

Fernández, M., Mollinedo-Gajate, I., & Peñagarikano, O. (2018). Neural circuits for social cognition: Implications for autism. Neuroscience, 370, 148-162.

Ficocelli, M., Terao, J., & Nejat, G. (2015). Promoting interactions between humans and robots using robotic emotional behavior. In Proceedings of IEEE transactions on cybernetics.

Finnoff, W., Hergert, F., & Zimmermann, H. G. (1993). Improving model selection by nonconvergent methods. Neural Networks, 6(6), 771–783. doi:10.1016/s0893-6080(05)80122-4

Fiore, S. M., Wiltshire, T. J., Lobato, E. J. C., Jentsch, F. G., Huang, W. H., & Axelrod, B. (2013). Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior. Frontiers in Psychology, 4. doi:10.3389/fpsyg.2013.00859

Fodor, J. A. (1975). The language of thought (Vol. 5). Harvard University Press.

Fodor, J. A. (1983). The Modularity of Mind; Cambridge, a Bradford Book.

Fonagy, P. (2018). Affect regulation, mentalization and the development of the self. Routledge.

Fox, R. (2006). Animal behaviours, post-human lives: Everyday negotiations of the animal–human divide in pet-keeping. Social & Cultural Geography, 7(4), 525-537.

François, D., Powell, S., & Dautenhahn, K. (2009). A long-term study of children with autism playing with a robotic pet: Taking inspirations from non-directive play therapy to encourage children's proactivity and initiative-taking. Interaction Studies, 10(3), 324-373.

Freedman, E. G., & Sparks, D. L. (2000). Coordination of the eyes and head: movement kinematics. Experimental brain research, 131(1), 22-32.

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. Psychonomic bulletin & review, 5(3), 490-495.

Frith, C. D., & Frith, U. (1999). Interacting minds--a biological basis. Science, 286(5445), 1692-1695.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. Neuron, 50(4), 531-534.

Frith, C. D., & Frith, U. (2008). Implicit and explicit processes in social cognition. Neuron, 60(3), 503-510.

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. Annual review of psychology, 63, 287-313.

Frith, U. (2003). Autism: Explaining the enigma. Blackwell Publishing.

Geisen, E., & Bergstrom, J. R. (2017). Usability testing for survey research. Morgan Kaufmann.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. Trends in cognitive sciences, 7(7), 287-292.

German, T. P., & Johnson, S. C. (2002). Function and the origins of the design stance. Journal of Cognition and Development, 3(3), 279-300.

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

Ghiglino, D., De Tommaso, D., Willemse, C., Marchesi, S., & Wykowska, A. (2020b). Can I get your (robot) attention? Human sensitivity to subtle hints of human-likeness in a humanoid robot's behavior. In Cogsci 2020

Ghiglino, D., Willemse, C., Tommaso, D. D., Bossi, F., & Wykowska, A. (2020a). 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.

Gielniak, M. J., Liu, C. K., & Thomaz, A. L. (2013). Generating human-like motion for robots. The International Journal of Robotics Research, 32(11), 1275-1301.

González, A., Ramírez, M. P., & Viadel, V. (2012). Attitudes of the Elderly Toward Information and Communications Technologies. Educational Gerontology, 38(9), 585–594.

Graf, B., Hans, M., & Schraft, R. D. (2004). Care-O-bot II—Development of a next generation robotic home assistant. Autonomous robots, 16(2), 193-205.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. science, 315(5812), 619-619.

Green, M. F., & Horan, W. P. (2010). Social cognition in schizophrenia. Current Directions in Psychological Science, 19(4), 243-248.

Griffin, R. & Dennett, D.C. (2008). What does the study of autism tell us about the craft of folk psychology? In T. Striano & V. Reid (Eds.) Social Cognition: Development, Neuroscience, and Autism (pp. 254-280). Wiley-Blackwell.

Hampton, K. N., Sessions, L. F., Her, E. J., & Rainie, L. (2009). Social isolation and new technology. Pew Internet & American Life Project, 4.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Human factors.

Harald Baayen, R., & Milin, P. (2010). Analyzing reaction times. International Journal of Psychological Research, 3(2), 12–28.

Hauser, M. D. (1996). The evolution of communication. The MIT Press.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. The American journal of psychology, 57(2), 243-259.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. The American journal of psychology, 57(2), 243-259.

Hinz, N.-A., Ciardo, F., & Wykowska, A. (2019). Individual Differences in Attitude Toward Robots Predict Behavior in Human-Robot Interaction. Lecture Notes in Computer Science, 64–73.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). Eye tracking: A comprehensive guide to methods and measures. OUP Oxford.

Huang, C.M., Andrist, S., Sauppé, A., & Mutlu, B. (2015). Using gaze patterns to predict task intent in collaboration. Frontiers in Psychology, 6.

Itier, R. J., & Batty, M. (2009). Neural bases of eye and gaze processing: the core of social cognition. Neuroscience & Biobehavioral Reviews, 33(6), 843-863.

Jacquette, D. (1991). The origins of Gegenstandstheorie: Immanent and transcendent intentional objects in Brentano, Twardowski, and Meinong.

Johansson, R. S., Westling, G., Bäckström, A., & Flanagan, J. R. (2001). Eye–Hand Coordination in Object Manipulation. The Journal of Neuroscience, 21(17), 6917–6932.

John, O. P., and Srivastava, S. (1999). "The Big Five trait taxonomy: History, measurement and theoretical perspectives". In Handbook of personality: Theory and research. New York: Guilford.

Kahn, P. H., Ishiguro, H., Friedman, B., & Kanda, T. (2006, September). What is a human?-toward psychological benchmarks in the field of human-robot interaction. In Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on (pp. 364-371). IEEE.

Kajopoulos, J., Wong, A. H. Y., Yuen, A. W. C., Dung, T. A., Kee, T. Y., & Wykowska, A. (2015, October). Robot-assisted training of joint attention skills in children diagnosed with autism. In International Conference on Social Robotics (pp. 296-305). Springer, Cham.

Kamide, H., Kawabe, K., Shigemi, S., & Arai, T. (2015). Anshin as a concept of subjective well-being between humans and robots in Japan. Advanced Robotics, 29(24), 1624-1636.

Kaplan, S. C., Levinson, C. A., Rodebaugh, T. L., Menatti, A., & Weeks, J. W. (2015). Social anxiety and the Big Five personality traits: The interactive relationship of trust and openness. Cognitive behaviour therapy.

Kapron-king, A., Kirby, S., & Woensdregt, M. (2020). Modelling cultural evolution of pragmatic communication when language co-develops with perspective-taking. In proceedings of the 13th conference on The Evolution of Language, 220-222.

Karmiloff-Smith, A., Klima, E., Bellugi, U., Grant, J., & Baron-Cohen, S. (1995). Is there a social module? Language, face processing, and theory of mind in individuals with Williams syndrome. Journal of cognitive Neuroscience, 7(2), 196-208.

Kelemen, D. (1999). Function, goals and intention: Children's teleological reasoning about objects. Trends in Cognitive Sciences, 3(12), 461-468.

Kelley, T. D. (2006). Developing a psychologically inspired cognitive architecture for robotic control: The Symbolic and Subsymbolic Robotic Intelligence Control System (SS-RICS). International Journal of Advanced Robotic Systems, 3(3), 32.

Khatib, O., Warren, J., De Sapio, V., & Sentis, L. (2004). Human-like motion from physiologically-based potential energies. In On advances in robot kinematics (pp. 145-154). Springer, Dordrecht.

Kobayashi, H., & Kohshima, S. (2001). Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye. Journal of human evolution, 40(5), 419-435.

Kompatsiari, K., Ciardo, F., Tikhanoff, V., Metta, G., Wykowska, A. "On the role of eye contact in gaze cueing", Scientific reports, 2018, 8(1):17842.

Korkmaz, B. (2011). Theory of mind and neurodevelopmental disorders of childhood. Pediatr Res, 69(5 Pt 2), 101R-8R.

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.

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.

Lee, D. H., & Anderson, A. K. (2017). Reading What the Mind Thinks From How the Eye Sees. Psychological Science, 28(4), 494–503.

Lee, J., & Lee, K. H. (2006). Precomputing avatar behavior from human motion data. Graphical Models, 68(2), 158-174.

Leiner, D. J. (2016). SoSci Survey. Available at: https://www.soscisurvey.de

Leiner, D. J. (2018). SoSci Survey (Version 2.5.00-i1142) [Computer software]. Available at http://www.soscisurvey.com

Leite, I., Martinho, C., & Paiva, A. (2013). Social Robots for Long-Term Interaction: A Survey. International Journal of Social Robotics, 5(2), 291–308.

Lillard, A. S., & Kavanaugh, R. D. (2014). The contribution of symbolic skills to the development of an explicit theory of mind. Child development, 85(4), 1535-1551.

Liu, C., Conn, K., Sarkar, N., & Stone, W. (2008). Online affect detection and robot behavior adaptation for intervention of children with autism. IEEE transactions on robotics, 24(4), 883-896.

Lopes, P. N., Salovey, P., Côté, S., Beers, M., & Petty, R. E. (2005). Emotion regulation abilities and the quality of social interaction. Emotion, 5(1), 113.

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the Intentional Stance towards humanoid robots? Frontiers in psychology.

Martini, M. C., Buzzell, G. A., & Wiese, E. (2015). Agent Appearance Modulates Mind Attribution and Social Attention in Human-Robot Interaction. Lecture Notes in Computer Science, 431–439.

Matan, A., & Carey, S. (2001). Developmental changes within the core of artifact concepts. Cognition, 78(1), 1-26.

Mathôt, S., Schreij, D., & Theeuwes, J. (2011). OpenSesame: An open-source, graphical experiment builder for the social sciences. Behavior Research Methods, 44(2), 314–324.

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.

Mayer-Hillebrand, F. (1951). Verzeichnis der Manuskripte Franz Brentanos, Typoscript, Harvard.

McConnell, S. R. (2002). Interventions to facilitate social interaction for young children with autism: Review of available research and recommendations for educational intervention and future research. Journal of autism and developmental disorders, 32(5), 351-372.

Mele, A. R., & William, H. (1992). Springs of Action. Understanding Intentional Behavior. Philosophical Books, 34(2), 116–120.

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., … Montesano, L. (2010). The iCub humanoid robot: An open-systems platform for research in cognitive development. Neural Networks, 23(8-9), 1125–1134.

Metta, G., Sandini, G., Vernon, D., Natale, L., & Nori, F. (2008). The iCub humanoid robot: an open platform for research in embodied cognition. In Proceedings of the 8th workshop on performance metrics for intelligent systems. ACM.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: eye movements during noun phrase production. Cognition, 66(2), B25–B33.

Miller, L. E., & Saygin, A. P. (2013). Individual differences in the perception of biological motion: links to social cognition and motor imagery. Cognition, 128(2), 140-148.

Millikan, R.G. (1984) Language, Thought and Other Biological Objects, Cambridge, Mass.: MIT Press.

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. Frontiers in Robotics and AI.

Mori, M. (1970). The uncanny valley. Energy.

Moriguchi, Y., Ohnishi, T., Lane, R. D., Maeda, M., Mori, T., Nemoto, K., ... & Komaki, G. (2006). Impaired self-awareness and theory of mind: an fMRI study of mentalizing in alexithymia. Neuroimage, 32(3), 1472-1482.

Müller, S. L., & Richert, A. (2018). The Big-Five Personality Dimensions and Attitudes to-wards Robots: A Cross-Sectional Study. In Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference.

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

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

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

Nummenmaa, L., Hyönä, J., & Calvo, M. G. (2006). Eye movement assessment of selective attentional capture by emotional pictures. Emotion, 6(2), 257–268.

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.

Paul, F., Elena, C., Daniele, D., Ali, P., Giorgio, M., & Lorenzo, N. (2014). A middle way for robotics middleware. JOURNAL OF SOFTWARE ENGINEERING IN ROBOTICS, 5(2), 42-49.

Pelachaud, C., & Bilvi, M. (2003, September). Modelling gaze behavior for conversational agents. In International Workshop on Intelligent Virtual Agents (pp. 93-100). Springer, Berlin, Heidelberg.

Perez-Osorio, J., Marchesi, S., Ghiglino, D., Ince, M., & Wykowska, A. (2019, November). More Than You Expect: Priors Influence on the Adoption of Intentional Stance Toward Humanoid Robots. In International Conference on Social Robotics (pp. C1-C1). Springer, Cham.

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

Peters, R. J., & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

Platek, S. M., Critton, S. R., Myers, T. E., & Gallup Jr, G. G. (2003). Contagious yawning: the role of self-awareness and mental state attribution. Cognitive Brain Research, 17(2), 223-227.

Pollack, M. E., Brown, L., Colbry, D., Orosz, C., Peintner, B., Ramakrishnan, S., ... & Thrun, S. (2002, August). Pearl: A mobile robotic assistant for the elderly. In AAAI workshop on automation as eldercare (Vol. 2002, pp. 85-91).

Pollard, N. S., Hodgins, J. K., Riley, M. J., & Atkeson, C. G. (2002). Adapting human motion for the control of a humanoid robot. In Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on (Vol. 2, pp. 1390-1397). IEEE.

Prolific, Oxford, UK (2015). Available at: https://www.prolific.co

Putnam, H. (1967). Psychological predicates. Art, mind, and religion, 1, 37-48.

PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eye tracking experiments. Behavior research methods, 46(4), 913-921.

Radach, R., Huestegge, L., & Reilly, R. (2008). The role of global top-down factors in local eye-movement control in reading. Psychological Research, 72(6), 675–688.

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. Psychological methods.

Robins, B., Dautenhahn, K., Te Boekhorst, R., & Billard, A. (2005). Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills?. Universal Access in the Information Society, 4(2), 105-120.

Roncone, A., Pattacini, U., Metta, G., Natale L.: A cartesian 6-DoF gaze controller for humanoid robots. In: Proceedings of Robotics: Science and Systems, Ann Arbor, MI, 18–22 June 2016.

Ruhland, K., Peters, C. E., Andrist, S., Badler, J. B., Badler, N. I., Gleicher, M., ... & McDonnell, R. (2015, September). A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. In Computer Graphics Forum (Vol. 34, No. 6, pp. 299-326).

Russell, B. (1905) 'On denoting', Mind, 14: 479-93; reprinted in A.P. Martinich (ed.) (1996) The Philosophy of Language, Oxford: Oxford University Press.

Sabbagh, M. A. (2004). Understanding orbitofrontal contributions to theory-of-mind reasoning: Implications for autism. Brain and cognition, 55(1), 209-219.

Sallinen-Kuparinen, A., McCroskey, J. C., & Richmond, V. P. (1991). Willingness to communicate, communication apprehension, introversion, and self-reported communication competence: Finnish and American comparisons. Communication Research Reports, 8(1), 55-64.

Salthouse, T. A., & Ellis, C. L. (1980). Determinants of eye-fixation duration. The American journal of psychology, 207-234.

Scassellati, B., Admoni, H., & Matarić, M. (2012). Robots for use in autism research. Annual review of biomedical engineering.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. Trends in cognitive sciences, 19(2), 65-72.

Schoemaker, M. M., & Kalverboer, A. F. (1994). Social and affective problems of children who are clumsy: How early do they begin?. Adapted physical activity quarterly, 11(2), 130-140.

Scholl, B. J., & Leslie, A. M. (1999). Modularity, development and 'theory of mind'. Mind & Language, 14(1), 131-153.

Schweinberger, S. R., Pohl, M., & Winkler, P. (2020). Autistic traits, personality, and evaluations of humanoid robots by young and older adults. Computers in Human Behavior.

Scott-Phillips, T. C. (2010). The evolution of communication: Humans may be exceptional. Interaction Studies, Social Behaviour and Communication in Biological and Artificial Systems, 11(1), 78–99.

Searle, J. (1979). What is an intentional state?. Mind, 88(349), 74-92.

Searle, J. (1980), "Minds, Brains and Programs", Behavioral and Brain Sciences, 3: 417–457

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. Trends in Cognitive Sciences, 10(2), 70–76.

Shanahan, M. (2006). A cognitive architecture that combines internal simulation with a global workspace. Consciousness and cognition, 15(2), 433-449.

Sharkey, A., & Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. Ethics and information technology, 14(1), 27-40.

Smith, K. A., Barstead, M. G., & Rubin, K. H. (2017). Neuroticism and conscientiousness as moderators of the relation between social withdrawal and internalizing problems in adolescence. Journal of youth and adolescence.

Sparrow, R. (2002). The march of the robot dogs. Ethics and information Technology, 4(4), 305-318.

Stergiou, N., & Decker, L. M. (2011). Human movement variability, nonlinear dynamics, and pathology: is there a connection?. Human movement science, 30(5), 869-888.

Syrdal, D. S., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2009). The negative attitudes towards robots scale and reactions to robot behavior in a live human-robot interaction study. Adaptive and emergent behaviour and complex systems.

Taylor, M., & Carlson, S. M. (1997). The relation between individual differences in fantasy and theory of mind. Child development, 68(3), 436-455.

Terada, K., Shamoto, T., & Ito, A. (2008, August). Human goal attribution toward behavior of artifacts. In Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on (pp. 160-165). IEEE.

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, 1962.

Thepsoonthorn, C., Ogawa, K. I., & Miyake, Y. (2018). The relationship between robot's nonverbal behaviour and human's likability based on human's personality. Scientific reports.

Troscianko, T., & Hinde, S. (2011). Presence While Watching Movies. i-Perception, 2(4), 216–216.

Turing, A. M. (1950). Mind. Mind, 59(236), 433-460.

Turkle, S., Taggart, W., Kidd, C. D., & Dasté, O. (2006). Relational artifacts with children and elders: the complexities of cybercompanionship. Connection Science, 18(4), 347-361.

Vaidya, A. R., Jin, C., & Fellows, L. K. (2014). Eye spy: The predictive value of fixation patterns in detecting subtle and extreme emotions from faces. Cognition, 133(2), 443–456.

Vernon, D., Metta, G., & Sandini, G. (2007, July). The icub cognitive architecture: Interactive development in a humanoid robot. In Development and Learning, 2007. ICDL 2007. IEEE 6th International Conference on (pp. 122-127). Ieee.

Veruggio, G. (2005). The birth of roboethics.

Vivanti, G., & Nuske, H. J. (2017). Autism, attachment, and social learning: Three challenges and a way forward. Behavioural brain research, 325, 251-259.

Wallach, W. (2010). Robot minds and human ethics: the need for a comprehensive model of moral decision making. Ethics and Information Technology, 12(3), 243-250.

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: using neuroscientific methods to make robots appear more social. Frontiers in psychology.

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. PloS one, 7(9), e45391.

Willemse, C., & Wykowska, A. (2019). In natural interaction with embodied robots, we prefer it when they follow our gaze: a gaze-contingent mobile eye-tracking study. Philosophical Transactions of the Royal Society.

Williams, D. (2010). Theory of own mind in autism: Evidence of a specific deficit in self-awareness?. Autism, 14(5), 474-494.

Wing, L., & Gould, J. (1979). Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification. Journal of autism and developmental disorders, 9(1), 11-29.

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), e94339.

Yates, K., & Le Couteur, A. (2016). Diagnosing autism/autism spectrum disorders. Paediatrics and Child Health, 26(12), 513-518.

Young, S. N., & Leyton, M. (2002). The role of serotonin in human mood and social interaction: insight from altered tryptophan levels. Pharmacology Biochemistry and Behavior, 71(4), 857-865.

Zheng, Z., Zhang, L., Bekele, E., Swanson, A., Crittendon, J. A., Warren, Z., & Sarkar, N. (2013, June). Impact of robot-mediated interaction system on joint attention skills for children with autism. In Rehabilitation Robotics (ICORR), 2013 IEEE International Conference on (pp. 1-8). IEEE.