UNIVERSITY OF GENOVA

PH.D. PROGRAM IN BIOENGINEERING AND ROBOTICS

Vicarious Sense of Agency in Human-Robot Interaction

by

Cecilia Roselli

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

December 2021

Prof. Agnieszka Wykowska	Supervisor
Dr. Francesca Ciardo	Co-Supervisor
Prof. Giorgio Cannata	Head of the Ph.D. Program

Dibris

Department of Informatics, Bioengineering, Robotics and System Engineering

Declaration

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

Cecilia Roselli

December 2021

Acknowledgments

I would like to thank all my colleagues, past and present, who I had the opportunity to spend the three years of my Ph.D. with. Each in their own way, they all represented a source of scientific enthusiasm, expertise, and never ending stimulating discussion. Among all, I would like to thank Davide, Serena, and Lorenzo: you always have been my safe space, and the ground to place my feet on every time I risked flying away.

However, the greatest gratitude goes to my supervisors Prof. Agnieszka Wykowska and Dr. Francesca Ciardo, for showing me the kind of scientist I would like to be: passionate, brilliant, and far-reaching.

A mia madre.

Nel giorno che sfugge,

il tempo reale sei tu.

Abstract

Sense of Agency (SoA) is the feeling of having control over one's actions and outcomes. In humans' daily life, SoA shapes whether, and how, people feel responsible for their actions, which has profound implications for the organization of human societies. Thus, SoA has received considerable attention in psychology and cognitive neuroscience, which tried to identify the cognitive mechanism underlying the emergence of the individual experience of agency. However, humans are inherently social animals, who are deeply immersed in social contexts with others. Thus, investigations of SoA cannot be limited to understanding the individual experience of agency, as SoA also affects the way people experience others' actions: this is how SoA becomes "vicarious". Humans can experience vicarious SoA over another human's actions and outcomes; however, the mechanisms underlying the emergence of vicarious SoA are still under debate. In this context, focusing on artificial agents may help shed light on the vicarious SoA phenomenon. Specifically, robots are an emerging category of artificial agents, designed to assist humans in a variety of tasks- from elderly care to rescue missions. The present Ph.D. thesis aimed at investigating whether, and under which conditions, robots elicit vicarious SoA in humans in the context of Human-Robot Interaction (HRI). Moreover, we aimed at assessing whether vicarious SoA may serve as an implicit measure of intentionality attribution towards robots. The link between vicarious SoA and intentionality attribution was based on the idea that, in some contexts, humans can perceive robots as intentional agents, and it may "boost" the "vicarious" control that they experience over robot's actions and outcomes- as well as it happens with other humans.

In three studies, we employed the Intentional Binding (IB) paradigm as a reliable measure of implicit SoA. Participants performed an IB task with different types of robots varying in their degree of anthropomorphic features and human-like shape (i.e., the Cozmo robot and the iCub robot). Specifically, our goal was to assess whether the emergence of vicarious SoA in humans was modulated by (1) the possibility to represent robot's actions using one's own motor schemes, (2) the attribution of intentionality towards robots, and (3) the human-like shape of the robot. Our results suggested that the interplay of these three factors modulates the emergence of vicarious SoA in HRI. In conclusion, the findings collected in the present thesis contribute to the field of research on the vicarious SoA phenomenon in HRI, providing useful hints to design robots well-tailored to humans' attitudes and needs.

Index

SECTION I- INTRODUCTION

1.1. Being an Agent: philosophical definitions and psychological implications of the Sense of Agency

How do I know myself? It is certainly a provocative question, as the "Self" is elusive and seems to escape introspection (De Vignemont & Fourneret, 2004). As Hume pointed out (1739), when humans look inside themselves they can never find the "Self", but only a collection of perceptions. Therefore, the question can be translated into how self-consciousness can be reduced to the consciousness of this collection of perceptions; in other words, how do I recognize my mental and bodily state as my own?

For instance, imagine that you are in a dark room, and you want to turn on the light. To do so, you move your arm to touch the switch, and after a short time, the room lights up. As your hand reached the switch, you know that the light turned on, as a consequence of your action. How do you know that? Furthermore, it may happen that the light does not turn on when you press the switch, or that you are too slow, and a friend turns on the light before you. In both cases, you still see the room lighting up, but you will not consider yourself as the one who caused it. In these cases, how can you know that the "authorship" of that action belongs to someone or something else? The ability to refer to oneself as the author of one's actions, and thus to "make something happen", is defined as *Sense of Agency* (SoA) (Haggard, 2017). As humans are agents intentionally moving in the world, our actions cannot be reduced to motor acts alone, as their consequences affect the surrounding environment. Humans tend to establish a temporal and causal connection between voluntary motor acts and external sensory events in the environment (Haggard, 2008), which are considered action outcomes or consequences. Thus, humans do not experience events like the change of luminosity in the room as passively happening to them, but they feel in charge of controlling the course of external events (Moore, 2016).

Section I- Introduction

In the literature, SoA received considerable attention in philosophy, given that it is profoundly intertwined with the philosophical debate about free will and attribution of responsibility (e.g., Gallagher, 2000; De Vignemont & Fourneret, 2004; Pacherie, 2008; Nichols, 2011; Frith, 2014). Most human societies, if not all, require that their members are held responsible for what they do. This, in turn, allows individuals' behavior to be legitimately managed through punishment or reward, for the benefit of the social group and the promotion of social cohesion. Legal systems have translated the concept of "behavioral management" into the notion of "legal responsibility", which assumes that healthy adult people have voluntary control over their actions and that they understand the relationship between their actions and the consequences of these actions (Christensen, Yoshie, Di Costa, & Haggard, 2016; Haggard, 2017). In other words, an adult person claiming to not have a conscious experience of what he/she was doing when committing a criminal act could not be held responsible for the consequences of this act, although the lack of agency is still difficult to be demonstrated objectively (Haggard, 2017). The assumption behind it is that it only makes sense to hold someone responsible for their actions if they are free in control of them. Indeed, free will is defined as the idea that human beings possess the power to make free choices and are hence able to bear responsibility for their actions (see Moore, 2016 for a more extensive discussion). In the attempt to identify specific cognitive processes allowing a person to relate the origin of an action to an agent (Georgieff & Jeannerod, 1998), SoA has become a phenomenon of interest for both psychology and cognitive neuroscience. From a neuropsychological point of view, there is evidence for specific neural correlates of SoA, which might reflect different aspects of agency and/or sub-processes related to agency (see Balconi, 2010 for a more general overview). In a clinical context, some psychiatric and neurological disorders causing abnormalities of self-awareness provided knowledge about the experience of agency. For example, schizophrenic

patients typically exhibit symptoms like acoustic-verbal hallucinations, through insertion or withdrawal, and delusion of alien control, which make them experience a loss of control over their actions, or the feeling to be controlled by external forces (Jeannerod, 2009). In a similar vein, also psychotic patients often declare that their actions do not belong to them, but rather that they are forced to act by some other agents (see Moore & Fletcher, 2012 for a review). Notably, some other disorders seem to involve a significant impairment of SoA, such as Obsessive-Compulsive Disorder (Gentsch, Schütz-Bosbach, Endrass, & Kathmann, 2012), or anosognosia for hemiplegia (Fotopoulou et al., 2008). In this framework, it appears crucial to investigate SoA in order to shed light on processes underlying the conscious experience over one's actions and their consequences. This should enable the definition of boundaries of individual responsibility that, as humans, we can apply to our everyday life.

1.2. Measures of Sense of Agency

Despite the increasing attention given to the topic of SoA in the past years (see Moore, 2016 for a review), an objective measure of SoA is difficult to achieve. Indeed, several methodological issues arise in the context of measuring SoA: first, many voluntary actions are "phenomenally thin" (Haggard, 2017, p.1). In other words, in real life, humans are quite unlikely to have a strong and stable experience of SoA, and usually navigate the world having a minimal experience of agency for voluntary actions that have become part of their daily routine, e.g. driving the car to get back home after work. Since the subjective awareness that accompanies these actions is not particularly vivid (e.g., I would feel minimally aware of my foot stepping on the clutch pedal), it is not an easy endeavor to develop measures able to capture this elusive experience, especially in a laboratory setting.

For many years, the most common approach to quantify SoA involved the use of explicit measures (Moore, 2016). For example, participants were asked to describe their agentic experience in the form of judgments or self-reports (e.g., Nahab et al., 2011; Preston & Newport, 2010). Sometimes these subjective reports have been related to physiological measures, such as the readiness potentials (RPs) calculated from the electroencephalography (EEG) or muscular activity (e.g., Haggard & Eimer, 1999; Haggard, Clark, & Kalogeras, 2002). Although explicit measures have significantly contributed to investigating factors potentially influencing the conscious experience of SoA, these measures seem to be vulnerable to some limitations. Indeed, explicit measures are self-biased, as shown by the consistent tendency to overestimate one's agency over external events across studies (e.g., Daprati et al., 1997; Tsakiris, Hesse, Boy, Haggard & Fink, 2007). The typical result is that people tend to misattribute to oneself events that are unrelated to one's actions (Wegner & Wheatley, 1999; Tsakiris, Haggard, Franck, Mainy, & Sirigu, 2005). This tendency seems to be stronger when the outcome of an action is positive, compared to when it is neutral or negative, suggesting that a powerful "self-serving" bias plays a role in distorting the subjective experience of SoA (Bandura, 1982). In this framework, implicit measures provided a valid alternative to quantify SoA, with no explicit requests about the agentic experience (Moore, 2016). This is especially valid if we consider that, in our everyday life, we generally feel in charge of what we are doing without reflecting upon it, i.e. without an explicit and conscious experience of SoA. So far, several implicit measures have been developed, usually assessing a behavioral or (electro-) physiological correlate of voluntary actions (Moore, 2016). One example is paradigms based on sensory attenuation, according to which the perceived intensity of a sensory outcome is reduced when it is caused by one's actions compared to externally triggered actions (Blakemore, Wolpert, & Frith, 1998; 1999; 2000). Within the context of the present thesis, however, the most

relevant implicit measure is the Intentional Binding (IB) paradigm, which is based on estimations of temporal duration (Haggard et al., 2002; see Moore and Obhi, 2012 for a review) either through the interval estimation procedure (Engbert, Wohlschläger, Thomas, & Haggard, 2007), or through reporting an occurrence of an event in time (Libet, Gleason, Wright, & Pearl, 1983). The interval estimation procedure requires participants to estimate the perceived length of the time interval between two types of events (usually a voluntary action and an auditory tone). The other procedure requires participants to estimate when a given event occurred by means of the Libet clock (Libet et al., 1983). Both methods have the advantage to be a reliable implicit measure of SoA, due to a lack of explicit awareness of the perceptual shift between a voluntary action and its subsequent outcome. At the same time, both methods suffer from some limitations. For instance, the interval estimation method can only make inferences about the overall IB effect; in contrast, the Libet clock method is able to disentangle between *Action Binding* and *Outcome Binding*, as the onset of action and outcome can be estimated separately according to the event of interest (Tanaka, Matsumoto, Hayashi, Takagi, & Kawabata, 2019). However, the use of the Libet clock method has several limitations as well. For instance, the instructions of whether to report the onset or the end of one's movement may influence participants' estimations, as well as the luminance of the clock hand and its size (e.g. Pockett & Miller, 2007). Additionally, tasks employing the Libet clock method are visually and cognitively demanding, as participants have to continuously follow the clock hand on the screen to make accurate judgments (Muth, Wirth, & Kunde, 2021). Moreover, when used in a shared social context with another agent, it may be that the social aspect may be less preserved during the experiment, as participants would focus their attention on the clock hand rotating rather than on the co-agent they are performing the task with.

However, the Libet clock method seems to be more robust and sensitive to capture implicit SoA compared to the interval estimation method (see Tanaka et al., 2019 for more information). Throughout all studies reported in this thesis, the Libet clock was used. Therefore, the next paragraph focuses on describing and reviewing evidence of SoA collected with paradigms based on the Libet clock method.

1.3. The Intentional Binding paradigm

In 2002, Haggard and colleagues conducted an innovative study to investigate the consciousness of action (Haggard et al., 2002). To this aim, they developed a novel paradigm, i.e. the so-called Intentional Binding (IB) paradigm, based on the Libet clock method (Libet et al., 1983). Participants' task was to observe a clock with conventional intervals (5, 10, 15 minutes, etc.), and a clock hand rotating. Then, they reported the time of occurrence of a given event (e.g., an action or a subsequent sensory event), by indicating where the clock hand was when the event of interest occurred. Haggard and colleagues (2002) designed four *Baseline* conditions, in which a single event of interest occurred, one for each separate block. In the (1) *Voluntary Action* condition, participants performed a voluntary index-finger keypress at the time of their choosing, and then they reported the position of the clock hand when the keypress occurred. In the (2) *TMS-Induced Action* condition, the Transcranial Magnetic Stimulation (TMS) induced involuntary twitches of the hand, and participants had to report the position of the clock hand when the twitch (i.e., an involuntary action) occurred. In the (3) *Sham-TMS* condition, the TMS stimulation produced an audible click, without any recordable muscular activity or abnormal perceptual experiences. Participants reported the position of the clock hand when they heard the click. Finally, in the (4) *Auditory Stimulus* condition, participants heard a tone occurring randomly during the rotation of the clock hand. Their task was to report the position of the clock hand when the tone occurred.

Haggard and colleagues (2002) also designed three *Operant* conditions, in which the auditory tone followed (1) voluntary actions, (2) TMS-induced movements, or (3) sham-TMS stimuli with a fixed action-effect interval of 250 ms. Participants' task was to report the position of the clock hand when either the first event (i.e., voluntary actions, TMS- generated motor responses, or sham TMS stimuli) or the subsequent tone occurred. Interestingly, results showed different patterns according to the critical event to judge (i.e., action or tone) and the nature of the causing action (see **Figure 1**).

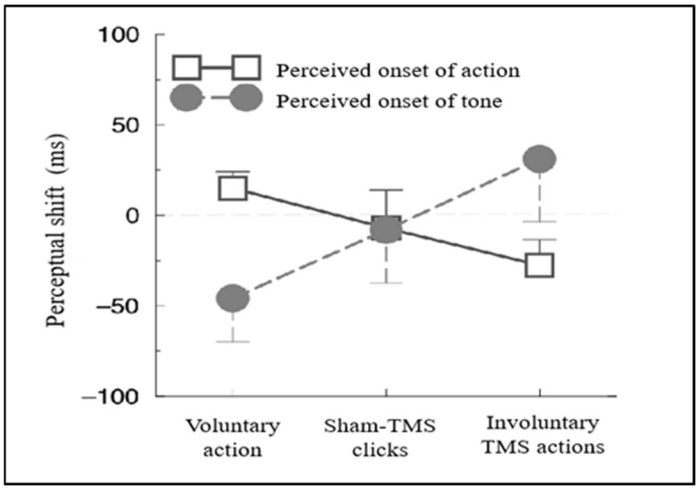

Figure 1. The classic pattern of the IB effect, according to the nature of the causing action and the critical event to judge (redrawn from Haggard et al., 2002). The onset of voluntary actions was perceived as shifted later on the timeline (white squares), whereas the onset of tone was perceived as shifted earlier on the timeline (grey dots). This was not the case for involuntary actions induced by TMS, which showed the opposite effects. Notably, the onset of the sham-TMS clicks and the tone following such clicks were not shifted in time.

These findings showed that, when humans perform self-generated voluntary actions causing a sensory consequence (e.g., a tone), these two events are perceptually linked to each other; this perceptual shift is called the *IB effect* (Haggard et al., 2002). It leads to the question of whether the IB paradigm can be considered as a reliable measure to investigate implicit SoA. Thus, in the next paragraph, I will briefly describe current models underlying cognitive mechanisms of SoA, and their link with the IB paradigm.

1.4. How does the Sense of Agency come about? Models of Agency

Several models have been proposed to explain cognitive mechanisms underlying SoA. They can be divided into two main categories, according to the most relevant source of information for the experience of agency.

Predictive models. One category comprises the so-called *predictive models*, which emphasize the central contribution of the motor system for the experience of agency: these are the *Comparator Model* (e.g., Wolpert, Ghahramani, & Jordan, 1995; Frith, Blakemore, & Wolpert, 2000; Frith, 2005), and the *Ideomotor Theory* (e.g., Prinz, 1997; Hommel, Müsseler, Aschersleben, & Prinz, 2001; Hommel, Alonso, & Fuentes, 2003; Massen & Prinz, 2009). The *Comparator Model* mainly focuses on the role of the prediction towards the outcome of an action, whereas the *Ideomotor Theory* mainly focuses on the role of the internal representation of the action-outcome link. Both models share the assumption that the predictive processes related to the sensory consequences of an action are the central mechanisms for the experience of agency.

Thus, SoA would arise only *retrospectively*, i.e., after the action has been performed.

The Comparator Model. The key assumption of the *Comparator Model* is that the experience of agency for a given action arises from the internal motor representation of that action, which is used to predict its sensory consequences (e.g., Kawato, 1999; Blakemore, Wolpert, & Frith, 2002). Specifically, an internal prediction of the sensory consequences of an action is generated based on an *efference copy* of the motor commands. These predicted sensory consequences are then compared with the actual sensory consequences after the action has been initiated. If the actual sensory consequences match the predicted ones, then the agent will perceive the sensory consequences as caused by his/her action. As a result, SoA will arise. Conversely, in the case of a

mismatch, the agent will perceive the action as externally caused, and SoA will be reduced or absent (see **Figure 2**).

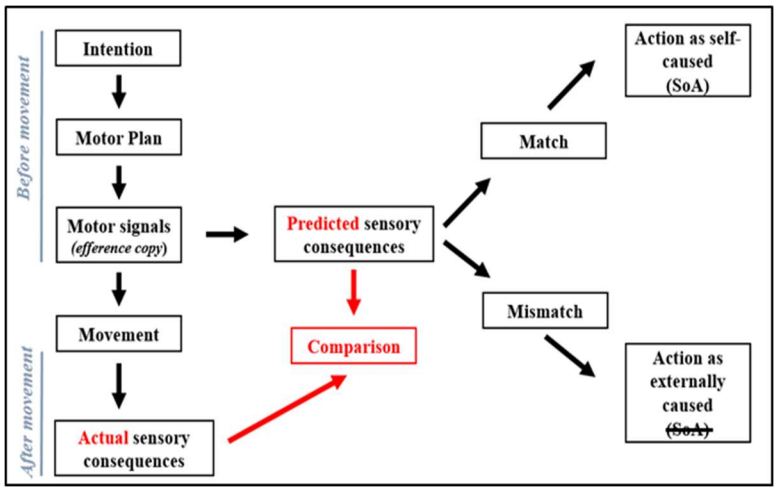

Figure 2. Schematic representation of the *Comparator Model*, redrawn from David, Newen, & Vogeley (2008). It is important to note that it is not a purely predictive account, as SoA here requires sensory feedback (and thus also a postdictive component) to arise. However, the main emphasis of the *Comparator Model* remains on the predictive sensorimotor processes, and not the postdictive inferences, for the experience of agency.

Many experimental investigations supported the *Comparator Model*, in terms of both behavioral and neurophysiological evidence (e.g., Daprati et al., 1997; Fink et al., 1999; Sirigu, Daprati, Pradat-Diehl, Franck, & Jeannerod, 1999; Slachevsky et al., 2001; Farrer & Frith, 2002; Farrer et al., 2003; MacDonald & Paus, 2003). This model also received support when investigating abnormalities of action awareness in clinical contexts (e.g., Frith & Done, 1989; Franck et al., 2001; Haggard, Martin, Taylor-Clarke, Jeannerod, & Franck, 2003), to explain, for example, delusion of control mechanisms in schizophrenic patients (e.g., Frith & Done, 1989; Franck et al., 2001; Blakemore et al., 2002; cf. Gallagher, 2004).

The Ideomotor Theory. The key assumption of the Ideomotor Theory is that intentional behaviors are based on the bidirectional association between actions and the subsequent sensory consequences (e.g., Prinz, 1997; Hommel et al., 2001; Massen & Prinz, 2009; Haering & Kiesel, 2014). Thus, actions could be represented through their perceivable sensory outcomes, and could

be triggered by anticipating or thinking about these outcomes (Prinz, 1997). SoA is inferred by noticing "perceptual conjunctions" (i.e., similarities) between the "idea" of the action and the actual outcome of the action. In other words, the "idea" of the action triggers an action whose outcome is similar to this idea. If the actual outcome matches the predicted one, according to the action that has been previously selected based on the "idea" of the action, then SoA arises (Chambon & Haggard, 2013). Conversely, in case of mismatch- i.e., when the action does not produce the expected outcome-, SoA is reduced (see **Figure 3**).

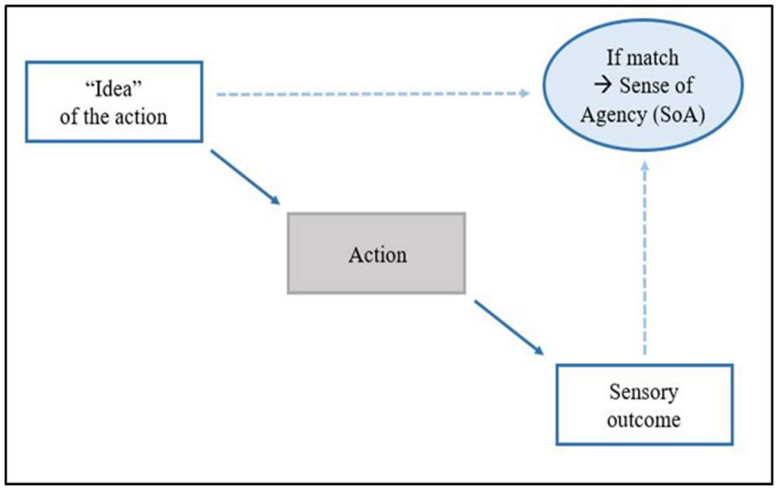

Figure 3. Schematic representation of the *Ideomotor Theory*, redrawn from Chambon & Haggard (2013). When an action repeatedly results in the same outcome, people would form an association between the action and the subsequent outcome. Therefore, when an agent intends to produce that outcome, the anticipation of the acquired effect automatically triggers the appropriate motor behavior, i.e., the action that usually produces that outcome.

A recent study investigating SoA at both implicit and explicit levels demonstrated the similarity between the *Comparator Model* and the *Ideomotor Theory* to explain the experience of agency (Barlas & Kopp, 2018). Despite the common emphasis on the general processing of actions, the comparator approach translates into operations the cognitive architecture suggested by the ideomotor approach; whereas the ideomotor approach provides a cognitive architecture that focuses on the processes that the comparator approach explicitly targets (for an extensive discussion, see Hommel et al., 2001; Chambon & Haggard, 2012; Chambon & Haggard, 2013;

Hommel, 2015). In other words, the *Comparator Model* can be considered as a translation of the *Ideomotor Theory* into processing terms, which makes these approaches two complementary views explaining the emergence of SoA in "motor" terms (e.g., Hommel, 2015).

Postdictive models. The second category of SoA models comprises the so-called *postdictive models*, which emphasize not only the contribution of the motor system for the experience of agency, but also the role of information external to the motor systems, in establishing the links between actions and their outcomes. These are the *Theory of Apparent Mental Causation* (Wegner & Wheatley 1999; Wegner, 2002), and the *Cue Integration Theory* (Moore, Wegner & Haggard, 2009; Moore & Fletcher, 2012; Synofzik, Vosgerau & Voss, 2013). The *Theory of Apparent Mental Causation* proposes that the experience of agency minimally relies on motor prediction, and mostly derives from information external to the motor system such as intentions and prior thoughts. The *Cue Integration Theory* tries to overcome this apparent dichotomy between internal and external information, proposing that SoA arises from a multilevel integration of both. The main assumption of *postdictive models* is that the experience of agency derives from a *post-hoc* inference occurring both before and after the action has been performed. Thus, SoA would be the result of a postdictive inference rather than the result of direct access to one's cognitive motor preparation processes preceding one's actions.

The Theory of Apparent Mental Causation. According to this theory, two causal pathways determine the experience of SoA (Wegner, 2002). One is the causal pathway that is responsible for the action, which corresponds to the working of the motor control system; the other is the one responsible for the associated thoughts about the actions, i.e., the intentions. There are some events we are conscious of, i.e., the intention to act and the act itself; SoA would be determined by the relationship between the intention and the actions (Wegner, 2002). Specifically, the experience of

agency would arise when three principles are met: (1) *priority*, (2) *consistency*, and (3) *exclusivity*. In more detail, (1) *priority* means that the agent has to have prior thoughts or plans about the action that he/she/they is going to perform, (2) *consistency* relates to the idea that the occurred action has to match with the action that was planned, and (3), *exclusivity* refers to the agent's actions being the exclusive cause of the subsequent outcome. Therefore, the simple co-occurrence of outcomes coherent with the agent's intentions would be sufficient for SoA to emerge, thanks to the postdictive inference (see **Figure 4**).

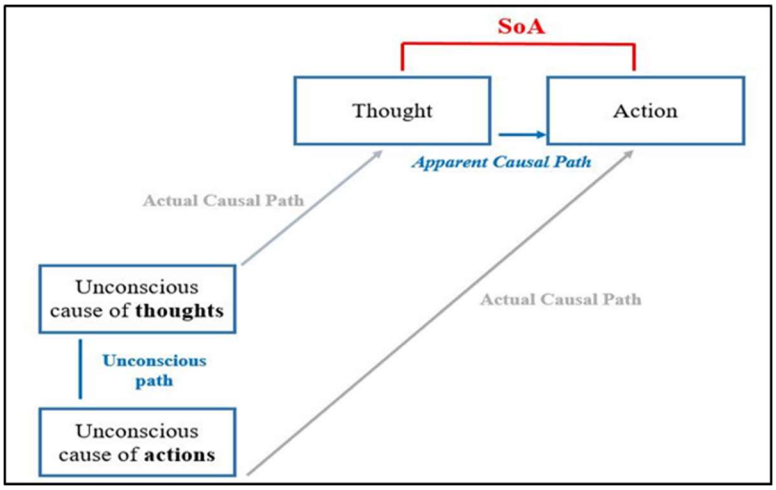

Figure 4. Schematic representation of the *Theory of Apparent Mental Causation*, redrawn from Wegner (2002). As shown by the picture, the experience of agency only partially relies on the working on motor control system (unconscious cause of action). Rather, it mostly relies on the inferential product of the integration of both motor and cognitive priors (unconscious cause of thoughts).

Previous evidence supported the *Theory of Apparent Mental Causation*. For example, Wegner and Wheatley (1999) conducted a study in which participants reported an illusory SoA for actions they did not perform; indeed, they were primed with thoughts related to a movement made by the confederate, which made them rate actions as self-caused. These findings would suggest that people have limited access to information related to their actions, and thus they rely on inferential processes to make sense of them. Consequently, information that is external to the motor system would be the key to the emergence of SoA (Wegner & Wheatley, 1999).

***The Cue Integration Theory*.** This theory proposes that the brain integrates information from multiple internal and external cues within a Bayesian model (Moore et al., 2009; Moore & Fletcher, 2012). Each cue has a relative influence, depending on its availability and reliability. Therefore, the experience of agency is generated when the highest weight is attributed to the cue that appears to be the most reliable in a given context. Thus, the interplay of interacting cues can result in two levels of the agency. The first is a basic, perceptual level (i.e., the Feeling of Agency, *FoA*), at which the agent can only determine whether an action is self- or not self-generated. The second is an explicit, high-level of agency, (i.e., the Judgment of Agency, *JoA*), at which it is possible to detect the actual agent of an action based on contextual and environmental information, i.e., one's background beliefs about who (or what) is the cause of the action or information about the environment (Synofzik, Vosgerau, & Newen, 2008) (see **Figure 5**).

Moore and Haggard (2008) showed evidence in support of this theory. The authors highlighted the role of both internal sensorimotor processes and inferential "sense-making" processes that occur after the action (Moore & Haggard, 2008). This theoretical account might also elucidate the processes underpinning SoA in disease; for example in schizophrenic patients, who rely more on visual feedback than on sensorimotor processes, which are unreliable and compromised (Synofzik, Thier, Leube, Schlotterbeck, & Lindner, 2010; see also Voss et al., 2010).

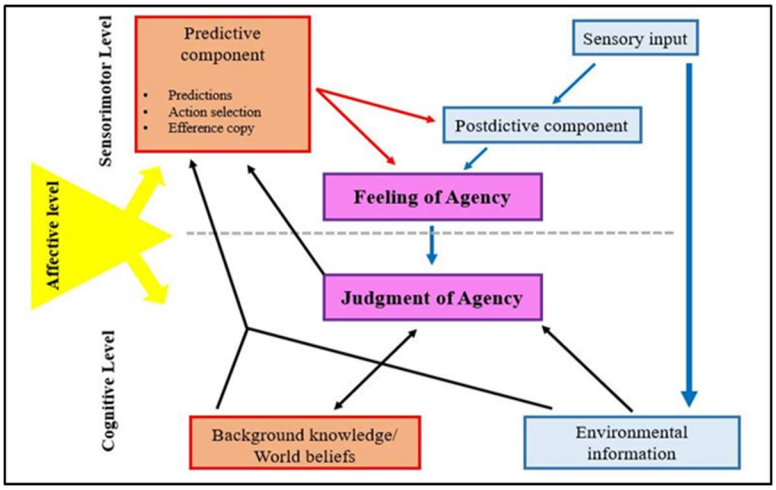

Figure 5. Schematic representation of the *Cue Integration Theory*, redrawn from Synofzik et al. (2008). As shown by the picture, SoA is the result of the interplay of various sources of information, which gains a different weight depending on a given context. For example, when internal predictions generated by the motor system are not precise, or weak, they receive a lower weight, with the most reliable cue being more high-level contextual or environmental information; or vice versa.

1.5. Intentional Binding and the Sense of Agency

To understand whether the IB effect can be used as a reliable measure of implicit SoA, we need to consider the models explained in the previous paragraph and see whether they could contribute to explaining the IB effect.

Previous evidence demonstrated that the IB effect could be explained in terms of *predictive mechanisms*, whether they are (1) the motor prediction emphasized by the *Comparator Model* (e.g., Wolpert et al., 1995; Frith, 2005), or the (2) representation of the action-outcome link emphasized by the *Ideomotor Theory* (e.g., Prinz, 1997; Hommel et al., 2001; Massen & Prinz, 2009). Regarding (1) the role of motor prediction, in their IB study, Haggard and Clark (2003) asked participants to perform either voluntary actions, i.e., keypresses, or involuntary TMS-induced actions. These kinds of movements were identical, apart from the intention to act that was lacking in movements triggered by the TMS stimulation. Results showed that the IB effect emerged only for voluntary actions, i.e., when participants had the intention to produce the tone.

If the intention was interrupted by an imposed involuntary movement, followed by an identical auditory tone, the IB effect did not occur (Haggard & Clark, 2003). On the one hand, these findings showed that the intention to act is a major factor affecting the IB effect. It was confirmed by further evidence showing that the IB effect is enhanced when actions are intentional compared to when they are not (e.g., Buehner, 2015). On the other hand, these results showed that the IB effect requires a specific match between intentions, actions, and outcomes to arise. Regarding (2) the representation of the action-outcome link, as postulated by the *Ideomotor Theory* (e.g., Prinz, 1997; Hommel et al., 2001), Barlas and Kopp (2018) investigated the role of action-tone congruency for the emergence of the IB effect. Congruency was manipulated varying the choice-level of actions from 1 to 4. Specifically, participants performed either an instructed or a freely selected action among four alternatives (i.e., right, left, up, and down keypresses). Visual outcome comprised as many alternatives (i.e., right, left, up, and down-pointing arrows). Each action could produce either a congruent or an incongruent outcome, depending on the matching between the direction of the keypress and the direction of the arrow outcome. As results showed that the IB effect was stronger in trials when there was congruency between the voluntary action and the subsequent outcome, the authors argued that the action-outcome association plays a fundamental role for the IB effect to occur (Barlas & Kopp, 2018).

However, the IB effect could be explained also in terms of *postdictive mechanisms*, i.e., based on information external to the motor system, as suggested by the *Theory of Apparent Mental Causation* (e.g., Wegner, 2002). For example, by varying the probability with which a keypress produces an auditory outcome, Moore and Haggard (2008) demonstrated that, even if the sensorimotor prediction is weak (i.e., the probability that the keypress produces a tone is low), the IB effect still occurs when the actions caused a tone. Moore and colleagues' study (2009) further

demonstrated the dual contribution of both internal motor signals and external, situational cues for the emergence of the IB effect, in line with the *Cue Integration Theory* (e.g., Moore et al., 2009; Synofzik et al., 2013). They used primes to modulate implicit SoA for voluntary and involuntary actions, by modifying the content of participants' conscious thoughts before the movement. The authors found a twofold result. First, the IB effect was stronger when primes were congruent with the outcome. Second, the IB effect was different according to the type of movement (Moore et al., 2009). When the movement was involuntary- and participants could not rely on internal motor commands-, external cues (i.e., primes) had a greater effect on the perception of the action-outcome interval, resulting in a diminished IB effect. Conversely, when the movement was voluntary- and internal motor commands were present-, primes played a reduced role, resulting in a stronger IB effect. These findings suggested that the IB effect (and, by extension, the experience of agency) is based on a combination of internal motor signals and external sensory cues, whose relative influence is weighted according to the context.

The evidence presented so far highlighted two important aspects. First, the IB effect can be considered as a reliable measure of implicit SoA, resulting from a weighted combination of both internal and external information that determines the experience of agency. Second, the IB effect is not only a mere temporal conjunction between an action and an outcome, but a causal action-outcome link that is strengthened when the causing action is intentional.

The "neurobiological" level of Sense of Agency

Despite remarkable, evidence presented in the previous paragraph explored the link between the *Intentional Binding* paradigm (e.g., Haggard et al., 2002) and SoA only from a behavioral perspective. However, it would be worthwhile to mention that other evidence supported the

existence of this link also at the neurobiological level. In 2010, Moore and colleagues conducted a study to investigate the neural substrates involved in SoA by measuring the effects of local inhibitions of brain activity on the IB effect (Moore, Ruge, Wenke, Rothwell, & Haggard, 2010). Specifically, the authors used theta-burst TMS to inhibit brain activity in the sensorimotor hand area (SMHA), which is primarily concerned with motor execution and sensorimotor feedback (e.g., Weiller et al., 1996), and in the pre-supplementary motor area, (pre-SMA), which is involved in more cognitive aspects of internal motor generation (e.g., Picard & Strick, 2001) as well as in the conscious "urge to act" (e.g., Fried et al., 1991). Both target regions were neural substrates if SoA, but at the same time, they were distinguishable in terms of function and time of contribution, The authors employed the Libet-clock method (Libet et al., 1983), in which the voluntary action was followed by a somatosensory stimulus (i.e., a mild shock to the right little finger). Results showed that only pre-SMA inhibition led to a significant reduction of the IB effect overall, and particularly in the IB effect for sensory outcomes towards the actions. The same result did not emerge from inhibition of SMHA (Moore et al., 2010). A subsequent study by Kühn and colleagues (2013) employed an interval estimation paradigm to quantify implicit SoA, in combination with functional Magnetic Resonance Imaging (fMRI) to overcome the fact that TMS did not explore the effect of stimulating different brain regions within the SMA complex (Kühn, Brass, & Haggard, 2013). This study comprised an *Active Condition* in which participants performed a voluntary right index-finger keypress, followed by an auditory tone, and a *Passive Condition* in which participants' action was involuntary (i.e., the experimenter pushed down participants' index finger to perform the action). At the behavioral level, results replicated the classical IB effect, with a perceived shorter interval between voluntary actions and subsequent tones compared to physically comparable passive movements. At the neural level, a cluster in the

left SMA emerged, extending to the dorsal pre-motor cortex, whose activity was significantly enhanced when participants estimated the time interval between a voluntary action and its tone outcome compared to when the action was passively executed (Kühn et al., 2013).

In sum, these findings showed that supplementary motor complex might contribute to the implicit experience of agency, which is associated with distortions of time perception. Notably, this relationship seems to be confirmed also at the explicit level, by other evidence showing a positive correlation between pre-SMA activation and explicit judgments of SoA (e.g., Miele, Wager, Mitchell, & Metcalfe, 2011).

1.6. Feeling control in social contexts: how Sense of Agency becomes "vicarious" Sense of Agency

So far, I focused on the experience of agency at the individual level, referring to various theories that, across years, tried to explain what characterizes the feeling of "authorship" over one's actions and outcome. It was meant to give a general overview of the SoA phenomenon, which however is not limited to the sole individual experience. Indeed, humans are inherently social; already Aristotle stated that "*man is by nature a social animal; an individual who is unsocial naturally and not accidentally is either beneath our notice or more than human*" (Aristotle, Politika, ca. 328 BC). In the context of SoA, it would mean that, in addition to the sense of self-agency over one's actions and outcomes, (i.e., the sense that "I did it"), people can also experience a sense of *Joint Agency* (i.e., the sense that "We did it") (e.g., Pacherie, 2012). However, when an action is performed by another agent, SoA may still occur at the individual level (Wegner, Sparrow, & Winerman, 2004). This is how individual SoA becomes "*vicarious*" SoA, i.e., the feeling of authorship for others' actions and outcomes (Wegner et al., 2004). Nielsen was the first to document an experience of vicarious agency (Nielsen, 1963). He asked participants to draw a line

on a piece of paper. Participants' hand was inserted in a viewing box, so that they could see their hand when drawing the line, or, unbeknownst to them, the experimenter's hand, which appeared in the position where they would expect their hand to be. When the experimenter's hand deviated from the line, participants did not spot this deception; instead, they corrected the tracing by aligning their hand with the experimenter's one. This result showed that participants experienced a "vicarious" control over actions executed by another agent, as demonstrated by the fact that they adjusted their movements to the false visual feedback without being aware of the adjustment, or being aware that the hand that was deviating was not theirs.

In this framework, also abnormalities of SoA may contribute to shedding light on mechanisms underlying vicarious SoA. For example, some schizophrenic patients who suffer from hallucinations experience hearing voices, apparently because they have thoughts that do not follow consistently their prior thinking. As a result, these inconsistent thoughts are difficult to attribute to oneself, being instead ascribed to imaginary agents and often causing authorship confusion (Hoffman, 1986). Some insight also came from a study of patients with apraxia, which disrupts the ability to produce voluntary movements (Sirigu et al., 1999). Patients were asked to make finger movements, e.g., crossing the middle over the index finger, and to discriminate whether, in a series of videos, what they observed were their fingers or those of an experimenter. When the observed movement matched the instructed movement, they often claimed that the movement was their own- although their motor disturbance prevented them from moving their fingers as in the observed video. Thus, the authors concluded that thinking of the correct action, and observing it, yielded an illusion of mental causation even when the direct sensation of disturbed movement contradicted this experience, causing an "illusory" feeling of authorship over actions that were not performed by them (Sirigu et al., 1999). These findings helped to show that vicarious SoA is a

pervasive experience of humans' daily life, which could have implications not only for the experience of conscious will and perception of one's agency, but also for the way people experience others' actions more generally (Wegner et al., 2004). Therefore, in the next paragraphs, I will describe the vicarious SoA phenomenon when people share a social context with various types of agents: other humans, or artificial systems such as computers and robots.

1.7. Vicarious Sense of Agency over other humans

Several studies investigated whether, and how, people can experience SoA over another human's actions and outcomes. Indeed, some experience of agency may persist over someone else's actions, even if no common goal is shared (Sahaï, Pacherie, Grynszpan, & Berberian, 2017), or people know that the action has been performed by another agent (Wegner et al., 2004). In Wohlschläger and colleagues' study (2003), the authors used the Libet clock methodology to understand whether the IB effect for actions arises similarly in self- and other-generated actions (Wohlschläger, Haggard, Gesierich, & Prinz, 2003). The study comprised two conditions involving humans. In the *Self-Action* condition, participants were placed in front of a lever that they could press at the time of their choosing. In the *Other-Action* condition, participants observed another agent (i.e., the experimenter) pressing the lever. In both conditions, participants had to estimate the onset time of the lever press. Results showed that the IB effect was indistinguishable between self- and other-generated actions, so that the perceived onset time of the action in the *Self-Action* condition was perceived in a similar way as in the *Other-Action* condition. Therefore, participants always experienced SoA, regardless of the author of the action (Wohlschläger et al., 2003). The authors argued that the other-generated action was simulated in the observer's predictive system, and thus vicarious SoA occurred because of the match between the predicted state- derived from the *efference copy*- and the actual state- derived from the observation of the real movement

(Wohlschläger et al., 2003). The idea of common processing between self- and other-generated actions was also supported by an ERP study investigating the neural processes involved in the implicit SoA over both one's and other's actions (Poonian, McFadyen, Ogden, & Cunnington, 2015). Using an interval estimation paradigm as a measure of the IB effect, participants completed three separate conditions. In the *Self* condition, participants performed a keypress at the time of their choosing; in the *Other* condition, a video was presented of another person performing a keypress. In both conditions, the keypress triggered a tone after a certain delay. In the *Control* condition, no keypress was required, but two tones were presented one after the other, each presentation separated by a certain delay. For all conditions, the participant's task was to estimate the time interval between the first event (i.e., the action, or the tone in the *Control* condition) and the second event (i.e., the subsequent tone). In addition, the authors used electroencephalography (EEG) to record participants' brain activity. Specifically, they aimed to investigate the N1 component, i.e., a component with a negative amplitude arising 100 ms after the onset of an auditory stimulus (see Näätänen & Picton, 1987, for an overview), and whose suppression has been considered as an indicator of implicit SoA (e.g., Gentsch & Schütz-Bosbach, 2011; Gentsch et al., 2012). Results showed that the IB effect occurred both in the *Self* and the *Other* condition, as participants underestimated the time interval between the observed action and the subsequent tone. Conversely, the time interval between the two tones in the *Control* condition was overestimated (Poonian et al., 2015). Furthermore, EEG results showed an N1 suppression for both self- and other-generated tone, with no differences in suppression between the two. In the *Control* condition, N1 was enhanced during the perception of the second tone compared to the one, because the sensory outcome had been already heard once. Consequently, Poonian and colleagues argued that sensory outcomes of self- and observed actions are processed in a similar way (Poonian et al.,

2015). Specifically, they suggested that top-down processes (and, in particular, predictions about the consequences of actions within an internal comparator model) would act upon the neural responses to sensory outcomes in the same way for outcomes caused by both self-and other-generated actions (Poonian et al., 2015). At the neurobiological level, evidence of an action/observation matching system were first found in the premotor area of the monkeys, namely the F55 area (e.g., Di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Rizzolatti, Fogassi, & Gallese, 2001; Umiltà et al., 2001). Specifically, these studies showed that the neurons of the F5 area, i.e., the *mirror neurons*, activated both when people executed a goal-directed, voluntary action and when they observed the very same action performed by another individual (see Fogassi et al, 2005 for more information). Later, an analogous system has been discovered in humans, and called *mirror system* (Rizzolatti & Craighero, 2004). The *mirror system* has been demonstrated to support our understanding of the low-level motor intentions of others (Rizzolatti & Sinigaglia, 2010). In other words, the mirror system would simulate the motor commands allowing the simulation content to be used to predict the consequences of the action, enhancing action understanding (Pacherie & Dokic, 2006). Notably, there is evidence that these predictive mechanisms are not solely involved in action generation and understanding, but also in the emergence of SoA (see Sahaï et al. 2017 for an overview). Therefore, *mirror neurons* may represent the neural substrate of vicarious SoA that would allow understanding actions of others, based on shared predictive mechanisms between self- and other-generated actions.

Results supporting the idea of common processing between self- and other-generated actions were also found by Wegner and colleagues (2004), using an explicit judgment methodology. In Experiment 1, participants were paired with confederates. Both participants and confederates wore headphones. Participants watched themselves in a mirror while the confederate, hidden from view,

extended hands forward on each side, appearing in the position of participants' arms both from the participants' normal viewpoint and in the participants' mirror reflection. Participants were asked to watch the mirror and not to move their arms. Through the headphones, confederates were told that they would hear a sequence of instructions of what movements to make. Conversely, participants were told that they might or might not hear anything through their headphones, and whatever they heard might or might not relate to the movement of the confederate. In the *Preview* condition, participants heard instructions as the hands of the confederate followed them, whereas in the *No-Preview* condition participants heard nothing through the headphones. Afterward, participants' task was to rate their feeling of "vicarious" control over the confederate's actions. Results showed that, when participants were primed by headphones instructions about the confederate's hand (*Preview* condition), they reported a higher degree of control over the observed actions compared to when no preview of the confederate's movement was given through the headphones (*No-Preview* condition). In Experiment 2, the authors added a new condition (i.e., *Inconsistent Preview* condition), so that participants could also hear instructions timed to coincide with the confederate's instructions, but not matching the movement that the confederate performed. Participants' task was to rate their feeling of control over the confederate's movement. In a similar vein as in Experiment 1, participants reported a higher degree of control when they were primed with instructions compared to where they did not. However, consistent instructions prompted reports of higher control, relative to inconsistent instructions, whereas no differences emerged between inconsistent instructions and no instructions at all (Wegner et al., 2004). Wegner and colleagues (2004) interpreted these results in the light of the *Theory of Apparent Mental Causation* (Wegner & Wheatley, 1999; Wegner, 2002). Indeed, the results of Experiment 1 showed that priming thoughts about the upcoming action fostered the vicarious SoA for that action,

as shown by a higher degree of control reported by participants when they could hear instructions previewing the confederate's movement compared to when they heard no instructions (Experiment 1). Furthermore, the results of Experiment 2 demonstrated the validity of the consistency principle described by the *Theory of Apparent Mental Causation* (Wegner & Wheatley, 1999; for a description, see paragraph 1.4, p. 12). In other words, thoughts consistent with the following action, as in the case of consistent instructions given to participants, yielded a stronger feeling of control compared to thoughts inconsistent with actions. It would suggest that is not sufficient to be primed with any instructions, rather, compatibility between thoughts and following actions is needed to experience vicarious SoA over another agent's actions. It was also demonstrated by the fact that inconsistent instructions led to the same reduced level of vicarious SoA produced by no instructions at all (Wegner et al., 2004).

The evidence presented so far demonstrated that humans experience vicarious SoA over another human's actions and outcomes. However, the mechanisms underlying the feeling of "vicarious control" are still under debate. Indeed, when interacting or observing another human agent acting, it may be difficult to disentangle between the role of the intentionality attributed to the other human and the role of predictive sensorimotor processes for the emergence of vicarious SoA. In other words, it may be hard to disentangle whether we feel in control over someone else's actions because we recognize the other as a human able to act intentionally, as well as we do, or because we can form a representation of the other human's actions using our motor system. In this context, artificial systems, such as computers and robots, may contribute to shedding light on the vicarious SoA phenomenon, as these factors can be orthogonally manipulated (or measured) in the case of artificial agents.

1.8. Vicarious Sense of Agency over artificial systems

Vicarious SoA over artificial agents' actions has been long recognized as a key factor in how people experience interactions with technology (Limerick, Coyle, & Moore, 2014). Indeed, a better understanding of mechanisms underlying vicarious SoA over these agents would allow users to feel that they are in charge of the system, and that the system responds to their actions (Shneiderman & Plaisant, 2010). Therefore, it is crucial to understand whether, and how, humans can experience vicarious SoA (and, by extension, control) over actions executed by an artificial system, and their subsequent outcomes.

In the next section, evidence of vicarious SoA in social contexts with artificial agents will be described according to the type of agent involved. First, the focus will be on studies involving computer agents, in the form of simple devices such as mechanical levers (e.g., Wohlschläger et al., 2003), or computer programs (e.g., Obhi & Hall, 2011; Sahaï, Desantis, Grynszpan, Pacherie, & Berberian, 2019). Then, the focus will be on studies involving embodied robots displaying different degrees of human-like features, such as non-anthropomorphic robots (e.g., Ciardo, Beyer, De Tommaso, & Wykowska, 2018; 2020) or human-like, robotic arms (e.g., Khalighinejad, Bahrami, Caspar, & Haggard, 2016).

1.8.1. The case of computers

In the context of vicarious SoA over artificial agents, an interesting question is whether the occurrence of this phenomenon is different in a social context with another human, compared to a computer partner. Wohlschläger and colleagues' study (2003; Experiment 1) addressed this question by investigating whether the IB effect arises similarly in self-, other human-, and machine-generated actions. Specifically, in this last condition, participants observed a mechanical lever that was programmed to move automatically to perform a keypress (*Machine* condition)

(Wohlschläger et al., 2003). Results indicated that the perceived onset of the machine-generated actions differed from both self- and other human-generated actions, which were indistinguishable from each other. Specifically, the perceived onset of both self-and other human-generated actions was estimated as occurring *later* than it did, signaling that the action was shifted towards the subsequent tone; hence, the IB effect emerged. However, it was not the case of machine-generated actions, as participants estimated those actions as happening *earlier* than they did. The authors hypothesized that such a result was due to a lack of visual information associated with the hand movement that was present in both the self and the other-human conditions. As a control, in the subsequent experiment, the authors used a rubber hand in the *Machine* condition (Wohlschläger et al., 2003; Experiment 2). The rubber hand was placed on the mechanical lever, in such a way that the index finger of the rubber hand would always move with the lever. Results showed that, with the rubber hand, the anticipatory effect reported in Experiment 1 was reduced; however, it did not induce the IB effect. Taken together, these findings confirmed that, when interacting with a machine, humans do not experience vicarious SoA over its actions (Wohlschläger et al., 2003). As an explanation, it has been proposed that failures in predictive mechanisms during the observation of machines may cause difficulties in experiencing vicarious SoA over machine-generated actions (Sahaï et al., 2017). Similar results have been reported by Poonian and Cunnington (2013; Experiment 2). The authors asked participants to watch a video and estimate the time interval between the observed keypress and the subsequent tone. Critically, the video could display either another human performing the keypress (*Observed Agent* condition) or a keypress performed automatically by the keyboard, which was pushed down without any visible agent acting (*Observed No-Agent* condition). Results showed that the IB effect was diminished when the tone was triggered by the automatic keypress (*Observed No-Agent* condition) compared to when the

human agent was present (*Observed Agent* condition) (Poonian & Cunnington, 2013). To explain these results, Sahaï and colleagues (2017) proposed two alternative, but non-exclusive, hypotheses (Sahaï et al., 2017). According to the first one, vicarious SoA may rely on a low-level motor prediction system. Specifically, when observing a machine acting, the predictive system fails to predict the sensory consequences of its actions because humans cannot simulate the machine motor schemes using their motor systems. Alternatively, vicarious SoA may rely on more high-level conceptual beliefs, based on which the other agent is a machine that has no control over the task. Therefore, when people observe actions that are generated by a mechanical agent, their beliefs about the fact that machines' actions are not driven by intentionality would inhibit or suppress vicarious SoA.

Other evidence showed that, when sharing a task with a computer program, the IB effect does not arise. In Obhi and Hall's study (2011), participants performed an IB task in which they had to perform a keypress at the time of their choice, which triggered a tone after a certain delay. They were partnered with another human or with a computer, separated from them by a curtain. Participants were told that, during the task, their partner could also trigger the tone, but in fact, it was always the participants who triggered the tone. Results showed that the IB effect occurred when participants believed that the other human caused the tone, as well as when they believed to be the one who caused the tone. However, the IB effect did not occur when participants believed that the computer caused the tone (Obhi & Hall, 2011). To explain such a result, the authors argued that the experience of agency for the actions performed by another agent is based on one's ability to understand that the other has intentions to act that is similar to one's intentions, which is not the case of computers; therefore, the IB effect did not occur (Obhi and Hall, 2011).

Similar results have been found when participants observed a computer program acting. For example, Sahaï and colleagues (2019) asked participants to perform a Joint Simon task (e.g., Ferraro, Iani, Mariani, Milanese, & Rubichi, 2011; Ciardo & Wykowska, 2018; Yamaguchi, Wall, & Hommel, 2018; see Dolk et al., 2014 for a review), in which usually pairs of participants perform spatial compatibility go/no-go task. The typical result of the task is that, when performing the go/no-go task together with another agent, a spatial compatibility effect emerges, whereas it does not when the very same task is performed alone. Such a result has been interpreted as an index of the fact that, when in a social context, humans tend to represent tasks as shared, thus including other's actions in their task representation (Sebanz, Knoblich, & Prinz, 2003; Tsai, Kuo, Hung, & Tzeng, 2008). Sahaï and colleagues (2019) combined this task with an interval estimation paradigm. Thus, after a response was given during the go/no-go task, a tone was presented and participants were asked to verbally report the time interval between the keypress and the tone, both for their and co-agent's response. Notably, the time interval estimation served as an implicit measure of vicarious SoA over the actions performed by the observed co-agent. The authors manipulated whether participants performed the task with another human agent (*Passive Observation Human* condition) or with a computer (*Passive Observation Computer* condition). Results showed that both spatial compatibility effect and vicarious SoA occurred only when the co-agent was the other human, and not when it was a computer (Sahaï et al., 2019). Such a result led the authors to suppose that, in the *Passive Observation Computer* condition, the Joint Simon effect and vicarious SoA might share common underlying mechanisms (Sahaï et al., 2019). Specifically, humans would automatically activate their sensorimotor representation when partnered with another human (Sebanz et al., 2003), because they experienced the action of the human co-agent as if was their action. It would be consistent with previous studies in which the

authors suggested that the social identity of the partner, i.e., the fact that the partner was perceived as an intentional agent or not, influences individuals' action representation abilities during a joint task (e.g., Stenzel et al., 2012; Wen and Hsieh, 2015). However, the same would not occur with a computer, as it would not be perceived as intentional. Thus, humans' sensorimotor system would not be activated to simulate and try to understand computer-generated actions. As a result, vicarious SoA would not occur.

This collection of evidence from the Human-Computer Interaction (HCI) field highlighted that people do not experience vicarious SoA over computer co-agents. However, it is still unclear whether the lack of vicarious SoA over computer's actions and outcomes is because people cannot use their motor schemes to represent computer's actions, given its disembodied nature, or because they do not perceive computers as intentional agents- or both. In this context, robots represent a useful, new tool allowing us to examine these factors separately, as they would allow manipulating both motor repertoire and degree of attributed intentionality independently from each other. This way, it would be possible to disentangle the specific contribution of these factors for the emergence of vicarious SoA.

Given that, the next paragraph will be dedicated to exploring recent findings on vicarious SoA in a social context with robotic artificial agents.

1.8.2. The case of robots

By definition, a robot is a programmable automated machine that can perceive and interpret information from the physical environment, and can act upon the environment (Russel & Norvig, 2002). Robots have three types of components: (1) a control system, such as the controller board, (2) the sensors, which can read the information from the surrounding environment, and (3) the actuators, which produce an effect on the environment of the robot. This allows for investigating

separately the role of action representation and intentionality attribution for vicarious SoA in HRI, as robots can display different motor repertoires and show various degrees of intentional behaviors. In a recent study, Ciardo and colleagues (2020) showed that when performing a risk-taking task with a robotic agent, participants reported reduced SoA over negative outcomes (Ciardo et al., 2020), as well as it is observed in contexts of interactions with other human agents (e.g., Beyer, Sidarus, Bonicalzi, & Haggard, 2017; Beyer, Sidarus, Fleming, & Haggard, 2018). Notably, reduced SoA seems to play a role in the diffusion of responsibility phenomenon, i.e., feeling less responsible for the consequences of one's actions (Bandura, 1991). In social situations, it would decrease the likelihood that people act in presence of other agents potentially able to act the same way they do.

In three experiments, the authors compared self-reported ratings of agency while human participants were interacting with the non-anthropomorphic Cozmo robot (Experiment 1), with a passive, non-agentic air pump (Experiment 2), or with another human being (Experiment 3). Participants' task was to rate the perceived control they felt on the outcome of their action while performing a task, in which they had to stop the inflation of a balloon before it would reach a pin and burst. Every action was associated with a negative outcome (i.e., losing a certain amount of points). Results showed that participants rated their SoA lower in trials in which the Cozmo robot was able to act compared to when they performed the task alone. However, the same results did not occur in Experiment 2, when the co-agent was a passive, non-agentic air pump. Results of Experiment 3 showed that the effect was in Cozmo and in the human co-agent (Ciardo et al., 2020) conditions.

Although this study did not directly address vicarious SoA towards a robotic agent, it showed that interacting with a non-anthropomorphic robotic agent affects SoA similarly to interacting with

other humans, but differently from non-agentic mechanical devices. Moreover, it gave two important hints regarding SoA during interactions with artificial agents. The first hint was that the attribution of intentional agency reduces individual SoA. Indeed, participants' SoA was affected only when interacting with the Cozmo robot (Experiment 1), and not with the non-agentic, passive device, i.e., the air pump (Experiment 2). As in both experiments participants were performing the task with another agent, the authors suggested that the reduction of SoA experienced when interacting with the Cozmo robot could not be interpreted as a consequence of the mere presence of an entity. This argued in favor of the interpretation that attribution of intentionality was crucial to reduce SoA, which was subsequently confirmed by the results of Experiment 3. Indeed, when performing the task in two sessions, i.e., one with Cozmo and once with another human, explicit SoA resulted to be similarly reduced in both conditions, indicating that attribution of intentional agency plays a crucial role in the reduction of SoA when interacting with a non-anthropomorphic robot (Ciardo et al., 2020).

The second hint was that, when studying SoA during social interactions with artificial agents, considering their embodiment may be critical, given that intentional agency is defined as the ability to plan and *act* (Ciardo et al., 2020). However, not only the embodiment seems to be important for the vicarious SoA phenomenon, but also the degree of human-like features displayed by the artificial co-agent one is interacting with. Khalighinejad and colleagues (2016) employed the IB paradigm to investigate whether observing the action of a human or of a non-human agent could influence individual implicit SoA (Khalighinejad et al., 2016). In their study, participants were asked to perform an IB task in which they observed a clock hand rotating; then, they had to report the position of the clock hand at the time of the action, or of the auditory tone following the action. As a between-subject manipulation, half of the participants performed the IB task with another

human agent, whereas the other half were partnered with a robotic hand having an anthropomorphic shape and remotely controlled by an experimenter. In both groups, participants were instructed to alternate with their co-agent in performing the task. During the task, for half of the participants, a barrier was placed between the two agents to prevent each member of the pair to see each other, whereas for the other half of participants no barrier was placed between the two agents, so that participants were allowed to look at co-agent's actions. Results showed that the IB effect occurred for both actions and tone outcomes when participants were partnered with the other human agent as well as when they were partnered with the robotic arm. Interestingly, the IB effect was enhanced when participants could see the other agent acting, independent of whether it was another human or a robotic hand (Khalighinejad et al., 2016). The authors argued that the possibility to see the other agent's hand might have allowed participants to represent its actions and the subsequent outcomes. The authors suggested that the comparable IB effect for actions performed by the human and the robotic hand might be because participants represented the robotic hand's actions as well as they did with the human hand's actions (Khalighinejad et al., 2016). The authors also argued that the comparable IB effect for human and robotic hand's actions might have been because participants were told that they were interacting with an "intentional", "human-like" agent (Khalighinejad et al., 2016), suggesting that SoA arises only when interacting with agents that are perceived as intentional. This might also explain the lack of the IB effect for actions executed by a mechanical lever (Wohlschläger et al., 2003). More specifically, Wohlschläger and colleagues (2003) argued that participants did not perceive the mechanical lever as an intentional system; consequently, vicarious SoA arose only when the co-agent was the other human, and not with the lever (Wohlschläger et al., 2003).

Therefore, it seems that the attribution of intentionality towards robotic agents is also a relevant aspect when investigating SoA. This is motivated by the fact that intentionality attribution is relevant for one's own SoA. Indeed, as demonstrated by Haggard and colleagues (2003), participants experienced implicit SoA, in the form of IB effect, only when their actions were voluntary, and not when they were involuntary (i.e., externally triggered by TMS stimulation (Haggard et al 2003). Thus, if intentionality attribution plays a role in the emergence of individual implicit SoA, it may be relevant also for vicarious SoA in the context of HRI.

Intentional robots?

In the past years, Daniel Dennett (1971, 1981) argued that the "default" strategy that humans use when interacting with mechanical systems is the *Design Stance*. In other words, people know that robots are designed to appear and behave in a certain way, and thus they explain and predict robots' behavior with reference to the way that the robots are designed to behave. However, people may also adopt the Intentional Stance (Dennett, 1971, 1981) towards robots, and explain their behavior with reference to mental states that humans attribute to the robots. In line with this, recent evidence demonstrated that, in some context, robots can be perceived as intentional agents, rather than pre-programmed artifacts (see Pérez-Osorio & Wykowska, 2020 for a review). For instance, in a recent study, Marchesi and colleagues (2019) presented to participants several fictional scenarios depicting the iCub robot (Metta, Sandini, Vernon, Natale, & Nori, 2008) while performing various activities. For each scenario, participants rated (by moving a slider on a scale) if they thought that iCub's behavior was motivated by a mechanical cause (such as malfunctioning or calibration) or by a mentalistic explanation (such as desire or curiosity). Although the results showed on average a slight bias towards the mechanistic explanation, a substantial number of mentalistic explanations were also given, suggesting that it is possible to induce the adoption of the *Intentional Stance* towards robotic agents (Marchesi et al., 2019).

The evidence reviewed so far suggested that, although little is still known about vicarious SoA towards robots, some aspects of the robotic agents may affect humans' vicarious SoA. First, as opposed to computers, robots are embodied agents, which may potentially lead humans to represent their actions as well as other humans' actions. This would be amplified by human-like features with which robots can be endowed. Second, people may attribute intentionality to robots, and perceive them as intentional agents. These two factors lead to the question of whether, and under which conditions, robots may elicit vicarious SoA in humans. These questions represented the core of my Ph.D. project.

1.9. The rationale of the Ph.D. Project

The aim of my Ph.D. project was twofold. The first aim (I) was to investigate whether humans can experience vicarious SoA over robot's actions and outcomes, and under what conditions. The second aim (II) was to disentangle whether vicarious SoA may serve as an implicit measure of intentionality attribution towards robots.

To address these questions, I conducted a series of studies employing the Intentional Binding (IB) paradigm (Haggard et al., 2002; see Moore & Obhi, 2012 for a review) as a reliable and well-established measure of implicit SoA.

The first study, reported in **Publication I**, was the first attempt to develop an IB paradigm adapted to an HRI protocol, to determine whether it was able to elicit the IB effect in the first place. In this study, we employed the non-anthropomorphic Cozmo robot (Anki Robotics), i.e., a small, portable robot to assess whether its social presence might affect individuals' SoA. In the subsequent study, reported in **Publication II**, the IB paradigm developed in the first study (Publication I) was used to investigate whether observing the Cozmo robot in action induced vicarious SoA in humans, and under what conditions. Specifically, in two experiments we wanted to disentangle the potential

role of (1) action representation, and (2) attribution of intentionality towards robots in the emergence of vicarious SoA towards robots. In the last study, reported in **Publication III**, we used the IB paradigm of the previous study (Publication II) with the humanoid iCub robot (Metta et al., 2008). The purpose was to assess whether (1) the human-likeness of the robot, and (2) attribution of intentionality towards it might play a role in the emergence of vicarious SoA towards robots.

Publication I – one's own SoA in the presence of a robot

The first study of my Ph.D. project, reported in Publication I, aimed at investigating whether the social presence of a robot affects one's own SoA by using an implicit measure as the IB effect. To this purpose, participants were asked to perform an Intentional Binding (IB) task, adapted to an HRI protocol, both alone (Individual Condition) and together with the non-anthropomorphic Cozmo robot (Social Condition). Participants' task was to observe a clock hand rotating, and to report the time of occurrence of both self- and robot-generated actions and outcomes. To test whether the presence of a robot as a potential agent affects individual agency we compared the IB effect for self-generated actions and outcomes between the two experimental conditions (Individual vs. Social Condition). We hypothesized that, if the social presence of the robot affected individual SoA, participants would have a smaller or null IB effect when Cozmo was in charge to perform the task (Social Condition) compared to when it did not (Solo Condition).

Publication II – vicarious SoA towards a non-anthropomorphic robot

The second study of my Ph.D. project aimed at investigating whether observing a non-anthropomorphic robot in action elicits vicarious SoA in humans, and under what conditions. In two experiments, we asked participants to perform an IB task alone (Solo Context) and with the non-anthropomorphic Cozmo robot (Social Context). Participants' task was to observe a clock

hand rotating and to report the time of occurrence of both self-and robot-generated actions and outcomes.

To assess the role of action representation, the Cozmo robot was programmed to perform either a physical action, i.e., a keypress (Experiment 1), or a "digital" action, i.e., "sending a signal" to the computer via Bluetooth. To assess the role of intentionality attribution, before both experiments participants filled out a questionnaire measuring individuals' tendency to attribute intentionality to robots. We hypothesized that, if action representation plays a role for vicarious SoA towards robots, the vicarious IB effect (i.e., vicarious SoA) would emerge only when the robot performed physical, and not digital, actions. Furthermore, we hypothesized that, if the attribution of intentionality plays a role for vicarious SoA in HRI, the magnitude of the vicarious IB effect would be predicted by the degree of individual tendency to attribute intentionality to robots, with a higher intentionality score predicting a larger vicarious IB effect.

Publication III – vicarious SoA towards a humanoid robot

The last study of my Ph.D. project aimed at investigating whether the vicarious SoA towards robots depends on (1) the human-like shape of the robot, (2) the intentionality attributed to it, or (3) both. To this purpose, participants performed an IB task both alone (Solo Context) and together with the humanoid robot iCub (Social Context). Based on the evidence collected in Publication II, here we focused our interest on vicarious SoA experienced for robot's actions only. Thus, participants were asked to report the time of occurrence of both self-generated and iCub's actions. As in the previous study (Publication II), before the experiment we also collected the individual tendency to attribute intentionality towards robots. We hypothesized that, if the human-like shape is the predominant factor contributing to the emergence of vicarious SoA towards robots, then a comparable IB effect should emerge for both self-generated and iCub's actions, with the intentionality attribution not

predicting the magnitude of the vicarious IB effect. In contrast, if attribution of intentionality plays the major role in the emergence of vicarious SoA towards robots, then the vicarious IB effect for iCub's actions would be predicted by the individual tendency to attribute intentionality to robots. Finally, if both human-like shape and intentionality attribution are crucial in the emergence of vicarious SoA, then we would expect a comparable IB effect for both self-generated and iCub's actions, with a higher intentionality attribution predicting a larger vicarious IB effect.

Publication I constitutes the manuscript of the conference paper "Roselli, C., Ciardo, F., & Wykowska, A. (2019). Robots improve judgments on self-generated actions: an Intentional Binding Study. In: Salichs M. et al. (eds) Social Robotics. ICSR 2019. Lecture Notes in Computer Science, vol.11876, pp. 88-97. Springer, Cham.".

Publication II constitutes the manuscript of the journal paper "Roselli, C., Ciardo, F., & Wykowska, A. (2021). Intentions with actions: The role of intentionality attribution on the vicarious sense of agency in Human-Robot interaction. Quarterly Journal of Experimental Psychology, 17470218211042003".

Publication III is the manuscript submitted for publication to Cognition, titled "Human-likeness and attribution of intentionality predict vicarious SoA over humanoid robot actions", authors: Roselli, C., Ciardo, F., Wykowska, A.

SECTION II- PUBLICATIONS

2.1. Publication I: Robots improve judgments on self-generated actions: an Intentional Binding study

Cecilia Roselli [1,2], Francesca Ciardo [1], and Agnieszka Wykowska [1]

[1] Social Cognition in Human Robot Interaction, Fondazione Istituto Italiano di Tecnologia, Center for Human Technologies, via Enrico Melen 83, Genova, Italy

[2] DIBRIS, Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi, Via all'Opera Pia 13, Genova, Italy

Authors Contribution

C.R. performed the study; analyzed the data; discussed and interpreted the results; and wrote the manuscript. F.C. and A.W. conceived and designed the study; discussed and interpreted the results; wrote the manuscript. All authors reviewed the manuscript.

2.1.1. Abstract

In near future, robots will become a fundamental part of our daily life; therefore, it appears crucial to investigate how they can successfully interact with humans. Since several studies already pointed out that a robotic agent can influence human's cognitive mechanisms such as decision-making and joint attention, we focus on Sense of Agency (SoA). To this aim, we employed the Intentional Binding (IB) task to implicitly assess SoA in Human-Robot Interaction (HRI). Participants were asked to perform an IB task alone (Individual condition) or with the Cozmo robot (Social condition). In the Social condition, participants were free to decide whether they wanted to let Cozmo press. Results showed that participants performed the action significantly more often than Cozmo. Moreover, participants were more precise in reporting the occurrence of a self-made action when Cozmo was also in charge of performing the task. However, this improvement in evaluating self-performance corresponded to a reduction in SoA. In conclusion, the present study highlights the double effect of robots as social companions. Indeed, the social presence of the robot leads to a better evaluation of self-generated actions and, at the same time, to a reduction of SoA.

Keywords: Human-Robot Interaction, Sense of Agency, Intentional Binding.

2.1.2. Introduction

In recent years, artificial agents (e.g. Siri, Alexa, Google Assistant) have started appearing more commonly in our houses. They are able to perform a huge number of autonomous activities: playing our favorite music, reminding us to take a pill, or helping us to reach a friend's place. However, since these artificial agents are not embodied, they are not able to manipulate the physical world. In the near future, also robots will take part in our everyday life, as useful supportive assistants at home or at work (Glasauer, Huber, Basili, Knoll, & Brandt, 2010). Thanks to their embodiment, robots not only will comply with our requests, but also will act directly on our environment and manipulate it (Wykowska, Chaminade, & Cheng, 2016). Therefore, it appears crucial to understand whether robots as social companions can affect basic cognitive mechanisms in humans. Some evidence showed that the mere presence of a robot influenced the behavior of a human partner, leading her/him to follow robot's recommendation in a decision-making task (Shinozawa, Naya, Yamato, & Kogure, 2005), or even to afford a greater peripersonal space (Bainbridge, Hart, Kim, & Scassellati, 2008).

Beyond the mere presence, the actions of a robotic agent actually have an impact on human's behavior. For example, when performing a target discrimination task with the iCub robot, participants' attentional orienting was biased by robot's gaze direction similarly to when human eyes are presented (Kompatsiari, Ciardo, Tikhanoff, Metta, & Wykowska, 2018; Kompatsiari, Pérez-Osorio, De Tommaso, Metta, & Wykowska, 2018). Similar findings have been shown also for joint actions. For instance, when performing a joint Simon task (Sebanz, Knoblich, & Prinz, 2003; Ciardo & Wykowska, 2018) participants coordinated their actions with a non-humanoid robot; however, this was true only when they believed that the robot was controlled by a human being (Stenzel et al., 2012).

Together, these findings demonstrated that robots are able to modify humans' cognitive mechanisms, such as decision-making and joint attention, in a similar fashion as they occur in human-human interaction. However, other mechanisms of human cognition are still poorly investigated in HRI; one of these is Sense of Agency (SoA). SoA has been defined as the feeling of control that we experience over our actions and their outcomes (Gallagher, 2000). Given its pivotal role in the embodied nature of the Self, disruption of SoA may lead to unpleasant consequences, like the misrecognition of ourselves as the authors of our thoughts, feelings, and actions. For instance, disruption of SoA has been reported in schizophrenia patients, who find difficult to distinguish between self- and externally generated events (Daprati et al., 1997; Gallagher, 2012).

In the context of HRI, Ciardo and colleagues demonstrated that SoA over self-generated actions is reduced when performing a task with a robotic agent (Ciardo, De Tommaso, Beyer, & Wykowska, 2018). Specifically, when participants performed costly actions (i.e. losing a various amount of points) together with a robot, they rated their SoA lower compared to when they performed the same task alone. However, although interesting, these results have been obtained with explicit measures only, thus it remains unclear whether the social presence of a robot can affect SoA in humans also at an implicit level.

One of the most common implicit measures to investigate SoA is based on recording variations in time perception related to action effects (see Haggard, 2017 for a review). The typical result is to perceive the time interval between a self-voluntary action and its sensory consequence (e.g., a tone) as shorter in time than its actual duration. This effect is known as *Intentional Binding* (IB) (Haggard, Clark, & Kalogeras, 2002), and it occurs only when people perform self-voluntary actions. Indeed, if the same action-effect chain is produced by involuntary movements, then the

IB effect is not reported (Haggard & Clark, 2003; Tsakiris & Haggard, 2003). In terms of SoA in a social context, Sahaï and colleagues (Sahaï, Desantis, Grynszpan, Pacherie, & Berberian, 2019) showed that IB for self-generated actions does not differ whether people perform the task alone or interact with another human, but it decreases when they interact with a computer.

2.1.3. Aim

In the present study, we aimed to investigate whether the social presence of the robot can affect SoA in humans by using an implicit measure. To this end, participants were asked to perform an IB task (Strother, House, & Obhi, 2010; Obhi & Hall, 2011a) alone (Individual Condition) or with the Cozmo robot (Anki Robotics) (Social Condition). Notably, both participants and Cozmo performed the task by executing the action, i.e. the keypress, at the time of their choice. We hypothesized that, if the social presence of the robot can actually influence participants' SoA, inducing a reduction of SoA as in Ciardo and colleagues' study (Ciardo et al., 2018), then participants' performance is expected to differ across conditions. Specifically, in Social blocks participants would be better at judging the position of the clock hand compared to the Individual blocks. As a consequence, smaller or null IB was expected in the Social compared to the Individual Condition.

2.1.4. Materials and Methods

Participants. Eighteen right-handed young adults (mean age =22.47, SD= 3.14, 5 males) have been recruited to take part in the study. Sample size has been estimated according to previous experiments (Strother et al., 2010; Obhi & Hall, 2011a) and to *a priori* power analysis performed in G*Power v. 3.1.9.1 (see Faul, Erdfelder, Lang, & Buchner, 2007 for more information). It indicated that a sample size of N = 12 was needed in order to detect a medium effect size [Cohen's d for repeated measures (Dz) = 0.63, alpha (one-tailed) =.05 and power = 0.80] for within-subjects

comparisons. The study has been conducted in accordance with the ethical standards laid down in the 2013 Declaration of Helsinki and has been approved by the local ethical committee (Comitato Etico Regione Liguria). All participants gave written informed consent prior to the experiment. They received an honorarium of 10 € per hour for their participation. The experimental session lasted around 60 minutes. At the end of the experiment, participants were debriefed about the purpose of the study.

Apparatus and Stimuli. The experimental setup consisted of a mobile Android device with the standard Cozmo application running in 'SDK enabled option' (https://cozmosdk.anki.com/docs/install-win-dows.html); one computer connected with Cozmo through the Android Debug Bridge (cosmosdk.anki.com/docs/adb.html); one 21' inches screen (1920 x 1080 pixels) to display the task; two keyboards to collect responses during the experiment. Participants and Cozmo were seated side by side at approximately 60 cm away from the computer screen. A keyboard was placed in front of the participants and in front of Cozmo (see **Figure 1**). Stimuli presentation, response collection, and the Cozmo robot were controlled with OpenSesame (see Ciardo et al., 2018 for the procedure of how to integrate Cozmo).

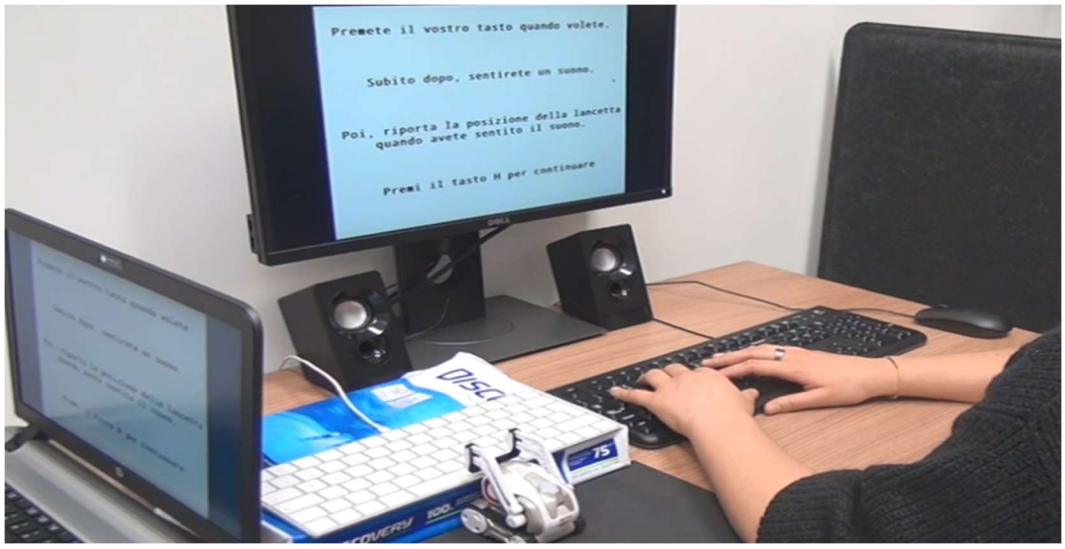

Figure 1. Experimental setup.

Procedure. Participants were asked to perform the Intentional Binding task alone (Individual condition) or with Cozmo (Social Condition). They were presented with an image of the clock (10.6 ° visual angle) with a red rotating clock hand. Each trial started with a black fixation dot on a white background for 1000 ms, followed by the image of a clock with a static clock hand for 500 ms. Then, the clock hand started to rotate randomly from one of the 12 five-minutes positions of the clock, in order to complete a unique full rotation in 2560 ms. At the end of each trial, participants were asked to report the position of the clock hand at the time of the event of interest (either keypress or tone play). The task comprised four different types of blocks: two types of Baseline blocks (tone or action), and two types of Operant blocks (tone or action) (see **Table 1**). For both the Individual and the Social Condition, each block was repeated twice; following the procedure of Obhi & Hall (2011b), only the Baseline Tone block was performed once, since no action was required.

At the beginning of each block, participants were informed whether they were performing the task alone (Individual Condition) or with Cozmo (Social Condition). If the block belonged to the Individual condition, Cozmo moved away from its keyboard and entered into the sleep mode. When a block of the Social Condition started, Cozmo was programmed to wake up and reach the keyboard. In those blocks belonging to the Social Condition, participants were instructed that if they wanted they could let Cozmo do the task alone in their place. Cozmo was programmed to tap the bar during the clock hand rotation at a random time (see video "**Publication I**": https://osf.io/23jmt/?view_only=f58dfc2c426f45ba93a7eff5f931c43f). The task comprised 14 blocks of 24 trials each, for a total number of 336 trials. Blocks were randomly assigned to either Individual (7 blocks) or Social Condition (7 blocks). Therefore, once in the Individual and once in the Social Condition, each participant performed 1 BT, 2 BA, 2 OT and 2 OA blocks (see **Table**

1). In order to prevent any habituation effect, both Individual and Social blocks were presented in a random order within participants. A practice session of the entire task (i.e. 14 trials, one per condition) was administered.

Block type	**Task**
Baseline Tone (BT)	A tone (440 Hz, 700 ms) is played at a random time while the clock hand (length= 170 pixels) is rotating. Participants have to judge at which time on the clock the sound event occurs. No action is required.
Baseline Action (BA)	Participants have to press the spacebar at any moment while the clock hand is rotating. They have to report the position of the clock hand when they act. No auditory feedback occurs.
Operant Tone (OT)	Participants have to press the spacebar at the time of their choice while the clock hand is rotating. 250 ms after the keypress, the tone (440 Hz, 700 ms) is presented while the clock hand is still rotating. Participants have to judge at which time on the clock the sound event occurs.
Operant Action (OA)	Participants have to press the spacebar at any moment during the clock hand rotation. 250 ms after the keypress, the tone (440 Hz, 700 ms) is presented while the clock hand is still rotating. Participants have to report the position of the clock hand when they act.

Table 1. The Baseline and the Operant blocks.

2.1.5. Data analysis

Our dependent measure was the percentages of humans' and Cozmo's responses in the Social blocks and the judgment errors (JEs) in reporting the critical event across the Individual and the Social Blocks. Related to the percentages, only Social blocks were considered; then, the percentages of responses for each agent were estimated only for those trials in which an action occurred and compared with a chi-square test. JEs were estimated as the difference between the position of the clock hand reported by participants and the actual onset of the critical event (i.e. action or tone play). For Social blocks, JE was estimated only for those trials in which participants acted. Then, for each block we calculated the average JE and its standard deviation; trials in which the JE deviated more than ± 2.5. SD from the participants' mean were excluded from the analysis. After the outliers' removal, participants with a total number of valid trials lower than 24 were

excluded from the subsequent analysis. According to this criterion, for action blocks four participants were excluded, whereas, for tone blocks, data of eight participants were not analyzed. Therefore, for 14 participants in action blocks, and for 10 participants in tone blocks, the IB effect was estimated as the difference between the mean JE for the Baseline and the mean JE for the corresponding Operant block, for both Individual and Social conditions.

Given that JEs were not normally distributed, Wilcoxon signed-rank tests were used to compare JEs and IB across conditions (Individual, Social) and Block Types (Baseline, Operant). The threshold for level of significance was set at $p < .05$, rank-biserial correlation coefficient (r_b) is reported as an index of the effect size. 95% confidence intervals of the means are reported.

2.1.6. Results

In Social blocks, participants acted in 75.3% of trials (95% CI [73.5%, 77%]), letting Cozmo perform the task in their place in the remaining 24.7% of trials (95% CI [23%, 26.5%]). This difference was statistically significant from chance ($\chi^2 = 4.468$, $p < .001$). Interestingly, the Human/Cozmo action ratio was constant across all the Social blocks in which an action was required (see **Figure 2**).

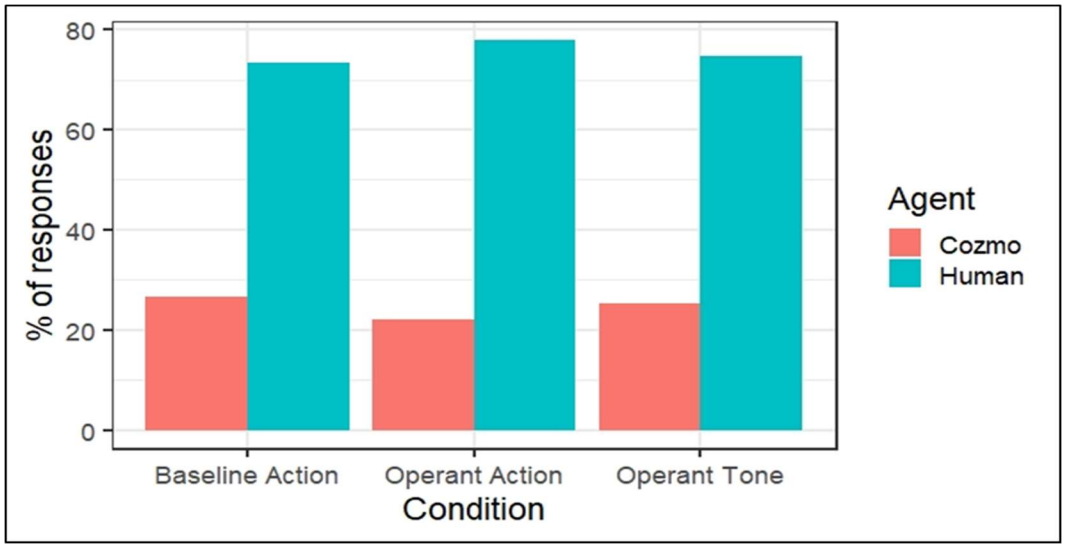

Figure 2. Percentages of responses in each Social Condition.

Action blocks. Results showed that when participants performed the task alone, JEs were smaller in the Baseline (M= -28.25, SD=33.39, 95% CI [-43.94, -12.43]) than in the Operant blocks (M= 26.82, SD=112.61, 95% CI [-30.16, 83.67]), W= 13.00, p= .011, r_b = - .75. However, this was not true for the Social Condition, as no differences occurred in JEs between the Baseline (M= -15.25, SD= 38.08, 95% CI [-35.58, 4.19]) and the Operant blocks (M= 5.08, SD= 68.73, 95% CI [-30.77, 41.46]), W=25.00, p= .091, r_b = - .52 (see **Figure 3, Panel A**). Moreover, results showed that JEs marginally differed across conditions for Baseline blocks only (Individual: M= -28.25, SD= 33.39, 95% CI [-43.94, -12.43]; Social: M= -15.25, SD= 38.08, 95% CI [-35.58, 4.19]), W= 21.00, p= .049, r_b = - .60. No differences in JEs across conditions occurred for Operant blocks, (Individual: M= 26.82, SD= 112.61, 95% CI [-30.16, 83.67]; Social: M= 5.08, SD= 68.73, 95% CI [-30.77, 41.46]), W= 49.00, p= .85, r_b = - .07. Finally, the comparison between IB effects across conditions showed a larger IB in the Individual (M= 55.08, SD= 106.3, 95% CI [2.10, 107.29]) compared to the Social Condition (M= 20.33, SD= 39.45, 95% CI [0.44, 39.52]), W= 85.00, p= .04, r_b = .61.

Tone blocks. Results showed that when participants performed the task alone, JEs were larger in the Baseline (M= 44.68, SD= 44.63, 95% CI [18.66, 70.85]) than in the Operant blocks (M= -42.02, SD= 89.45, 95% CI [-96.03, 9.91]), W= 55.00, p= .002, r_b = - .75. This was the case also for the Social Condition, with larger JEs in the Baseline (M= 23.61, SD= 38.95, 95% CI [1.58, 45.48]) compared to Operant blocks (M= -59.52, SD= 105.6, 95% CI [-119.99, -0.04]), W= 48.00, p= .04, r_b = - .75 (see **Figure 3, Panel B**). No differences in JEs occurred across conditions both for Baseline, W= 21.00, p= .049, r_b = - .60 (Individual: M= 44.68, SD= 44.63, 95% CI [18.66, 70.85]; Social: M= 23.61, SD= 38.95, 95% CI [1.58, 45.48]), and Operant blocks, W= 49.00, p= .85, r_b = - .07 (Individual: M= -42.02, SD= 89.45, 95% CI [-96.03, 9.91]; Social: M= -59.52 , SD= 105.6 95% CI [-119.99, -0.04]). No differences in the IB effect occurred across conditions, W=

30.00, p= .846, r_b = .091 (Individual: M= -86.70, SD= 71.69, 95% CI [-128.43, -46.03]; Social: M= -83.13, SD= 99.11 95% CI [-142.26, -24.25]).

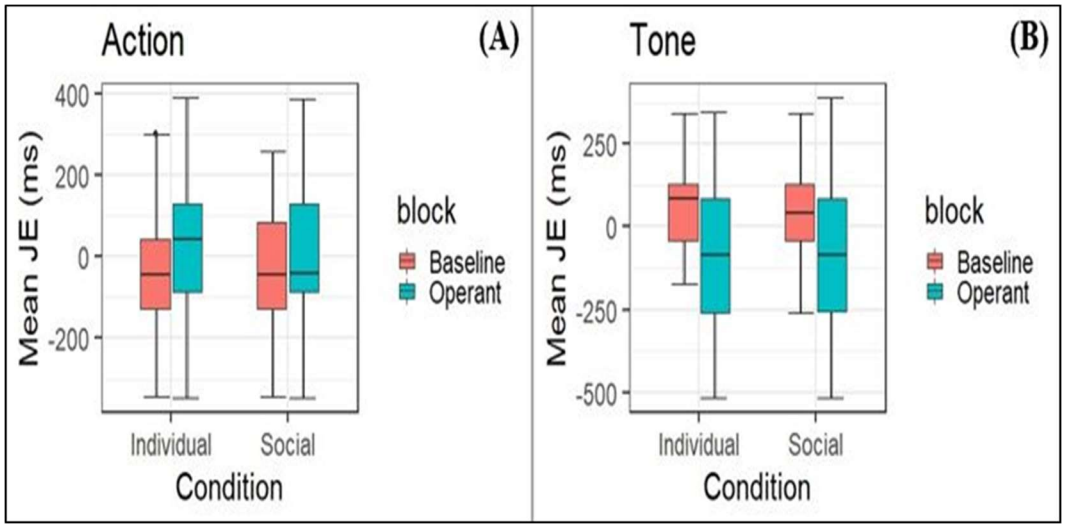

Figure 3. **Panel (A)**: Mean JEs for Baseline and Operant block as a function of Condition (Individual, Social) when the critical event was the action (action blocks). **Panel (B)**: Mean JEs for Baseline and Operant block as a function of Condition (Individual, Social) when the critical event was the tone (tone blocks).

2.1.7. Discussion

In this experiment, we sought to determine whether the social presence of the robot might influence SoA for self-generated actions. To this end, we employed an Intentional Binding (IB) task as an implicit measure to investigate SoA. As in the standard versions of the IB task where participants knew they always were the initiator of the action (Obhi & Hall, 2011b), we let participants choose whether they wanted to press the spacebar, or let Cozmo press in their place. We predicted that, if the social presence of the robot can actually have an impact on human's SoA, participants' performance would be different across conditions, i.e. they would judge better the occurrence of self-voluntary actions when doing the task with the robot compared to when they perform the task alone. Otherwise, if the robot does not influence humans' SoA, participants' performance in Social Condition would have mirrored the one in Individual Condition.

Overall participants let Cozmo respond only in the 30% of the Social trials. This trend in responses was constant across different block types, suggesting a general attitude toward the robot. It may be possible that they perceived the Social Condition as a competition between themselves and the robot. If this has been the case, then participants may have been triggered to act faster than Cozmo, resulting in a lower percentage of Cozmo trials. Another possible explanation is that given the nature of the robot we used, participants might have perceived the Cozmo robot as not competent enough to perform the task, thus they took the responsibility of the task and acted above the level of chance. Future studies should investigate these possibilities by addressing how individual differences in attitude towards the robot and perceived competence may affect the possibility to let it act. In line with previous studies (Haggard & Clark, 2003; Obhi & Hall, 2011b), when participants performed the task alone (Individual Condition) JEs for action events were smaller for the Baseline than for the Operant blocks, whereas, when the critical event was a tone, JEs were smaller for the Operant than for the Baseline blocks. The reversed pattern can be explained by the fact that the direction of the IB effect depends on when the critical event occurred on the temporal line in the Operant blocks.

When the critical event was the action (i.e. the keypress), since it preceded the sensory effect (i.e. the sound) on the temporal line, the actual occurrence of the action was bound to the occurrence of the tone, leading to perceive the temporal interval between the action and the sensory outcome as longer. Thus, in Operant blocks, participants reported the action as delayed compared to a condition in which it did not produce any sensory outcome. When participants had to judge when the sound occurred, given that the tone followed the keypress on the temporal line, its occurrence was bound to the preceding event (i.e. the action). Therefore, participants perceived the temporal interval between the action and the sensory outcome as shorter. Thus, in Operant blocks,

participants reported the tone earlier compared to when the sound is passively presented. Our results showed that the presence of the robot affected JEs only when the critical event is an action, as indicated by the lack of IB in the Social Condition only. The robot did not affect JEs when the critical event was a tone. Interestingly, previous studies showed that a lack of IB is typically reported for unintentional self-generated actions (Haggard et al., 2002; Chambon, Moore, & Haggard, 2015). In our Social Condition for action blocks, the lack of IB was driven by a reduction in JEs both in the Baseline and in the Operant blocks, suggesting that when the robot was also in charge to perform the task, participants were better in reporting when their own action occurred. These latter results can be due to a well-known phenomenon in psychology named social facilitation, i.e. the fact that in the presence of a social companion, humans' performance is enhanced (Zajonc, 1965).

Our results suggest that the social presence of Cozmo had a double effect. Indeed, on one hand the presence of the robot led to an improvement in the evaluation of self-generated actions. On the other hand, it might bring a reduction of self-agency, similarly to what occurred in human-human contexts (Beyer et al., 2017).

2.1.8. Conclusions

In the present study, we raised the question of whether a robot, as a social companion, can have an influence on human's SoA. According to our results, participants were better at judging the occurrence of their own actions when Cozmo was also in charge to perform the task, confirming that the social presence of the robot actually influences human's agency. On the other side, in action blocks a lack of IB in Social Condition compared to Individual condition corresponds to a reduction of SoA. Future studies should provide further evidence about the influence of robots on human's SoA, in order to build new robotic agents well-tailored on human's cognition.

2.1.9. Funding

This work has received support from the European Research Council under the European Union's Horizon 2020 research and innovation program, ERC Starting Grant, G.A. number: 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.

2.1.10. References

Android Debug Bridge, cozmosdk.anki.com/docs/adb.html, last accessed 2019/5/22.

Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008, August). The effect of presence on human-robot interaction. In RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication (pp. 701-706). IEEE.

Beyer, F., Sidarus, N., Bonicalzi, S., & Haggard, P. (2017). Beyond self-serving bias: diffusion of responsibility reduces sense of agency and outcome monitoring. Social cognitive and affective neuroscience, 12(1), 138-145.

Chambon, V., Moore, J. W., & Haggard, P. (2015). TMS stimulation over the inferior parietal cortex disrupts prospective sense of agency. Brain Structure and Function, 220(6), 3627-3639.

Ciardo, F., De Tommaso, D., Beyer, F., & Wykowska, A. (2018, November). Reduced sense of agency in human-robot interaction. In International conference on social robotics (pp. 441-450). Springer, Cham.

Ciardo, F., & Wykowska, A. (2018). Response coordination emerges in cooperative but not competitive joint task. Frontiers in psychology, 9, 1919.

Cozmo SDK Installation for Windows, https://cozmosdk.anki.com/docs/install-win-dows.html, last accessed 2019/5/22.

Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J., & Jeannerod, M. (1997). Looking for the agent: an investigation into consciousness of action and self-consciousness in schizophrenic patients. Cognition, 65(1), 71-86.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. Trends in cognitive sciences, 4(1), 14-21.

Gallagher, S. (2012). Multiple aspects in the sense of agency. New ideas in psychology, 30(1), 15-31.

Glasauer, S., Huber, M., Basili, P., Knoll, A., & Brandt, T. (2010, September). Interacting in time and space: Investigating human-human and human-robot joint action. In 19th International Symposium in Robot and Human Interactive Communication (pp. 252-257). IEEE.

Haggard, P. (2017). Sense of agency in the human brain. Nature Reviews Neuroscience, 18(4), 196-207.

Haggard, P., & Clark, S. (2003). Intentional action: Conscious experience and neural prediction. Consciousness and cognition, 12(4), 695-707.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. Nature neuroscience, 5(4), 382-385.

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

Kompatsiari, K., Pérez-Osorio, J., De Tommaso, D., Metta, G., & Wykowska, A. (2018, October). Neuroscientifically-grounded research for improved human-robot interaction. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3403-3408). IEEE.

Obhi, S. S., & Hall, P. (2011a). Sense of agency in joint action: Influence of human and computer co-actors. Experimental brain research, 211(3-4), 663-670.

Obhi, S. S., & Hall, P. (2011b). Sense of agency and intentional binding in joint action. Experimental brain research, 211(3-4), 655.

Sahaï, A., Desantis, A., Grynszpan, O., Pacherie, E., & Berberian, B. (2019). Action co-representation and the sense of agency during a joint Simon task: Comparing human and machine co-agents. Consciousness and Cognition, 67, 44-55.

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: just like one's own?. Cognition, 88(3), B11-B21.

Shinozawa, K., Naya, F., Yamato, J., & Kogure, K. (2005). Differences in effect of robot and screen agent recommendations on human decision-making. International journal of human-computer studies, 62(2), 267-279.

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

Strother, L., House, K. A., & Obhi, S. S. (2010). Subjective agency and awareness of shared actions. Consciousness and cognition, 19(1), 12-20.

Tsakiris, M., & Haggard, P. (2003). Awareness of somatic events associated with a voluntary action. Experimental brain research, 149(4), 439-446.

Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1693), 20150375.

Zajonc, R. B. (1965). Social facilitation. Science, 149(3681), 269-274.

2.2. Publication II: Intentions with actions: the role of intentionality attribution on the vicarious sense of agency in Human-Robot Interaction

Cecilia Roselli [1,2], Francesca Ciardo [1], and Agnieszka Wykowska [1]

[1] Social Cognition in Human Robot Interaction, Fondazione Istituto Italiano di Tecnologia, Center for Human Technologies, via Enrico Melen 83, Genova, Italy

[2] DIBRIS, Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi, Via all'Opera Pia 13, Genova, Italy

Authors Contribution

C.R., F.C., and A.W. conceived and designed the study; C.R. performed the study and analyzed the data; C.R., F.C., and A.W. discussed and interpreted the results and wrote the manuscript. All authors reviewed the manuscript.

2.2.1. Abstract

Sense of Agency (SoA) is the feeling of control over one's actions and their consequences. In social contexts, people experience a "vicarious" SoA over other humans' actions; however, the phenomenon disappears when the other agent is a computer. The present study aimed to investigate factors that determine when humans experience vicarious SoA in human-robot interaction (HRI). To this end, in two experiments we disentangled two potential contributing factors: (1) the possibility of representing the robot's actions, and (2) the adoption of Intentional Stance toward robots. Participants performed an Intentional Binding (IB) task reporting the time of occurrence for self- or robot-generated actions or sensory outcomes. To assess the role of action representation, the robot either performed a physical keypress (Experiment 1) or "acted" by sending a command via Bluetooth (Experiment 2). Before the experiment, attribution of intentionality to the robot was assessed. Results showed that when participants judged the occurrence of the action, vicarious SoA was predicted by the degree of attributed intentionality, but only when the robot's action was physical. Conversely, digital actions elicited reversed effect of vicarious IB, suggesting that disembodied actions of robots are perceived as non-intentional. When participants judged the occurrence of the sensory outcome, vicarious SoA emerged only when the causing action was physical. Notably, intentionality attribution predicted vicarious SoA for sensory outcomes independently of the nature of the causing event, physical or digital. In conclusion, both intentionality attribution and action representation play a crucial role for vicarious SoA in HRI.

Keywords: Vicarious Sense of Agency, Human-Robot Interaction, Intentional Binding, Intentional Stance.

2.2.2. Introduction

Sense of Agency (SoA) is the experience of identifying oneself as the author of an action and its consequences (Gallagher, 2000); it allows distinguishing self-generated actions from those generated by others or by external events (David, Newen, & Vogeley, 2008).

Traditionally, SoA has been measured by recording the perceived duration of the time interval between a self-generated voluntary action and its sensory outcome (e.g., a tone). The typical result is that the perceived time of the action is shifted towards the perceived time of the tone, and the tone is shifted back towards the action that caused it. This temporal compression leads to the Intentional Binding (IB) effect (Haggard & Clark, 2003; see Moore & Obhi, 2012 for a review).

In a social context[1] with other humans, IB occurs not only for self-generated actions but also for the actions of the partner (Strother, House, & Obhi, 2010), leading to a "vicarious" SoA. Notably, it does not occur when the co-agent is a computer (see Limerick, Coyle, & Moore, 2014 for a review). For instance, in a task when people evaluated IB for self-, other- (human) and machine-generated actions (Wohlschläger, Haggard, Gesierich, & Prinz, 2003), results showed a similar IB effect for self- and other- (human) generated actions, whereas no IB effect occurred for machine-generated actions.

Vicarious SoA can be explained with reference to the ideomotor theory, according to which actions are represented in terms of their perceivable sensory effects (Prinz, 1997). From the perspective of ideomotor theory, motor representations are anticipations of the sensory feedback from the action they represent. When observing others' actions, although to a lesser extent than during self-action

[1] Vicarious SoA has been investigated in different social scenarios, ranging from mere observation of a co-agent acting (Wohlschläger et al., 2003) to joint actions performed together with a co-agent, whether another human (Capozzi et al., 2016) or an artificial system, such as a computer (Obhi & Hall, 2011; Sahaï, Desantis, Grynszpan, Pacherie, & Berberian, 2019) or a robot (Grynszpan et al., 2019). Following APA's dictionary (https://dictionary.apa.org/social-context), when using the term "social context" we refer to a shared physical environment in which humans can observe, perceive, and evaluate other agents' actions.

execution, the perception of his/her action recruits, in the observer, the representational structures that are also involved in one's planning and control of those actions (Massen & Prinz, 2009). Consequently, humans can formulate accurate predictions about observed action's outcome as well as when they do for themselves (Springer, Hamilton, & Cross, 2012).

In this framework, vicarious SoA has been suggested to be the consequence of the activation at the neural level of action representation while observing other humans' actions. Indeed, evidence showed that implicit agency over actions generated by another human may depend on one's abilities to represent the partner's actions (Sahaï, Pacherie, Grynszpan, & Berberian, 2017). Notably, similar effects have been demonstrated on robot action observation, and specifically when the motion of the robot appeared to be biologically plausible (e.g., Chaminade, Franklin, Oztop, & Cheng, 2005; Liepelt, Prinz, & Brass, 2010).

In line with that, one possible explanation to account for the lack of vicarious SoA in human-computer interaction is the following. When the co-agent is a computer, the disembodied nature of the agent, together with a lack of perceivable action effects (i.e., an effector moving in the environment), would not allow the activation of action representation, affecting humans' ability to represent the cause-effect link between action and its sensory consequences (Ramnani & Miall, 2004; Sahaï et al., 2017). Therefore, accurate predictions about an outcome based on its cause (i.e., the computer program) could not be formulated, and people would not experience SoA over computer-generated actions (Obhi & Hall, 2011).

In line with this speculation, recent evidence showed that the observation of an action performed by a human-like automaton, i.e., an anthropomorphic hand with servo-actuated fingers, induced vicarious SoA similarly to the observation of another human performing the same action (Khalighinejad, Bahrami, Caspar, & Haggard, 2016). We might hypothesize that the human-like

hand, in terms of shape and physical motion, would have allowed people to represent the machine-generated actions; hence, vicarious SoA occurred. If it was the case, it would suggest a link between embodied physical actions and the possibility to represent them.

An alternative explanation for the lack of vicarious SoA in human-computer interaction is that people do not attribute intentionality to computers. According to Daniel Dennett (1971), humans adopt different strategies to predict and explain the behaviors of the system they are interacting with. When the system is a human being, people tend to adopt the Intentional Stance (Dennett, 1971) to explain his/her behavior referring to mental states such as desires, beliefs, and intentions. Conversely, when the system is artificial, people are more likely to adopt a Design Stance (Dennett, 1971) and explain its behavior referring to the way it was designed to behave. Thus, humans may not experience vicarious SoA in human-computer interaction because they do not attribute intentional agency to the system (Berberian, Sarrazin, Le Blaye, & Haggard, 2012; Sahaï et al., 2019, but see also Grynszpan et al., 2019 for different results).

This hypothesis is supported by findings showing that, for the IB effect to occur about one's actions, the actions must be voluntary and intentional. For instance, in a recent study investigating self-agency, Buehner (2015) disentangled the role of causality and intentionality in the occurrence of the IB effect[2]. Specifically, the author suggested that the effect is boosted when an agent acts intentionally, compared to when the action is driven by involuntary movements (e.g., Haggard & Clark, 2003; Haggard, Clark, & Kalogeras, 2002; Tsakiris & Haggard, 2003). In line with that,

[2] Haggard and collaborators (2002) were the first to interpret the IB phenomenon as the result of the perceptual attraction between intentional actions and their outcomes. Several studies demonstrated that the IB effect occurs in the absence of intentional actions, as long as the two events are perceived as causally linked. Therefore, the label "causal binding" seemed to be more appropriate than "intentional binding" (Buehner & Humphreys, 2009; Moore, Lagnado, Deal & Haggard, 2009) until the study of Buehner (2015). After having observed IB only in the voluntary action condition, Buehner (2015) proposed that the causal binding is strengthened when the cause of the action is intentional. Accordingly, although it is still an open question whether the IB effect and causal binding are equivalent, intentionality attribution seems to be at least partially involved in IB effect.

Obhi & Hall (2011) explained the lack of agency over computer's actions as caused by participants' disbelief in computers' ability to have intentions to act (as is the case of human intentional actions). In other words, authors speculated that, if humans do not attribute intentionality to the other agent, it is impossible to experience vicarious SoA (Obhi & Hall, 2011). Therefore, vicarious SoA would be unlikely to occur when the observed agent is perceived as a system that passively performs predetermined commands (Sahaï et al., 2019).

In this context, it is important to note that robots are ambiguous agents. On one hand, they are programmed artificial agents, which makes their actions involuntary and unintentional. On the other hand, through their embodiment, robots can physically act in the environment by executing actions that generate sensory effects that are similar to those of human actions. Consequently, their embodiment should allow for the activation of a representation of the cause-effect link for their actions (Massen & Prinz, 2009; Prinz, 1997). Moreover, it has been shown that, when facing a robot, in some contexts people could adopt the Intentional instead of Design stance to explain its behavior (Marchesi et al., 2019, but see also Perez-Osorio & Wykowska, 2020 for a review), so that humans may treat them as intentional agents.

2.2.3. Aims

The present study aimed to investigate factors that determine when humans experience vicarious SoA in human-robot interaction (HRI). To this end, we designed two experiments disentangling the role of two potential contributing factors: (1) the possibility of representing the robot's action at a neural level in terms of sensory consequences, and (2) the adoption of Intentional Stance toward robots. In the following, the paper is structured according to these two aims.

2.2.4. The role of action representation for vicarious SoA in HRI

In two experiments, participants performed an IB task (Haggard et al., 2002; Strother et al., 2010; Obhi & Hall, 2011) alone or with the Cozmo robot (Anki Robotics). Cozmo is a non-anthropomorphic robot that we decided to use in our paradigm to avoid any additional confounds driven by the physical similarity in appearance with humans (Epley, Waytz, & Cacioppo, 2007; Khalighinejad et al., 2016) inherent in humanoid robots. To determine whether the possibility of representing the physical cause-effect link contributes to the IB effect, we manipulated the way Cozmo executed the causing action across two experiments. In Experiment 1, the robot performed a keypress similarly to the human partner (i.e., an embodied and physically perceivable action). Conversely, in Experiment 2 participants were instructed that Cozmo "acted" by sending a command via Bluetooth to the keyboard (i.e., it performed a digital, non-embodied action). In this case, since digital actions would not generate similar sensory effects in the environment to those generated by physical actions, it would be more difficult to activate an action representation of the cause-effect link between action and outcome, thereby making it more difficult to represent the robot's actions. According to our reasoning, if vicarious SoA is mainly driven by the possibility to represent the cause-outcome link, then we would expect vicarious IB effect only in Experiment 1 when the robot performed a physical (embodied) action, and not in Experiment 2 when the robot sent the command via Bluetooth (i.e., digital action). For the Individual contexts, since participants were asked to voluntarily perform self-generated actions, we expected to always find an IB effect.

2.2.4.1. Experiment 1

2.2.4.1.1. Materials and Methods

Participants. Thirty-six right-handed young adults (range: 18-40 years old, M_{age} = 24.47, SD_{age} ± 4.87, 13 males) were recruited to take part in the study. All participants had a normal or

corrected-to-normal vision and were naïve to the purpose of the study. We estimated the sample size based on two different a priori power analyses performed with G*Power v. 3.1.9.1 (see Faul, Erdfelder, Lang, & Buchner, 2007 for more information). For the IB task, a priori power analysis estimated that a sample size of 34 was needed for sufficient power ($\beta = 0.80$) in order to detect a medium effect-size [Cohen's $D = 0.5$, α (two-tailed) = 0.05]. Since, before the experiment, we asked participants to fill out the Waytz scale, namely an intentionality subscale of the Waytz questionnaire (Waytz et al., 2010), we performed a priori power analysis investigating the relationship between the Waytz scale and the IB effect. It estimated that a sample size of 29 was needed for sufficient power ($\beta = 0.80$) to detect a large effect-size [$\rho = 0.5$, α (two-tailed) = 0.05]. The study has been conducted under the ethical standards laid down in the 2013 Declaration of Helsinki and approved by the local ethical committee (Comitato Etico Regione Liguria). All participants gave written informed consent before the experiment. They received an honorarium of 10 € per hour. At the end of the experiment, participants were debriefed about the purpose of the study.

Apparatus and Stimuli. The experimental setup consisted of a mobile Android Device in which the standard Cozmo application with 'SDK Enabled Option' run; one computer connected through the Android Debug Bridge; one 21' inches screen (refresh rate: 60 Hz, resolution: 1920x1080 pixels) to display the task; one keyboard and a customized one-key button. We used a customized button attached to the top of Cozmo's cube for two main reasons. Firstly, the robot could move independently toward the cube and tap it, which was not possible to be done with a standard keyboard due to the physical and mechanical constraints of the robot. Secondly, we wanted to ensure that Cozmo and participants' taps/keypress were collected with the same timing, which was not possible to be reached using just the Cozmo's cube due to a delay arising from the integration

of several components. Stimuli presentation, response collection, and the Cozmo robot were controlled with PsychoPy v.3.0.6 (see Ciardo, De Tommaso, Beyer, & Wykowska, 2018 for a similar procedure of how to integrate Cozmo).

Procedure. Before the experiment, participants filled out the Waytz scale (Waytz et al., 2010), which includes items related to the general tendency to attribute intentionality to robots. Subsequently, they performed the Intentional Binding (IB) task, both alone (Individual Context) and with the Cozmo robot (Social Context, see Khalighinejad et al., 2016 for a similar manipulation). The experiment has been designed to be a full factorial randomized study, with both Context (Individual, Social) and Block type (Baseline, Operant) being administered block-wise, with the order of blocks randomized.

Participants were seated approximately 70 cm away from the computer screen, which leaned in a horizontal position on the desk. Cozmo was placed between the participants and the screen, allowing them to have good visibility of both Cozmo performing the task and the screen. A keyboard was placed in front of participants, and a Cozmo's cube with the adapted one-key button on the top was placed in front of the robot. A mirror was positioned on the other side of the screen, letting participants see Cozmo acting from a frontal perspective as well (see **Figure 1**).

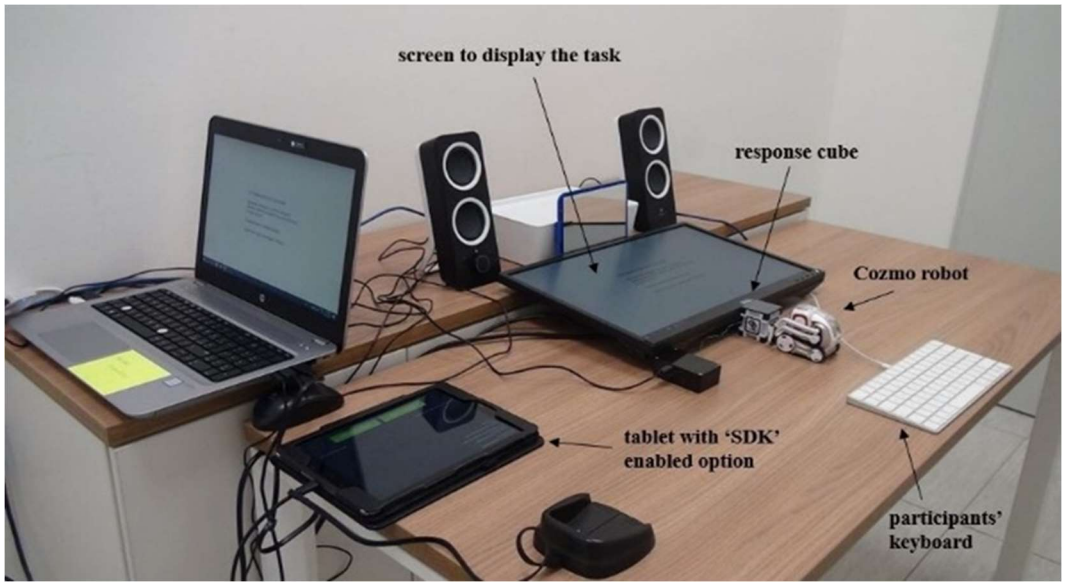

Figure 1. Experimental setup.

The task consisted of two contexts (Individual, Social) with four different types of blocks each: two Baseline blocks (Baseline Action and Baseline Tone), when a single event, either action or tone, occurred, and two Operant blocks (Operant Action and Operant Tone), when both events occurred:

- Baseline Action: the participant or Cozmo performed a keypress at any moment while the clock hand was rotating. No tone was subsequently played. Participants' task was to report at which point in time indicated on the clock the action occurred.
- Baseline Tone: a tone (440 Hz, 100 ms) was played at a random time while the clock hand was rotating. Participants' task was to report at which point in time indicated on the clock the tone occurred. No keypress was required by either the participant or Cozmo.
- Operant Action: the participant or Cozmo performed a keypress at any moment during the clock hand rotations. The tone was played 250 ms after the keypress, while the clock hand

was rotating. Participants' task was to report at which point in time indicated by the clock hand the action occurred, regardless of when they heard the tone.

- Operant Tone: the participant or Cozmo performed a keypress at any moment during the clock hand rotations. The tone was played 250 ms after the keypress. Participants' task was to report at which point in time indicated on the clock the tone occurred, regardless of when the action was executed.

In the Individual Context, participants were executing the task while Cozmo was resting, whereas in the Social Context the Cozmo robot was active and executed the action when required (see Video 1). At the beginning of the Individual blocks, Cozmo moved away from its cube and entered into a "sleep mode" (i.e., eyes closed, snoring). Therefore, in the Individual blocks, the critical event to judge was always the occurrence of a self-generated keypress or the subsequent outcome (i.e., action or tone event, respectively). In contrast, at the beginning of Social blocks, Cozmo opened its eyes and moved toward the cube. When in charge of acting, the robot was programmed to randomly tap the cube's surface during the clock hand rotations (see video "**Publication II_ Experiment1**":https://osf.io/23jmt/?view_only=f58dfc2c426f45ba93a7eff5f931c43f).

Participants were also informed that a red LED light appeared on Cozmo's back while it was performing the physical tapping. In these blocks, the critical event to judge was the occurrence of the robot-generated action or its subsequent outcome (i.e., the tone). Specifically, when the event to judge was Cozmo's tapping, we asked participants to focus on the onset of the movement. The task was designed in a way that, in Social blocks, Cozmo acted only in 90% of trials, and participants were thus instructed that they had to press their keyboard if Cozmo did not act within 10 full rotations; otherwise, from a starting amount of 120 points, they would lose 10 points for each missing response. This was made to ensure that participants attended Cozmo's performance

during the task. Consequently, in those trials in which Cozmo did not act, participants had to judge the occurrence of a self-generated event (either action or tone). Before the beginning of the task, the experimenter showed participants some functionalities of Cozmo, which moved around while making little sounds in front of them for several minutes. This was made to allow participants to understand the actual capacities of the robot, and it was unrelated to the IB task. Each trial started with a black fixation dot on a white background for 1000 ms, followed by the image of the clock (10.6° visual angle) with a red clock hand (length = 135 pixels) presented randomly on one of the 12 five-minute positions of the clock. After 500 ms, the clock hand started to rotate clockwise. To complete a full rotation, the clock hand took 2560 ms; notably, the clock hand stopped rotating randomly between 1500 and 2500 ms after the critical event (either keypress or tone). At the end of each trial, participants were asked to report the position of the clock hand of the event of interest (see **Figure 2**).

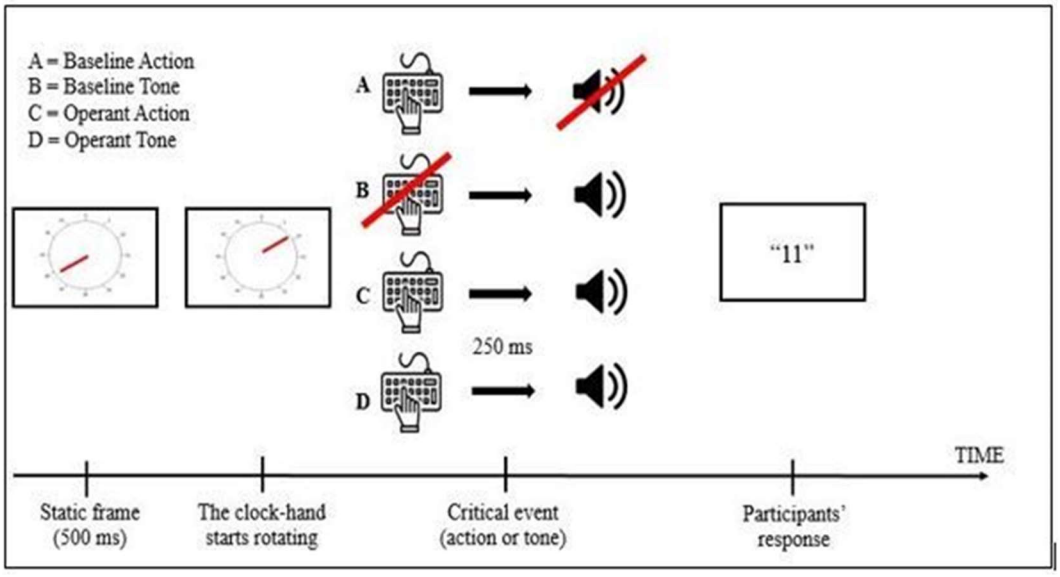

Figure 2. Trial sequence. Participants were instructed to observe the rotating clock hand displayed on the screen and to report its position at the occurrence of the critical event (action or tone). Note that, in Baseline blocks, only one event occurred, either action (A) or tone (B); in Operant blocks, keypress was always followed by a tone 250 ms thereafter, and participants had to judge the position of the clock hand when the event of interest occurred (C: action, D: tone).

When the action was required, participants were instructed to avoid responding in a stereotyped way, or predefined moments of the rotations. Moreover, they were asked to respond always after the first rotation was fully completed. Similarly, in Baseline Tone blocks, the tone (i.e., the critical event) was randomly played always after one full rotation was completed.

The task comprised eight blocks of 40 trials each, with 320 trials in total. Blocks were randomly assigned to either Individual (4 blocks) or Social Context (4 blocks). A practice session of the entire task (i.e., eight trials, one for each combination of Block type and Context) was administered before the task.

2.2.4.1.2. Statistical analyses

For each trial, we estimated the Judgment Error (JE) as the difference between the position of the clock hand reported by participants and the actual onset of the event (i.e., action or tone). A negative JE was interpreted as anticipatory awareness of the event (i.e., the event was perceived as happening earlier in time than it occurred), whereas a positive JE was interpreted as delayed awareness (i.e., the event was perceived as happening later in time than it occurred). Then, "minute" differences were transformed into "millisecond" differences (minute difference x 2560 ms/60). For each Block type (Baseline, Operant), we calculated the average JEs, including both negative and positive JEs, and their standard deviations. JEs that deviated more than ± 2.5 SD from the participants' mean in each Block type were considered as outliers and removed from the analyses (2.69% of the administered trials, mean JE = -22.99, SD = 672.98). Then, JEs were analyzed separately according to the critical event to be judged (i.e., action or tone). It is important to underline that, for the Social Context, we analyzed only trials in which Cozmo responded. The difference between JEs in the Baseline and JEs in the corresponding Operant Block has been defined as the IB effect (Ruess, Thomaschke, & Kiesel, 2018). Our analyses, however, have been

conducted on JEs, not on the IB effect as the dependent variable. Yet, when there was an effect of Block Type (Baseline vs. Operant) on the JEs, we refer to it as the IB effect.

Notably, the directionality of the IB effect has been demonstrated to vary according to the critical event to judge (Haggard et al., 2002). Typically, in a Baseline block (with only action or only tone occurring), when people are asked to judge the occurrence of an action (**Figure 3A**), they tend to underestimate the point in time when it occurred, reporting that the action happened earlier than it did (i.e., JE < 0, see **Figure 3B**). On the contrary, when they are asked to judge the occurrence of the tone (**Figure 3D**) they tend to overestimate the time point when it occurred, reporting that the tone happened later than it did (i.e., JE > 0, see **Figure 3E**). In the case of Operant blocks (i.e., when both action and tone events are present), when the critical event is the action, it is bound to the subsequent tone event. The result is that the subjective time point of occurrence is shifted toward the direction of the following tone, leading to a smaller ***underestimation*** compared to when the action occurred alone (see **Figure 3C**). When the event to be judged is the tone, it is bound to the preceding action. The result is that the subjective point of occurrence is shifted toward the direction of causing action, leading to a smaller ***overestimation*** compared to when the tone occurred alone (**Figure 3F**).

Therefore, when the critical event to judge is the occurrence of action, the IB effect is described as a smaller underestimation (i.e., less negative JEs) for Operant compared to Baseline block. Conversely, when the critical event to judge is the occurrence of tone, the direction of the IB effect is reversed, with a smaller overestimation (i.e., less positive JEs) for Operant compared to Baseline block (Haggard et al., 2002; Obhi & Hall, 2011).

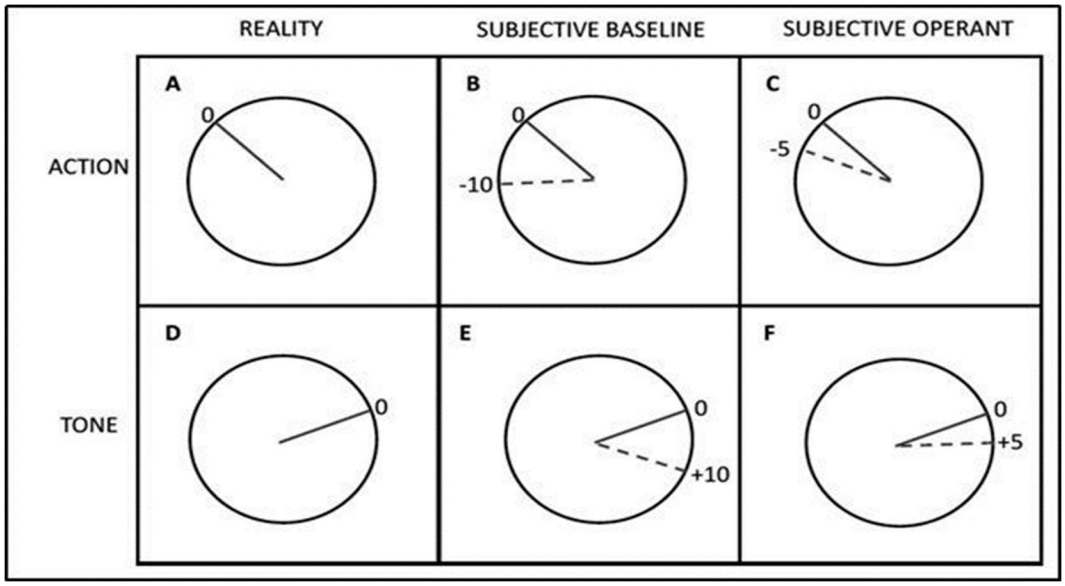

Figure 3. Visual representation of the over- and under-estimation of time of occurrence of tone and action events in an Intentional Binding task with a schematic representation of the clock stimulus, together with the clock hands. **Upper row**: action events. **Lower row**: tone events. **A**: time of occurrence of action event in reality (time point 0). **B**: Baseline block with only one event taking place: time of occurrence of action event in reality (solid line, time point 0) and subjective ***under-estimate*** of the time of occurrence (dashed line, time point -10). **C**: Operant block with two events, time of occurrence of action event in reality (solid line, time point 0) and subjective ***under-estimate*** of the time of occurrence (dashed line, time point -5). **D**: time of occurrence of tone event in reality (time point 0). **E**: Baseline block with only one event taking place: time of occurrence of tone event in reality (solid line, time point 0) and subjective ***over-estimate*** of the time of occurrence (dashed line, time point +10). **F**: Operant block with two events, time of occurrence of tone event in reality (solid line, time point 0) and subjective ***over-estimate*** of the time of occurrence (dashed line, time point +5). The depicted numbers are shown only for illustration of the directionality of the JEs and do not represent the actual sizes of the effects.

To investigate the effect of Context and Block type on participants' JEs, we run linear mixed-effect models, with JEs being modeled as a function of both Context (Individual, Social) and Block type (Baseline, Operant), plus their interactions, as fixed effects, and participants as a random effect. Notably, we run two identical models, one for action and one for tone blocks, separately:

mAction: JEs ~ Context * Block type, random = Participant

mTone: JEs ~ Context * Block type, random = Participant

Analyses were conducted by using the lme4 package (Bates, Mächler, Bolker, & Walker, 2014) in R v.3.0.6. (R Core Team, 2013). Parameters estimated (β) and their associated t-tests (t, p-value)

were calculated using the Satterthwaite approximation method for degrees of freedom (Kuznetsova, Brockhoff, & Christensen, 2017); they were reported with the corresponding bootstrapped 95% confidence intervals (Efron & Tibshirani, 1994). Following two-way significant interaction, pairwise comparisons were performed with the 'emmeans' package in R studio (Lenth, 2019). It computes contrasts with marginal means for factors combination in a variety of models (including linear mixed-effects models) and compares different slopes (beta estimates). Tukey correction has been applied.

To simplify reading the results section, we point out that, in general, the significant difference in JEs between Baseline and the corresponding Operant block (i.e., a main effect of Block) was defined as the IB effect; however, we further discuss in detail specific directions of the IB effect.

2.2.4.1.3. Results

Action events. When the critical event was the action, results showed a main effect of Context [β = -26.06, SE = 3.57, $t_{(5080.64)}$ = -7.29, $p < 0.0001$, CI = (-33.07; -19.05)]. Specifically, underestimation was smaller (i.e., less negative JEs) in Individual than in Social Context [β = 39.1, SE = 2.55, $t_{(5082)}$ = 15.29, $p < 0.0001$, CI = (34; 44.1); ($M_{Individual}$ = -39.1, $SE_{Individual}$ = 8.76; M_{Social} = -78.1, SE_{Social} = 8.79)]. Moreover, a main effect of Block type emerged [β = 27.69, SE = 3.43, $t_{(5080.28)}$ = 8.06, $p < 0.0001$, CI = (20.96; 34.42)], with significantly smaller underestimation (i.e., less negative JEs) in Operant than in the corresponding Baseline block [β = -14.7, SE = 2.55, $t_{(5081)}$ = -5.77, $p < 0.0001$, CI = (-19.7; -9.71); ($M_{Operant}$= -51.2, $SE_{Operant}$ = 8.78; $M_{Baseline}$= -65.9, $SE_{Baseline}$= 8.77)]. Finally, a significant Context * Block type interaction emerged [β = -25.98, SE = 5.08, t = -5.1, $p < 0.0001$, CI = (-35.95; -16)]. Given the significant two-way interaction, we further investigated the contrast between JEs in Baseline and in the corresponding Operant block (i.e., the IB effect) between Individual and Social context with pairwise comparisons (Tukey's

HSD correction for multiple comparisons). They showed that, for the Individual Context, the underestimation for JEs was significantly smaller in Operant compared to Baseline block, thereby indicating an IB effect [β = -27.69, SE = 3.43, $t_{(5080)}$ = -8.06, $p < 0.0001$, CI = (-36.5; -18.87); ($M_{Operant}$ = -25.2, $SE_{Operant}$ = 8.93; $M_{Baseline}$ = -52.9, $SE_{Baseline}$ = 8.92)]. However, this was not true for the Social Context, where no significant difference emerged between JEs in Operant and the corresponding Baseline block [β = -1.71, SE = 3.76, $t_{(5080)}$ = -0.45, p = 0.96, CI = (-11.4; 7.95); ($M_{Operant}$ = - 77.3, $SE_{Operant}$ = 9; $M_{Baseline}$ = -79, $SE_{Baseline}$ = 8.98)] (see **Figure 4**).

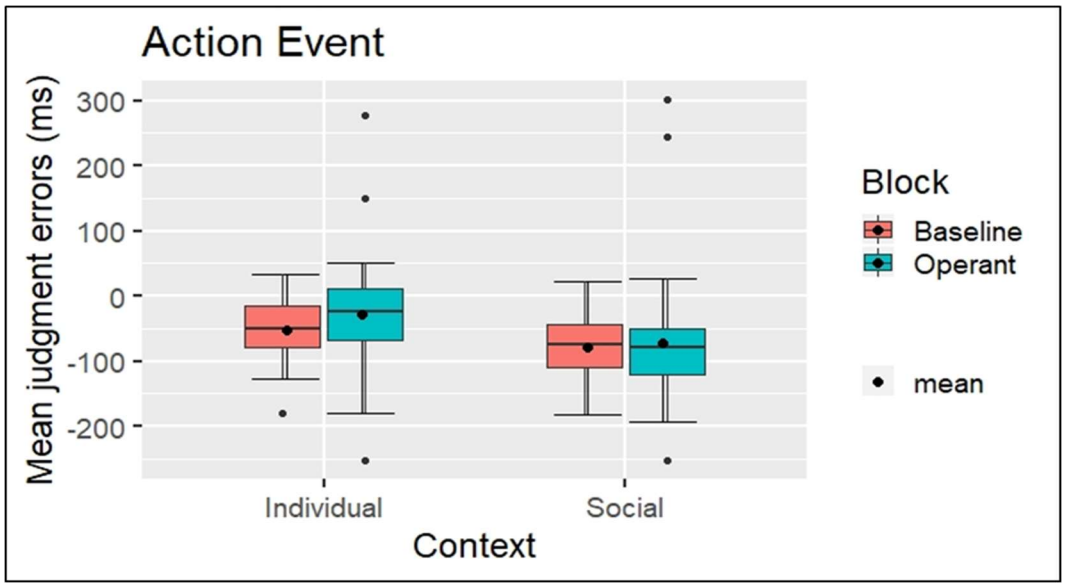

Figure 4. Experiment 1: Mean JEs plotted as a function of Context (Individual, Social) and Block type (Baseline, Operant) when the critical event to judge was the action. Error bars represent standard errors.

Tone events. When the critical event was the tone, results showed a significant main effect of Context [β = -8.99, SE = 4.39, t (5267.05) = -2.04, p = 0.04, CI = (-17.59; -0.38)]. Specifically, overestimation was larger (i.e., more positive JEs) in Individual compared to Social Context [β = 11.1, SE = 3.23, t (5268) = 3.42, p = 0.0006, CI = (4.73; 17.4); (M Individual = 68.3, SE Individual = 11.5; M Social = 57.3, SE Social = 11.5)]. Moreover, a significant main effect of Block type emerged [β = -61.79, SE = 4.41, t (5267.58) = -13.99, $p < 0.0001$, CI = (-70.44; -53.14)].

Specifically, overestimation was smaller (i.e., less positive JEs) in Operant compared to the corresponding Baseline block [β = 63.9, SE = 3.23, t (5268) = 19.79, p < 0.0001, CI = (57.5; 70.2); (M Operant = 30.9, SE Operant = 11.5; M Baseline = 94.7, SE Baseline = 11.5)]. The Context * Block type interaction was not significant [β = -4.14, SE = 6.45, t (5268.07) = -0.64, p = 0.52, CI = (-16.79; 8.5)], indicating the IB effect for both Individual and Social contexts (see **Figure 5**).

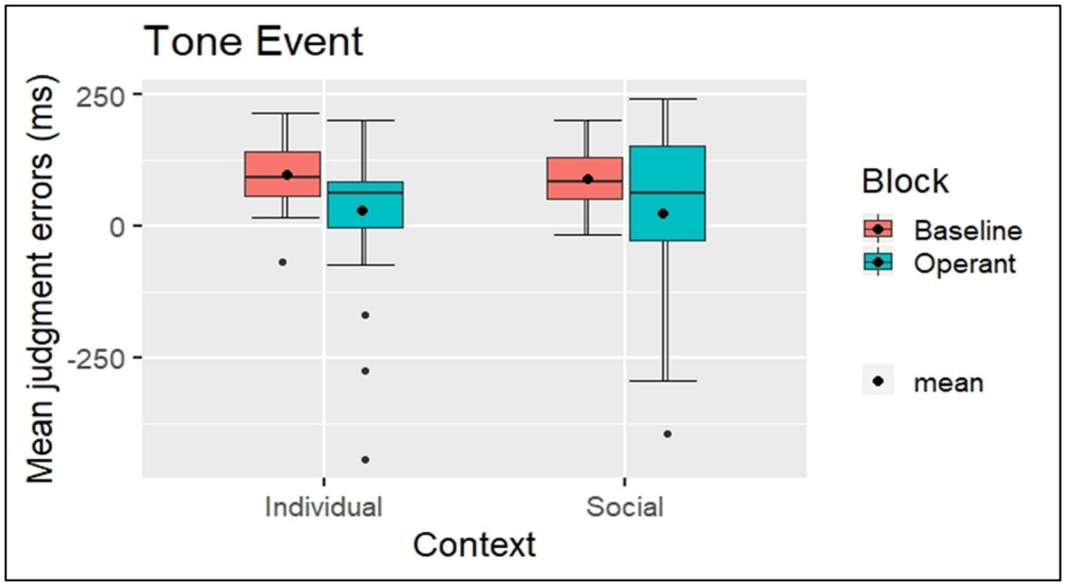

Figure 5. Experiment 1: Mean JEs plotted as a function of Context (Individual, Social) and Block type (Baseline, Operant) when the critical event to judge was the tone. Error bars represent standard errors.

2.2.4.2. Experiment 2

2.2.4.2.1. Materials and Methods

Participants. Thirty-six new participants took part in Experiment 2 (range: 18-40 years old, M age = 25, SD age ± 3.3, 11 males, 5 left-handed, 1 ambidextrous). All participants gave written informed consent and the study was conducted under the ethical protocol applied also in Experiment 1. The sample size was estimated as for Experiment 1.

Apparatus, Stimuli, and Procedure. The apparatus, stimuli, and procedure were the same as in Experiment 1, with the only exception that participants were instructed that in Social trials Cozmo performed the task by sending a command via Bluetooth. Thus, in the Social Context, Cozmo did not execute any embodied, perceivable physical action; otherwise, it was programmed to make a squeaking sound, indicating that it was sending a command to the keyboard, and, for participants, that the robot was executing the "digital" action (see video "**Publication II_ Experiment2**":https://osf.io/23jmt/?view_only=f58dfc2c426f45ba93a7eff5f931c43f). Different from Experiment 1, in the action blocks of the Social Context, participants were required to report the position of the clock hand when the digital action occurred (i.e., when they heard the onset of the squeaking sound). Notably, as in Experiment 1 participants saw a red LED light appearing on Cozmo's back while performing the digital action. Again as in Experiment 1, we asked participants to focus on the onset of the "digital action" (i.e., the squeaking sound). As in the previous experiment, the digital action was followed by a sensory outcome (i.e., a tone) 250 ms thereafter. Thus, the cause-effect relationship was not affected by the manipulation of the type of action that Cozmo performed across experiments.

2.2.4.2.2. Statistical analyses

For each trial, a Judgment Error (JE) was calculated with the same procedure as in Experiment 1. JEs that deviated more than ± 2.5 SD from the participants' mean in each Block type were considered as outliers and removed from the analyses (2.11 % of the administered trials, mean JE = -90.87, SD = 777.21). As in Experiment 1, we run linear mixed-effects models with JEs being modeled as a function of both Context and Block type, plus their interactions, as fixed effects, and participants as a random effect, in a separate way for each critical event (i.e., action and tone). When the critical event was the tone, one participant was excluded from the analyses due to a very

low number of valid trials in the Individual Context (< 10; 9 trials for the Baseline, 5 trials for the Operant block) upon outliers removal. Following two-way significant interactions, pairwise comparisons were performed with the 'emmeans' package in R studio (Lenth, 2019).

2.2.4.2.3. Results

Action events. When the critical event was the action, results showed a main effect of Context [β = 324.21, SE = 4.71, t (5282.16) = 68.75, p < 0.0001, CI = (314.97; 333.45)]. Specifically, the underestimation was larger (i.e., more negative JEs) in Individual compared to Social Context [β = -274, SE = 3.34, t (5282) = -82.26, p < 0.0001, CI = (-281; -268); (M Individual = -23.3, SE Individual = 11.8; M Social = 251.1, SE Social = 11.8)]. Moreover, a significant main effect of Block type emerged [β = 46.97, SE = 4.56, t (5282.08) = 10.3, p < 0.0001, CI = (38.04; 55.91)]; however, the difference between JEs in Baseline and in the corresponding Operant block resulted not to be significant [β = 2.88, SE = 3.33, t (5282) = 0.86, p = 0.38, CI = (-3.66; 9.41); (M Operant = 112, SE Operant = 11.8; M Baseline = 115, SE Baseline = 11.8)]. Finally, the Context * Block type interaction was significant [β = -99.7, SE = 6.66, t (5282.07) = -14.95, p < 0.0001, CI = (-112.77; -86.63)]. Therefore, pairwise comparisons (Tukey's HSD correction for multiple comparisons) were run further to investigate the contrast between JEs in Baseline and in the corresponding Operant block (i.e., the IB effect) for each Context separately. Results showed a significant difference in the Individual Context, with a smaller underestimation (i.e., less negative JEs) in Operant compared to the Baseline block, namely the IB effect [β = -47, SE = 4.56, t (5282) = -10.3, p < 0.0001, CI = (-59; -34.9); (M Operant = 0.17, SE Operant = 12; M Baseline = -46.8, SE Baseline = 12)]. A significant difference emerged also for the Social Context [β = 52.7, SE = 4.87, t (5282) = 10.83, p < 0.0001, CI = (39.9; 65.6)]. Specifically, the overestimation was smaller (i.e., less positive JEs) in Operant compared to Baseline block, thereby indicating a reversed IB

effect [β = 52.7, SE = 4.87, t (5282) = 10.83, p < 0.0001, CI = (39.9; 65.6); (M Operant = 224.68, SE Operant = 12.1; M Baseline = 277.41, SE Baseline = 12.1)] (see **Figure 6**).

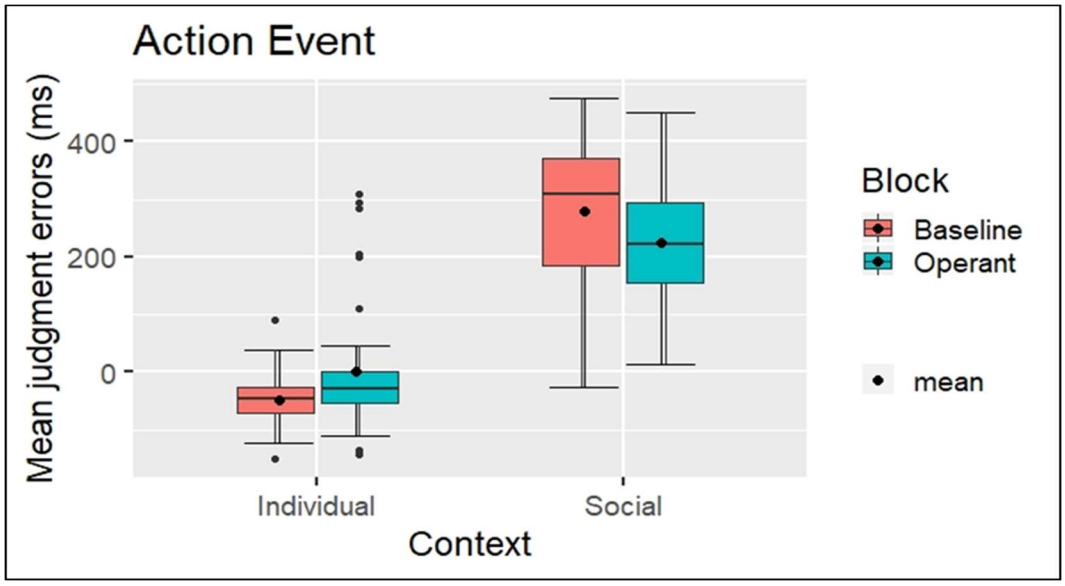

Figure 6. Experiment 2: Mean JEs plotted as a function of Context (Individual, Social) and Block type (Baseline, Operant) when the critical event to judge was the action. Error bars represent standard errors.

Tone events. When the critical event to judge was the tone, results showed a significant main effect of Context [β = -10.41, SE = 3.78, t (5322.01) = -2.75, p = 0.005, CI = (-17.82; -3)]. Specifically, overestimation was smaller (i.e., less positive JEs) in Individual compared to Social Context [β = -17.7, SE = 2.72, t (5322) = -6.51, p < 0.0001, CI = (-23; -12.4); (M Individual = 59.6, SE Individual = 9.17; M Social = 77.3, SE Social = 9.18)]. Moreover, a significant main effect of Block type emerged [β = -49.33, SE = 3.78, t (5322.01) = -13.04, p < 0.0001, CI = (-56.74; -41.92)]. Specifically, the overestimation was smaller (i.e., less positive JEs) in Operant compared to the corresponding Baseline block [β = 21.2, SE = 2.72, t (5322) = 7.79, p < 0.0001, CI = (15.9; -26.5); (M Operant = 57.8, SE Operant = 9.18; M Baseline = 79, SE Baseline = 9.17)]. Finally, the Context * Block type interaction was significant [β = 56.26, SE = 5.43, t (5322.09) =

10.34, p < 0.0001, CI = (45.6; 66.91)]. Pairwise comparisons (Tukey's HSD correction for multiple comparisons) were run further to investigate the contrast between JEs in Baseline and in the corresponding Operant block (i.e., the IB effect), for each Context separately. They revealed a significant difference for the Individual Context, with a smaller overestimation (i.e., less positive JEs) in Operant compared to the Baseline block, thereby indicating an IB effect [β = 49.33, SE = 3.78, t (5322) = 13.04, p < 0.0001, CI = (39.35; 59.31); (M Operant = 34.9, SE Operant = 9.36; M Baseline = 84.2, SE Baseline = 9.36)]. However, this was not true for the Social Context, where no significant differences emerged between JEs in Operant and in the corresponding Baseline block [β = -6.93, SE = 3.91, t (5322) = -1.77, p = 0.45, CI = (-17.2; 3.38); (M Operant = 80.7, SE Operant = 9.41; M Baseline = 73.8, SE Baseline = 9.36)] (see **Figure 7**).

Figure 7. Experiment 2: Mean JEs plotted as a function of Context (Individual, Social) and Block type (Baseline, Operant) when the critical event to judge was the tone. Error bars represent standard errors.

2.2.4.3. Discussion of the role of action representation for vicarious SoA in HRI

The first aim of the study was to determine the role of action representation in the emergence of vicarious SoA in HRI. To this end, we evaluated the IB effect in two experiments. In both experiments, participants performed the task alone (Individual Context) or with the Cozmo robot (Social Context). Across experiments, the Cozmo robot executed the action in two different ways. In Experiment 1, we programmed the robot to execute a physical, perceivable keypress, whereas, in Experiment 2, participants were instructed that the robot sent a command to the computer via Bluetooth (i.e., a non-embodied digital action). We programmed the robot to produce a squeaking sound that was supposed to mark the moment in which the command was sent to the computer for producing the outcome of the "action" (i.e., the tone). It is important to underline that, even if the event generating the tone outcome was different across the two experiments, the temporal contingency between the cause (physical or digital) and its outcome (the tone) was the same.

In each experiment, we first evaluated the IB effect for the Individual Context, to address whether participants experienced control over self-generated actions and outcomes. Although assessing the emergence of self-agency was beyond the scope of this paper, it was considered as a baseline, to ensure that our paradigm was able to elicit the IB effect in the first place. Notably, results showed that, in the Individual Context of both experiments, participants always experienced control over self-generated actions (i.e., action blocks) and their outcomes (i.e., tone blocks).

Then, we focused on the Social Context of both experiments, where results showed a different pattern for action and tone events. For action blocks, results showed that, in Experiment 1, participants did not experience vicarious SoA over physical robot's actions, as indicated by the lack of Social IB effect. However, in Experiment 2, when the robot-generated action was digital, JEs in the Operant block resulted to be underestimated compared to the Baseline block. The latter

suggests not only a lack of the IB effect in the typical direction (i.e., underestimated JEs in Baseline compared to Operant) but also a boosting of the effect in the opposite direction. This pattern of results might suggest that participants perceived the robot's digital action as occurring earlier than it occurred. A similar anticipatory awareness has been found also by Wohlschläger and colleagues (2003) when participants had to judge the occurrence of a machine-generated event, compared to both self- and other human-generated action conditions. As the authors hypothesized that the effect was due to the lack of the hand movement seen in the other conditions, the study was repeated using a rubber hand for the machine-action trials. As a result, the rubber hand reduced the anticipatory awareness but it did not produce the delayed awareness shown by both self- and other human-generated action conditions. Thus, authors suggested that participants perceived the machine-generated action as unintentional, compared to the human-generated ones (Wohlschläger et al., 2003). Interestingly, a similar reversed effect has been previously reported also for human involuntary actions, namely actions triggered by TMS impulses (Haggard et al., 2002). Certainly, our speculation needs to be further investigated, to determine whether actions perceived as unintentional have a common pattern and whether this pattern could be the same for both humans and artificial agents.

When the critical event to be judged was the tone, the vicarious SoA – in the form of vicarious IB effect, namely the difference between JEs in the Operant and the corresponding Baseline block- was observed only when the tone was the outcome of the robot's physical action (Experiment 1). However, when the sensory outcome was generated by the robot's digital, non-embodied and unobservable action (Experiment 2), the vicarious IB effect did not occur (see Supplementary Materials, point SM.1, p.126, for comparisons across experiments).

The lack of the vicarious IB effect in Experiment 2 might suggest that, perhaps, when the causing action is not embodied, and thus it doesn't generate sensory consequences in the environment, participants are less prone to form a causal link between the action and its subsequent outcome. It would not allow individuals to have a strong representation of the robot action, and, therefore, vicarious SoA would not arise.

Together, our results are only partially in line with the action representation account of vicarious SoA. According to the prediction, we found that, when the representation of the cause-effect link is weakened by the lack of an embodied physical action (Experiment 2), vicarious SoA never occurs. On the contrary, when the cause-effect link can be represented thanks to the embodied nature of the action (Experiment 1), vicarious SoA for robot-generated actions occurs, but only with reference to the sensory outcome (i.e., the tone), and not for the causal action (i.e., the keypress). Such a result might be explained in the light of the dissociation between action and outcome events, which have been demonstrated to be two distinct events in relation to implicit SoA. Indeed, repetitive TMS stimulation over pre-SMA resulted in disrupting the IB effect for self-generated actions but not for self-generated outcomes (Zapparoli et al., 2020). In Experiment 1, Cozmo acted through a keypress, in a similar way as the human participants. It might have allowed participants to represent the cause-effect link of Cozmo's actions, in a similar way as one's actions. However, the embodied physical action was executed by Cozmo and participants with different effectors (lift vs. hand, respectively). Thus, it might be that participants formed a less accurate representation of the robot action, and, consequently, vicarious SoA did not occur for the action event, despite it emerged for the tone event.

2.2.5. The role of adopting Intentional Stance for vicarious SoA in HRI

The second aim of the study was to determine the role of the adoption of the Intentional Stance in the occurrence of vicarious SoA. To this end, we evaluated the individual differences in attribution of intentionality toward robots before participants performed the task. We administered the intentionality subscale from Waytz et al. (2010) as a pre-task questionnaire. This scale measures whether people attribute to robots cognitive states that are considered uniquely human, such as intentions. We predicted that if adopting Intentional Stance plays a role in vicarious SoA, then the Waytz score should predict the magnitude of the IB effect in the Social Context. Specifically, a higher intentionality score should predict larger IB effects in the Social Context.

2.2.5.1. Statistical analyses

For the analyses related to the role of intentionality attribution in vicarious SoA, we focused only on the Social trials of both experiments analyzed together. We conducted two identical linear mixed-effect models, one for action and one for tone separately. To assess whether the degree of attribution of the intentionality (i.e., Waytz score) was predictive of the magnitude of the vicarious IB effect (namely, the difference in JEs between Operant and the corresponding Baseline block) and whether the relationship between vicarious IB and Waytz score changed across experiments, JEs were modeled as a function of Block type (Baseline, Operant), Waytz score, and Experiment (1, 2), plus their interactions as fixed effects; and participants as a random effect (see Supplementary Materials, point SM.2, p. 129, for more details). The Waytz score was calculated with reference to the paper of Ruijten and colleagues (2019), who employed the same reduced 7-items version of the scale that we decided to use.

2.2.5.2. Results

Action events. When the critical event was the action, results showed a significant main effect of Block type [β = -60.45, SE = 13.85, $t_{(4742.68)}$ = -4.36, $p < 0.0001$, CI = (-87.59; -33.31)]. Specifically, overestimation was smaller (i.e., less positive JEs) in Operant compared to the corresponding Baseline block [β = 27, SE = 3.17, $t_{(4742)}$ = 8.49, $p < 0.0001$, CI = (20.8; 33.2); ($M_{Operant}$ = 71.8, $SE_{Operant}$ = 10.9; $M_{Baseline}$ = 98.7, $SE_{Baseline}$ = 10.9)]. Moreover, a significant main effect of Experiment emerged [β = 414.48, SE = 65.33, $t_{(70.99)}$ = 6.34, $p < 0.0001$, CI = (288.35; 540.6)]. Specifically, the underestimation was larger (i.e., more negative JEs) for Experiment 1 compared to Experiment 2 [β = -329, SE = 21.5, $t_{(68)}$ = -15.31, $p < 0.0001$, CI = (-372; -286); ($M_{Experiment\ 1}$ = -79.5, $SE_{Experiment\ 1}$ = 15.2; $M_{Experiment\ 2}$ = 250, $SE_{Experiment\ 2}$ = 15.2)]. The Block type * Waytz score interaction resulted to be significant [β = 20.15, SE = 4.22, $t_{(4743.01)}$ = 4.76, $p < 0.0001$, CI = (11.87; 28.43)], as well as the Block type * Waytz score * Experiment interaction [β = -25.27, SE = 6.1, $t_{(4742.68)}$ = -4.14, $p < 0.0001$, CI = (-37.22; -13.31)]. To address the three-way interaction we run two separate mixed-effects models according to the Experiment (1, 2), with JEs in Social Context being modeled as a function of Block type and Waytz score, plus their interactions, as fixed effects, and participants as the random effect. Results showed that, in Experiment 1, a main effect of Block type emerged [β = -60.38, SE = 12.21, $t_{(2292.37)}$ = -4.94, $p < 0.0001$, CI = (-84.3; -36.4)] as well as a significant Block type * Waytz score interaction [β = 20.12, SE = 3.72, $t_{(2292.64)}$ = 5.4, $p < 0.0001$, CI = (12.81; 27.42)]. Specifically, the Waytz score predicted the JEs in Operant blocks [β = -12.85, SE = 4.02, $t_{(2378)}$ = -3.19, p = 0.001, CI = (-20.75; -4.95)] but only marginally in Baseline blocks [β = 5.15, SE = 2.66, $t_{(1188)}$ = 1.93, p = 0.053, CI = (-0.07; 10.38)]. In Experiment 2, only the main effect of Block type emerged [β = -38.2, SE = 14.62, $t_{(2450.16)}$ = - 2.61, p = 0.009,

CI = (-66.87; -9.54)] but not a Block type * Waytz score interaction [β = -5.12, SE = 4.83, $t_{(2450.14)}$ = -1.05, p = 0.28, CI = (-14.59; -4.35)] (see **Figure 8**).

Figure 8. Mean JEs plotted as a function of Block type (Baseline, Operant), Waytz score, and Experiment (1, 2) when the critical event to be judged was the action.

Tone events. When the critical event was the tone, results showed a main effect of Block type [β = -96.84, SE = 13.9, $t_{(5040.71)}$ = -6.96, p < 0.0001, CI = (-124.12; -69.61)]. Specifically, the overestimation was smaller (i.e., less positive JEs) in Operant compared to Baseline block [β = 29.1, SE = 3.12, $t_{(5035)}$ = 9.33, p < 0.0001, CI = (23; 35.2); ($M_{Operant}$ = 52.4, $SE_{Operant}$ = 8.87; $M_{Baseline}$ = 81.5, $SE_{Baseline}$ = 8.81)]. The Block type * Experiment interaction was significant [β = 69.66, SE = 19.75, $t_{(5037.25)}$ = 3.52, p = 0.0004, CI = (30.97; 108.39)]. In line with results of Experiment 1 and 2, pairwise comparisons (Tukey's HSD corrected) revealed a significant IB effect in Experiment 1 (i.e., significantly underestimated JEs in Operant compared to the Baseline block) [β = 66.75, SE = 4.47, $t_{(5036.6)}$ = 14.92, p < 0.0001, CI = (65.48; 114.8); ($M_{Operant}$ = 23.4, $SE_{Operant}$ = 12.5; $M_{Baseline}$ = 90.1, $SE_{Baseline}$ = 12.4)]. Conversely, no evidence of an IB effect were

found for Experiment 2 [β = -8.52, SE = 4.35, $t_{(5032.4)}$ = -1.96, p = 0.2, CI = (-19.7; 2.65); ($M_{Operant}$ = 81.3, $SE_{Operant}$ = 12.6; $M_{Baseline}$ = 72.8, $SE_{Baseline}$ = 12.5)]. Moreover, a Block type * Waytz score interaction was significant [β = 10.02, SE = 4.27, $t_{(5043.34)}$ = 2.34, p = 0.01, CI = (1.64; 18.4)]. Specifically, JEs were marginally predicted by the Waytz score in the Operant blocks [β = 6.76, SE = 3.62, $t_{(2315)}$ = 1.86, p = 0.06, CI = (-0.34; 13.86)] but not in Baseline blocks [β = -1.84, SE = 1.58 $t_{(2788)}$ = -1.2, p = 0.24, CI = (-4.95; 1.27)]. The Block type * Waytz score * Experiment interaction resulted not to be significant [β = 1.86, SE= 6.25, $t_{(5038.15)}$ = 0.29, p = 0.76, CI = (-10.4; 14.12)] (see **Figure 9**).

Figure 9. Mean JEs plotted as a function of Block type (Baseline, Operant) and Waytz score when the critical event was the tone.

2.2.5.3. Discussion of the role of adopting Intentional Stance for vicarious SoA in HRI

The second aim of the study was to determine the role of the adoption of the Intentional Stance in vicarious SoA. To this end, we investigated whether individual differences in the Waytz score predicted the occurrence of the IB effect in Social contexts. When the critical event was the action,

the Waytz score resulted to be predictive of the magnitude of the vicarious IB effect in Experiment 1, when Cozmo physically tapped the cube, but not in Experiment 2, when Cozmo's action was digital. This is an interesting result, as it sheds new light on the lack of vicarious SoA reported for Cozmo's actions in Experiment 1, c.f. paragraph 3.1.3. It would suggest that the attribution of intentionality plays a role in vicarious SoA for robot-generated actions. Indeed, the effect of the Waytz score emerged only in blocks when both events (action and tone) were present (i.e., Operant blocks), thereby allowing to form a cause-effect link between the causing action and the subsequent outcome. As an explanation, we may speculate that the attribution of intentionality acts as a reinforcement of this link, leading participants to perceive the robot's action as perceptually and temporally linked to the following tone. It would be in line with the Waytz score being not predictive of JEs in Social Context for Baseline blocks, where only one event is present (i.e., robot's action) and the lack of the following tone may hinder the formation of a cause-effect link. When participants were asked to judge the outcome (i.e., the tone), the relationship between vicarious IB and Waytz score appeared to be reversed. Specifically, results showed that, independent of the nature of the causing action (physical or digital), the more participants adopted the Intentional Stance toward robots the smaller was the IB effect reported in the Social Context. This new and intriguing result could be interpreted in the context of diffusion of responsibility (Bandura, 1991). Several studies showed that outcome monitoring is reduced when in the social context, leading to a generally lower SoA for self-generated action (Beyer et al., 2017; 2018). Specifically, it has been proposed that, in the presence of an intentional agent, mentalizing processes interfere with action selection and weaken the action-outcome link, resulting in a reduced SoA (Ciardo, Beyer, De Tommaso, & Wykowska 2020). In our study, this is supported by the significant relationship between Waytz score and JEs in Operant blocks; in other words, the

more participants attributed intentionality to robots the more they were accurate in estimating the occurrence of robots' actions (i.e., smaller differences in JEs between Baseline and Operant blocks). Thus, the more participants attribute intentionality toward robots the more they would perceive the outcome as disjointed from the action; consequently, they would perceive the tone outcome as an external event without any preceding cause, thereby leading to a reduced implicit SoA.

2.2.6. General Discussion

The present study aimed to investigate what are the contributing factors to vicarious SoA in HRI. Specifically, we addressed the contribution of two potential factors: (1) the possibility of representing the cause-effect link underlying the robot's actions, and (2) the adoption of the Intentional Stance to explain the robot's behavior.

To this purpose, we designed an IB task (Haggard et al., 2002; Strother et al., 2010; Obhi & Hall, 2011) that accounted for the robot's characteristics. Our dependent measure was the Judgment Error (JE), namely the difference between the estimated and actual position of the clock hand when participants judged the occurrence of the critical event (either action or tone). To evaluate the role of the action representations, participants performed the task alone (Individual Context) or with the non-anthropomorphic Cozmo robot (Social Context). Thanks to its embodiment, the robot executed either a physical embodied action (i.e., a keypress, Experiment 1) or a digital non-embodied action (i.e., sending a command by Bluetooth, Experiment 2). Moreover, we assessed the role of the attribution of intentionality toward robots by administering the Waytz scale (Waytz et al., 2010) before the task in both experiments.

When the critical event to judge was the physical action (Experiment 1), vicarious SoA did not emerge. As a possible explanation, we might speculate that, although the physical tapping

performed by the robot would have allowed participants to form an action representation of it, the effector used by Cozmo (i.e., the lift) was too dissimilar to the humans' effector used to perform the same tapping action (i.e., the hand). Thus, even if the representation of Cozmo's actions occurred, one possibility could have been that it was weakened by the fact the robot's effector has a non-anthropomorphic shape (Khalighinejad et al., 2016). If it was the case, people would represent Cozmo's action but the representation would not be sufficiently accurate to elicit vicarious SoA. However, there would be an alternative explanation related to attribution of intentionality. In our study, the attribution of intentionality predicted the magnitude of vicarious SoA (namely, the vicarious IB effect), in such a way that the more participants attributed intentionality to robots, the more they tended to experience vicarious agency over Cozmo's actions. This was not true for Experiment 2, where a reversed vicarious IB effect was found (i.e., underestimated JEs in Operant than in Baseline blocks). An analogous result was found by Wohlschläger and colleagues (2003), who suggested that this is because participants did not adopt Intentional Stance toward machine-generated actions, and it influenced the perception of those events. Therefore, it might be that digital actions performed by Cozmo were most probably perceived as unintentional. It should be noted that the lack of vicarious IB for the action event in Experiment 1 is in contrast with the results of Khalighinejad and colleagues (2016), who reported similar IB effects for an anthropomorphic hand with servo-actuated fingers and a human hand. However, in our study, we used a non-anthropomorphic robot, thus it is plausible that the action representation of Cozmo's keypress did not fully overlap with that of a self-generated action. Thus, attribution of intentionality may have acted by "boosting" the similarity between self- and robot-generated actions, but only when the embodied actions of the robot allowed activating action representation based on the causal link between actions and outcomes.

Taken together, these results suggest that both action representation and attribution of intentionality toward robots play a role for the vicarious SoA to emerge. Specifically, when a non-anthropomorphic robot performs a physical action, ascribing intentionality may be a prerequisite to link contingent events in the action-outcome chain.

When the critical event to judge was the outcome (i.e., the tone), results revealed that vicarious SoA occurred only following a physical action (Experiment 1). In contrast, vicarious SoA did not occur when the action was digital and thus disembodied (Experiment 2). In the case of digital action, the causal link (between action and its sensory consequence) might be difficult to represent. This might hinder the occurrence of vicarious SoA. Notably, it would support the crucial role of the action representation for vicarious SoA to emerge. Such a dissociation in the role of the action representation in IB effect for action and outcome events is in line with recent results from Zapparoli and colleagues (2020), who reported that interfering with the activation of a proper motor plan disrupted the IB effect for action events only, and not for their outcomes. Remarkably, in our study, when participants judged the occurrence of the tone event, the lack of a three-way interaction suggested that the relationship between the magnitude of the vicarious IB effect and the Waytz score did not change across experiments. As a possible explanation, we speculate that this relationship is not modulated by the nature of the robot's causing actions; therefore, further investigations will be needed to clarify the potential role of attribution of intentionality when people are focusing on the outcome, rather than on the action that generated it.

Interestingly, we found that, in both experiments, the more participants ascribed intentionality to robots the less they experienced vicarious SoA over robot's outcomes. Although this relationship shows an opposite direction compared to action blocks, it confirms previous findings showing that people experienced a generally reduced SoA when interacting with a non-anthropomorphic robot

(Ciardo et al., 2018; 2020). It would be in line with the model proposed by Beyer and colleagues (2017; 2018) to explain the reduction of SoA when people are engaged in a shared social context. In other words, the social presence of another agent involves the activation of mentalizing processes, which interfere with action selection processes by making them less fluent. Therefore, when interacting with intentional agents, mentalizing processes about their intentions make it more difficult for people to decide if and when to act, thereby reducing their ability to monitor and process action outcomes. As a consequence, they experience reduced SoA (Beyer et al., 2017; see also Vastano, Ambrosini, Ulloa, & Brass, 2020). Interestingly, the model has received support from recent evidence demonstrating that, at the electrophysiological level, being engaged in a joint task with the Cozmo robot reduces both outcome processing and monitoring (Hinz, Ciardo, & Wykowska, 2021), in line with previous findings investigating the effect of the social context when the co-agent was another human (Beyer et al., 2017; 2018).

2.2.7. Conclusions

In conclusion, through implementing an IB task in HRI, we examined the contribution of both action representation and adopting Intentional Stance to the emergence of vicarious SoA for artificial agents. When action monitoring was required (judgment of occurrence of the action event), vicarious SoA for robot-generated physical actions was predicted by the degree of attributed intentionality. In contrast, for robot's digital actions vicarious SoA never occurred. Such a result suggests that both action representation and attribution of intentionality are necessary but not sufficient to experience vicarious SoA over actions generated by a non-anthropomorphic robot. Conversely, representation of the action and intentionality attribution seems to play a different and independent role for outcome monitoring, with the former being necessary to the emergence of vicarious SoA, and the latter affecting the perceived link between cause and outcome.

Future studies should exploit the role of a robot as a social partner in affecting SoA for both self- and other- action-outcome monitoring, especially considering that, in the near future, robots will share human social spaces (e.g., schools, hospitals, companies). Therefore, it appears crucial to fully understand how their presence affects perception of authorship of action consequences and cognitive processes.

2.2.8. Acknowledgments

The authors are grateful to Dr. Frederike Beyer for her comments on a previous version of the manuscript.

2.2.9. Funding

This work has received support from the European Research Council under the European Union's Horizon 2020 research and innovation program, ERC Starting Grant, G.A. number: 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.

2.2.10. References

Android Debug Bridge. Online available at: cozmosdk.anki.com/docs/adb.html?highlight=adb. Last access: 7/8/2020

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50*(2), 248-287.

Berberian, B., Sarrazin, J. C., Le Blaye, P., & Haggard, P. (2012). Automation technology and sense of control: a window on human agency. *PLoS one, 7*(3), e34075.

Beyer, F., Sidarus, N., Bonicalzi, S., & Haggard, P. (2017). Beyond self-serving bias: diffusion of responsibility reduces sense of agency and outcome monitoring. *Social Cognitive and Affective Neuroscience, 12*(1), 138-145.

Beyer, F., Sidarus, N., Fleming, S., & Haggard, P. (2018). Losing control in social situations: how the presence of others affects neural processes related to sense of agency. *eneuro*, *5*(1).

Buehner, M. J. (2015). Awareness of voluntary and involuntary causal actions and their outcomes. Psychology of Consciousness*: Theory, Research, and Practice, 2*(3), 237.

Buehner, M. J., & Humphreys, G.R. (2009). Causal binding of actions to their effects. *Psychological Science, 20*(10), 1221-1228.

Capozzi, F., Becchio, C., Garbarini, F., Savazzi, S., & Pia, L. (2016). Temporal perception in joint action: This is MY action. *Consciousness and Cognition*, *40*, 26-33.

Chaminade, T., Franklin, D. W., Oztop, E., & Cheng, G. (2005, July). Motor interference between humans and humanoid robots: Effect of biological and artificial motion. In *Proceedings. The 4th International Conference on Development and Learning, 2005* (pp. 96-101). IEEE.

Ciardo, F., De Tommaso, D., Beyer, F., & Wykowska, A. (2018, November). Reduced sense of agency in human-robot interaction. In *International conference on social robotics* (pp. 441-450). Springer, Cham.

Ciardo, F., Beyer, F., De Tommaso, D., & Wykowska, A. (2020). Attribution of intentional agency towards robots reduces one's own sense of agency. *Cognition*, 194, 104109.

Cozmo SDK installation for Windows. Available: http://cozmosdk.anki.com/docs/install-windows.html. Last access: 5/8/2020.

David, N., Newen, A., & Vogeley, K. (2008). The "sense of agency" and its underlying cognitive and neural mechanisms. *Consciousness and Cognition, 17*(2), 523-534.

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

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Sciences, 4*(1), 14-21.

Grynszpan, O., Sahaï, A., Hamidi, N., Pacherie, E., Berberian, B., Roche, L., & Saint-Bauzel, L. (2019). The sense of agency in human-human vs human-robot joint action. *Consciousness and Cognition*, *75*, 102820.

Haggard, P., & Clark, S. (2003). Intentional action: Conscious experience and neural prediction. *Consciousness and cognition, 12*(4), 695-707.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience, 5*(4), 382-385.

Hinz, N. A., Ciardo, F., & Wykowska, A. (2021). ERP markers of action planning and outcome monitoring in human–robot interaction. *Acta Psychologica*, *212*, 103216.

Khalighinejad, N., Bahrami, B., Caspar, E. A., & Haggard, P. (2016). Social transmission of experience of agency: An experimental study. *Frontiers in Psychology, 7*, 1315.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software, 82(13)*, 1-26.

Lenth, R. (2019). emmeans: Estimated Marginal Means, aka Least-Squares Means.

R package version 1.4.1. https://CRAN.R-project.org/package=emmeans

Liepelt, R., Prinz, W., & Brass, M. (2010). When do we simulate non-human agents? Dissociating communicative and non-communicative actions. *Cognition*, *115*(3), 426-434.

Limerick, H., Coyle, D., & Moore, J. W. (2014). The experience of agency in human-computer interactions: a review. *Frontiers in Human Neuroscience, 8,* 643.

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.

Massen, C., & Prinz, W. (2009). Movements, actions and tool-use actions: an ideomotor approach to imitation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2349-2358.

Moore, J. W., & Obhi, S. S. (2012). Intentional binding and the sense of agency: a review. *Consciousness and Cognition, 21*(1), 546-561.

Moore, J. W., Lagnado, D., Deal, D. C., & Haggard, P. (2009). Feelings of control: contingency determines experience of action. *Cognition, 110*(2), 279-283.

Obhi, S. S., & Hall, P. (2011). Sense of agency in joint action: Influence of human and computer co-actors. *Experimental Brain Research, 211*(3-4), 663-670.

Perez-Osorio, J., & Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology, 33*(3), 369-395.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology*, *9*(2), 129-154.

Ramnani, N., & Miall, R. C. (2004). A system in the human brain for predicting the actions of others. *Nature Neuroscience, 7*(1), 85-90.

Ruess, M., Thomaschke, R., & Kiesel, A. (2018). Intentional binding of visual effects. *Attention, Perception, & Psychophysics, 80*(3), 713-722.

Ruijten, P. A., Haans, A., Ham, J., & Midden, C. J. (2019). Perceived human-likeness of social robots: testing the Rasch model as a method for measuring anthropomorphism. *International Journal of Social Robotics, 11*(3), 477-494.

Sahaï, A., Pacherie, E., Grynszpan, O., & Berberian, B. (2017). Predictive mechanisms are not involved the same way during human-human vs. human-machine interactions: A review. *Frontiers in Neurorobotics*, *11*, 52.

Sahaï, A., Desantis, A., Grynszpan, O., Pacherie, E., & Berberian, B. (2019). Action co-representation and the sense of agency during a joint Simon task: Comparing human and machine co-agents. *Consciousness and Cognition, 67*, 44-55.

Strother, L., House, K. A., & Obhi, S. S. (2010). Subjective agency and awareness of shared actions. *Consciousness and Cognition*, *19*(1), 12-20.

Springer, A., Hamilton, A. F. & Cross, E. S. (2012). Simulating and predicting others' actions. *Psychological Research 76(4):* 383–387.

Team, R. C. (2013). R: A language and environment for statistical computing. URL: https://www.R-project.org/. Last access: 5/8/2020.

Tsakiris, M., & Haggard, P. (2003). Awareness of somatic events associated with a voluntary action. *Experimental Brain Research*, *149*(4), 439-446.

Vastano, R., Ambrosini, E., Ulloa, J. L., & Brass, M. (2020). Action selection conflict and intentional binding: An ERP study. *Cortex*, *126*, 182-199.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., & Cacioppo, J. T. (2010). Making sense by making sentient: effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3), 410.

Wohlschläger, A., Haggard, P., Gesierich, B., & Prinz, W. (2003). The perceived onset time of self-and other-generated actions. *Psychological Science*, *14*(6), 586-591.

Zapparoli, L., Seghezzi, S., Zirone, E., Guidali, G., Tettamanti, M., Banfi, G., ... & Paulesu, E. (2020). How the effects of actions become our own. *Science Advances*, 6(27), eaay8301.

2.3. Publication III: Human-likeness and attribution of intentionality predict vicarious sense of agency over humanoid robot actions

Cecilia Roselli[1,2], Francesca Ciardo[1], Davide De Tommaso[1], and Agnieszka Wykowska[1]

[1] Social Cognition in Human Robot Interaction, Fondazione Istituto Italiano di Tecnologia, Center for Human Technologies, via Enrico Melen 83, Genova, Italy

[2] DIBRIS, Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi, Via all'Opera Pia 13, Genova, Italy

Authors Contribution

C.R. designed and performed the experiment, analyzed the data, and wrote the manuscript. F.C. designed the study, discussed and interpreted the results, wrote the manuscript. D.D.T programmed the experiment and integrated the iCub robot. A.W. designed the study, discussed and interpreted the results, wrote the manuscript. All the authors revised the manuscript.

2.3.1. Abstract

Sense of Agency (SoA) is the feeling of controlling one's actions and outcomes. In a social context, people can experience a "vicarious" SoA over another human's actions; however, it is still controversial whether the same occurs in Human-Robot Interaction (HRI). The present study aims at understanding whether humanoid robots may elicit vicarious SoA in humans, and whether its occurrence depends on the intentionality attributed towards it. We asked adult participants to perform an Intentional Binding (IB) task alone and with the humanoid iCub robot, reporting the time of occurrence of both self- and iCub-generated actions. Before the experiment, participants filled out the Waytz questionnaire (Waytz et al., 2010) to assess intentionality attribution towards robots. Results showed that participants experienced vicarious SoA over iCub-generated actions. Moreover, intentionality attribution positively predicted the magnitude of vicarious SoA, so that the more participants attributed intentionality to robots the more they experienced control over iCub's actions. In conclusion, our results highlight the interplay of human-likeness and intentionality attribution for the emergence of vicarious SoA during HRI.

Keywords: Vicarious Sense of Agency, Intentional Binding, Human-Likeness, Intentionality Attribution, Human-Robot Interaction.

2.3.2. Introduction

Sense of Agency (SoA) is the feeling of being in control of one's actions and outcomes (Haggard, 2017). The Intentional Binding (IB) paradigm has been extensively used to investigate implicit SoA (Haggard, Clark, & Kalogeras, 2002; see Moore & Obhi, 2012 for a review), demonstrating that voluntary actions and their sensory outcomes are perceived as shifted towards each other in time. This effect has been termed the Intentional Binding (IB) effect (Moore & Obhi, 2012). Notably, the IB effect can emerge not only in relation to one's actions, but also towards other humans' actions (Strother, House, & Obhi, 2010), leading to a "vicarious" SoA. Interestingly, recent evidence showed that also artificial agents could induce vicarious SoA in humans or have an impact on SoA in general (e.g., Ciardo, Beyer, De Tommaso, & Wykowska, 2020). For example, Barlas (2019) demonstrated that the more participants perceived the robot as anthropomorphic, the more they felt in control over outcomes of actions performed following robot's instructions (Barlas, 2019). Moreover, observing an embodied anthropomorphic hand in action showed similar IB effects as other humans did (Khalighinejad, Bahrami, Caspar, & Haggard, 2016). However, other evidence showed that people did not display the vicarious IB effect over actions that they believed to be performed by a robotic device (Grynszpan et al., 2019). Thus, it remains controversial whether robots can actually elicit vicarious SoA, and under which conditions.

One factor potentially contributing to the emergence of vicarious SoA towards robots is the human-like (anthropomorphic) shape of the robot. For example, Stenzel and colleagues (2012) showed that participants' beliefs that the robot was biologically inspired yielded to the emergence of the Social Simon effect (Sebanz, Knoblich, & Prinz, 2003), which was a measure of participants' ability to co-represent one's own and partner's actions. Notably, the effect was diminished when the robot was introduced to participants as machine-like (Stenzel et al., 2012).

As the Social Simon effect and vicarious SoA seems to share common underlying mechanisms (Sahaï, Desantis, Grynszpan, Pacherie, & Berberian, 2019), it may be plausible that also vicarious SoA is modulated by the perceived human-likeness of the robot.

Specifically, it may be that the more robots display human-like features, the more humans would be able to accurately represent their actions at the sensorimotor level, and, consequently, the more probable would vicarious SoA to be observed. It would be based on the idea that implicit SoA seemed to depend on one's ability to form a sensorimotor representation of the partner's actions, regardless of whether it was another human (Sahaï, Pacherie, Grynszpan, & Berberian, 2017) or a robot perceived as "biologically plausible" (e.g., Chaminade, Franklin, Oztop, & Cheng, 2005; Liepelt, Prinz, & Brass, 2010). A recent IB study speaks in favor of this interpretation (Roselli, Ciardo, & Wykowska, 2021). Participants were asked to report the time of occurrence of actions and outcomes generated by the non-anthropomorphic Cozmo robot, which was programmed to perform either physical or digital actions. Results showed that vicarious SoA for robot's outcomes emerged only when the causing actions were physical, and not in the "digital" action condition. This was interpreted as resulting from a lack of representation (at the sensorimotor level) of digital action-outcome links generated by artificial agents. However, another factor potentially contributing to the emergence of vicarious SoA towards robots is the attribution of intentionality, which might occur despite robots' artificial mechanical nature (Marchesi et al., 2019; see Perez-Osorio and Wykowska, 2020 for a review). In the context of vicarious SoA, Roselli and colleagues (2021) found that the degree of attributed intentionality to robots predicted vicarious SoA over robot actions (Roselli et al., 2021).

2.3.3. Aims

The present study aimed at investigating whether vicarious SoA toward a humanoid robot can be observed and whether this phenomenon depends on the attribution of intentionality towards it.

Participants performed an IB task (Haggard et al., 2002; Strother et al., 2010; Obhi & Hall, 2011) alone and with the humanoid robot iCub (Metta et al., 2010). *Humanoid robot* means that it is relatively similar to a human shape, with similar effectors as humans. To test the potential role of the attribution of intentionality towards robots, before the experiment participants filled out the Waytz questionnaire (Waytz et al., 2010), which measures the individual level of likelihood of attribution of intentionality towards robots.

We hypothesized that, if the human-like shape of the robot and its effectors is sufficient to induce the vicarious SoA, then a comparable IB effect should emerge for both self-generated and iCub's actions, with no relation between the IB effect and the individual likelihood of attributing intentionality to robots. In contrast, if attribution of intentionality plays a role in vicarious SoA, one would expect that the higher attributed intentionality to robots in general, the stronger the vicarious IB effect at the individual level.

2.3.4. Materials and Methods

Participants. Thirty-four participants were recruited to participate in the study (age range: 18-45 years old, M_{age} = 26.5, SD_{age} = 6.14, 4 left-handed, 16 males). All participants had a normal or corrected-to-normal vision, and they were naïve to the purpose of the study. Sample size was determined based on an *a priori* power analysis estimating the sample needed to obtain reliable results. We used the *pwr* package (Champely et al., 2018) in R Studio v.4.0.5. (R Core Team, 2013), considering f^2 as the most reliable effect size measured for mixed-effects models (Selya, Rose, Dierker, Hedeker, & Mermelstein, 2012), which planned for the analyses. We used a medium-to-large effect size (f^2 = 0.3); the significance level (alpha) was set to 0.05, and the power of the test was set to 0.95. Results showed that a sample size of N = 30 was needed to obtain reliable results. We finally tested 34 participants to account for possible need to exclude

participants from analyses. The study was conducted with the approval of the Local Ethical Committee (Comitato Etico Regione Liguria) and under the ethical standards laid down in the 2013 Declaration of Helsinki. All participants gave written informed consent before the experiment, and they were all paid 15 € for their participation. After the experiment, all participants were debriefed about the purpose of the study.

Apparatus and Stimuli. The experimental apparatus comprised the iCub robot (Metta et al., 2010), a workstation equipped with two 27' inches screens to display the task (resolution 1920 x 1200), two sets of speakers and two identical QWERTY keyboards, one for participants and one for iCub (see **Figure 1**). Presentation of stimuli and response collection were controlled using Psychopy v2021.2.0 (Peirce, 2007). The humanoid robot iCub (Metta et al., 2010) was connected to the workstation using a peer-to-peer Ethernet connection (see Supplementary Materials, point SM.2, p. 138, for more information).

The Waytz questionnaire was programmed using OpenSesame v.3 (Mathôt, Schreij, & Theeuwes, 2012).

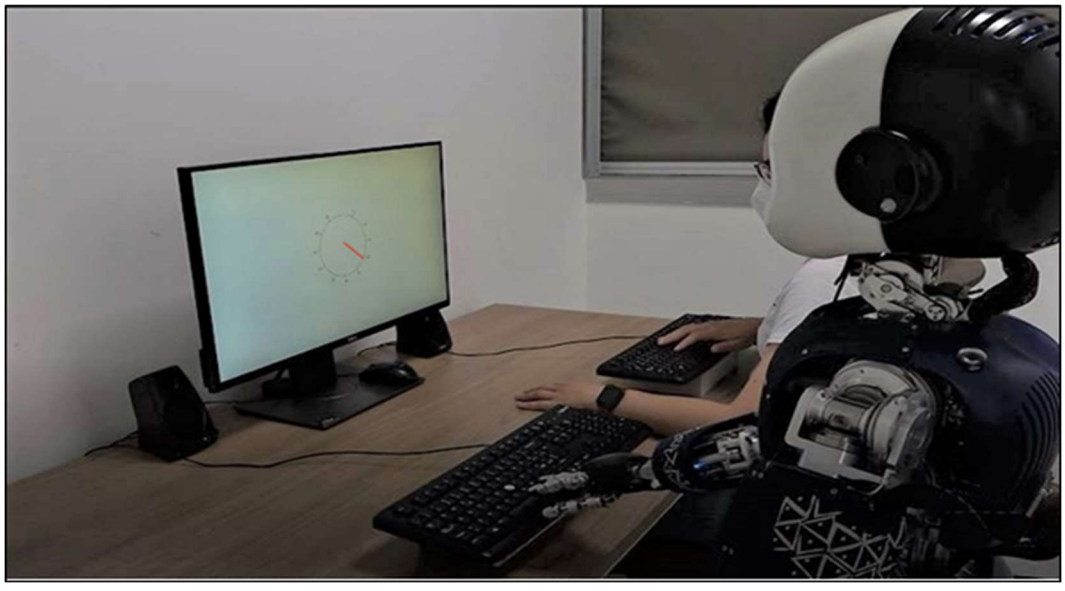

Figure 1. Experimental setup.

Procedure. Before the task, participants filled out the Waytz questionnaire (Waytz et al., 2010). Subsequently, they performed the IB task both alone (Solo Context) and with the iCub robot (Social Context).

Each Context (Solo, Social) included two types of (sub-) blocks of 40 trials each, i.e., a Baseline and an Operant block, order randomized. In the Baseline block, the critical event (i.e., action) did not produce any tone outcome, whereas in the Operant block the action produced a tone outcome 250 ms thereafter (440 Hz, 100 ms; See Supplementary Materials, point SM.1, p. 136, for more information). In the Solo Context, participants executed the task alone, with iCub being in a separate room. In the Social Context, participants entered the room where the robot was already activated, with its hand placed over its keyboard. In the Solo Context, participants' task was to perform a keypress at the time of their choosing, and subsequently report the time at which the keypress was made. In the Social Context, the task was to report the time at which iCub performed a keypress. A practice session (i.e., sixteen trials, four for each combination of Block and Context) was administered before the task.

Trial sequence. Participants were seated at approximately 70 cm from the computer screen. At the beginning of each trial, a fixation dot appeared on the screen for 1000 ms, followed by the image of a clock with a red clock hand (length = 135 pixels) in a static position for 500 ms. Afterward, the clock hand started rotating clockwise, with each rotation lasting 2560 ms. For each trial, the maximum number of rotations was set to 10. The clock hand stopped rotating randomly between 1500 and 2500 ms after the action occurred. In the Solo Context participants were instructed to wait until the end of the first full rotation of the clock hand, and then to perform a keypress at the time of their choosing. In the Social Context, the iCub robot was programmed to perform a keypress at a random time after the first full rotation of the clock hand, within a predefined time

window (2500-8000 ms) (see video "**Publication III**": https://osf.io/23jmt/?view_only=f58dfc2c426f45ba93a7eff5f931c43f). At the end of each trial, participants' task was to report the time indicated by the clock hand when they – or iCub – performed the keypress. To make sure that participants were attending iCub's actions, the robot was programmed to press in 90% of trials. Participants were instructed that if iCub did not act before the end of the tenth rotation, they had to execute the keypress themselves, otherwise, they would lose 10 points from a starting amount of 120 points.

2.3.5. Vicarious Sense of Agency

Statistical analyses. For each trial, we estimated the Judgment Error (JE), namely the "minute" difference between the position of the clock hand on the clock display reported by participants and its actual position. Then, "minute" JEs were transformed into "millisecond" JEs (minute JEs x 2560 ms/ 60). For each Block type (Baseline, Operant) we calculated the mean JEs and their standard deviations. JEs that deviated more than ± 2.5 SD from the participants' mean for each type of Block were considered as outliers and removed from the analyses (3.38 % of the total number of trials; mean JEs = 26.9 ms, SD = 435.77 ms). Data of three participants were excluded due to a low number of remaining trials in the Social Context subsequent to outliers' removal (< 30 trials in Baseline or Operant block, or both), resulting in a sample size of N = 31. Then, JEs were modeled as a function of Block type (Baseline, Operant) and Context (Solo, Social), plus their interactions, as fixed effects and participants as a random effect. Note that the IB for action events is defined as a smaller underestimation (i.e., less negative JEs) of the time of the action event for the Operant block, relative to the Baseline block (see Roselli et al., 2021 for a detailed description of the directionality of the IB effect for action events). Analyses were conducted using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2014) in R studio v. 4.0.5 (R Core Team,

2013). Parameters estimated (β) and their associated t-tests (t, p-value) were calculated using the Satterthwaite approximation method for degrees of freedom (Kuznetsova, Brockhoff, & Christensen, 2017); they were reported with the corresponding bootstrapped 95% confidence intervals (Efron & Tibshirani, 1994).

Results. Results showed a significant main effect of Block type [$\beta = 10.56$, $SE = 3.13$, $t_{(30)} = 3.37$, $p = 0.0007$, $CI = (4.42; 16.69)$], with less underestimated JEs in Operant compared to Baseline blocks [$\beta = -14.7$, $SE = 2.16$, $t_{(30)} = -6.81$, $p < 0.0001$, $CI = (-18.9; -10.5)$; ($M_{Operant} = -49.9$ ms, $SE_{Operant} = 8.74$; $M_{Baseline} = -64.6$ ms, $SE_{Baseline} = 8.74$;)]. Moreover, a significant main effect of Context emerged [$\beta = 16.01$, $SE = 3.04$, $t_{(30)} = 5.25$, $p < 0.0001$, $CI = (10.04; 21.99)$], with less underestimated JEs in Solo Context compared to the Social Context [$\beta = -20.1$, $SE = 2.16$, $t_{(30)} = -9.34$, $p < 0.0001$, $CI = (-24.4; -15.9)$; ($M_{Solo} = -47.2$ ms, $SE_{Solo} = 8.74$; $M_{Social} = -67.3$ ms, $SE_{Social} = 8.74$]. Notably, the two-way Block * Context interaction was not significant [$\beta = 8.24$, $SE = 4.31$, $t_{(30)} = 1.91$, $p = 0.05$, $CI = (-0.2; 16.69)$] (see **Figure 2**).

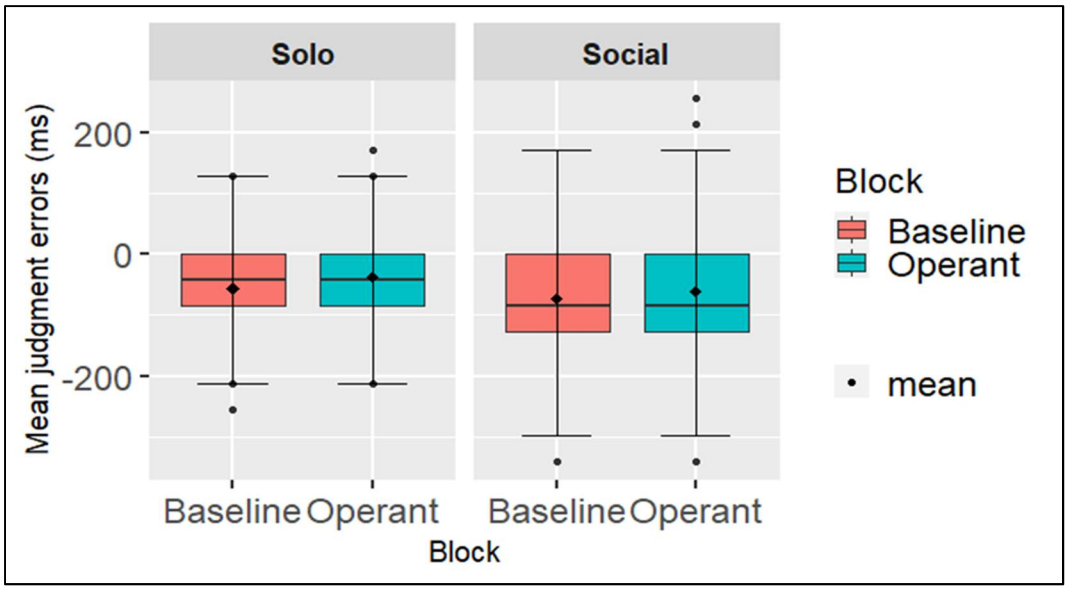

Figure 2. Mean JEs plotted as a function of Block (Baseline, Operant), separately for each Context (Solo, Social).

2.3.6. Intentionality attribution

Statistical analyses. To test the relationship between attribution of intentionality, as indexed by the Waytz score (Ruijten, Haans, Ham, & Midden, 2019) and the vicarious SoA, we selected only trials in the Social Context, i.e., when participants reported the time of occurrence of iCub's actions. Data of another participant were excluded due to unaccomplished completion of the Waytz questionnaire, resulting in a sample size of N = 30. Social JEs were modeled as a function of both Block type (Baseline, Operant) and Waytz score, plus their interactions, as fixed effects, and participants as a random effect. Analyses were conducted using the *lme4* package (Bates et al., 2014) in R studio v. 4.0.5 (R Core Team, 2013). Parameters estimated (β) and their associated t-tests (t, p-value) were calculated using the Satterthwaite approximation method for degrees of freedom (Kuznetsova et al., 2017), and then reported with the corresponding bootstrapped 95% confidence intervals (Efron & Tibshirani, 1994).

Results. Results showed no main effect of Block [$\beta = -5.85$, SE = 8.52, $t_{(29)} = -0.68$, $p = 0.49$, CI = (-22.55; 10.84)]. Moreover, no significant main effect of Waytz emerged [$\beta = 4.34$, SE = 10.09, $t_{(29)} = 0.43$, $p = 0.67$, CI = (-15.38; 24.07)]. Notably, the two-way Block * Waytz interaction resulted to be significant [$\beta = 6.4$, $t_{(29)} = 2.1$, $p = 0.03$, CI = (0.44; 12.35)]. Specifically, Waytz score predicted JEs only in the Operant block [$\beta = 10.41$, SE = 2.78, $t_{(29)} = 3.74$, $p = 0.0009$, CI = (4.96; 15.87)], and not in the Baseline block [$\beta = 4.07$, SE = 2.58, $t_{(29)} = 1.57$, $p = 0.11$, CI = (-0.99; 9.13)] (see **Figure 3**).

Figure 3. Mean JEs in the Social Context plotted as a function of Waytz score for both Baseline and Operant block.

2.3.7. General Discussion

The present study examined whether humanoid robots can elicit vicarious SoA, and whether it is related to the attribution of intentionality towards robots. Participants performed an IB task (Haggard et al., 2002) both alone (Solo Context) and with the humanoid iCub robot (Social Context). To assess the role of attribution of intentionality, participants filled out the Waytz questionnaire (Waytz et al., 2010) before the experiment. Our dependent measure was the Judgment Error (JE) in the IB task, i.e., the difference between the perceived and the actual position of the clock hand when the critical event (action) occurred.

Results showed that participants experienced SoA over both self-generated and iCub's actions, as demonstrated by the significant IB effect emerging in both Solo and Social contexts. The lack of the two-way Block * Context interaction suggests that participants experienced SoA over robot's actions similarly to their actions. Therefore, we might speculate that iCub's physical similarity with humans allowed participants to accurately represent iCub's actions at the sensorimotor level.

Such finding would align with previous evidence showing that vicarious SoA emerged only when the co-agent was an embodied, anthropomorphic robot (Khalighinejad et al, 2016), and not with a non-anthropomorphic robot whose effectors were too dissimilar to humans' ones (Roselli et al., 2021).

Regarding the role of attribution of intentionality, results showed that the magnitude of vicarious SoA was positively predicted by the degree of attributed intentionality. Notably, the Waytz score resulted to be predictive only of JEs in Operant block, i.e., when both events (actions and tone) were present, suggesting that attribution of intentionality led participants to perceive iCub's actions as linked to the subsequent outcome.

2.3.8. Conclusions

Taken together, our results indicated that both human-like shape and attribution intentionality are crucial factors playing a role in vicarious SoA towards robots. Specifically, the physical similarity between humanoid robots and humans may allow people to represent the robot's actions similarly to self-generated actions. However, the individual tendency of attributing intentionality to robots may additionally "boost" the vicarious SoA. These findings extend knowledge about mechanisms underlying SoA in a social context with other agents, but also help to design robots that we may treat as social companions in real-life scenarios.

2.3.9. Acknowledgments

The authors are grateful to Giulia Siri for her help in data collection.

2.3.10. Funding

This work has received support from the European Research Council under the European Union's Horizon 2020 research and innovation program ERC Starting Grant, G.A. number: ERC – 2016- StG- 715058, awarded to AW. 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.

2.3.11. References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Barlas, Z. (2019). When robots tell you what to do: Sense of agency in human-and robot-guided actions. *Consciousness and cognition*, *75*, 102819

Ciardo, F., Beyer, F., De Tommaso, D., & Wykowska, A. (2020). Attribution of intentional agency towards robots reduces one's own sense of agency. *Cognition*, 194, 104109.

Chaminade, T., Franklin, D. W., Oztop, E., & Cheng, G. (2005, July). Motor interference between humans and humanoid robots: Effect of biological and artificial motion. In *Proceedings. The 4th International Conference on Development and Learning, 2005* (pp. 96-101). IEEE.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., ... & De Rosario, M. H. (2018). Package 'pwr'. *R package version*, *1*(2).

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Grynszpan, O., Sahaï, A., Hamidi, N., Pacherie, E., Berberian, B., Roche, L., & Saint-Bauzel, L. (2019). The sense of agency in human-human vs human-robot joint action. *Consciousness and Cognition*, *75*, 102820.

Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews Neuroscience*, *18*(4), 196-207.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience, 5*(4), 382-385.

Khalighinejad, N., Bahrami, B., Caspar, E. A., & Haggard, P. (2016). Social transmission of experience of agency: An experimental study. *Frontiers in Psychology, 7*, 1315.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software, 82(13)*, 1-26.

Liepelt, R., Prinz, W., & Brass, M. (2010). When do we simulate non-human agents? Dissociating communicative and non-communicative actions. *Cognition*, *115*(3), 426-434.

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.

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., 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.

Moore, J. W., & Obhi, S. S. (2012). Intentional binding and the sense of agency: a review. *Consciousness and Cognition, 21*(1), 546-561.

Obhi, S. S., & Hall, P. (2011). Sense of agency in joint action: Influence of human and computer co-actors. *Experimental Brain Research, 211*(3-4), 663-670.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8-13.

Perez-Osorio, J., & Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology, 33*(3), 369-395.

Roselli, C., Ciardo, F., & Wykowska, A. (2021). Intentions with actions: The role of intentionality attribution on the vicarious sense of agency in Human–Robot interaction. *Quarterly Journal of Experimental Psychology*, 17470218211042003.

Ruijten, P. A., Haans, A., Ham, J., & Midden, C. J. (2019). Perceived human-likeness of social robots: testing the Rasch model as a method for measuring anthropomorphism. *International Journal of Social Robotics, 11*(3), 477-494.

Sahaï, A., Desantis, A., Grynszpan, O., Pacherie, E., & Berberian, B. (2019). Action co-representation and the sense of agency during a joint Simon task: Comparing human and machine co-agents. *Consciousness and Cognition*, *67*, 44-55.

Sahaï, A., Pacherie, E., Grynszpan, O., & Berberian, B. (2017). Predictive mechanisms are not involved the same way during human-human vs. human-machine interactions: A review. *Frontiers in Neurorobotics*, *11*, 52.

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: just like one's own?. *Cognition*, *88*(3), B11-B21.

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A practical guide to calculating Cohen's f2, a measure of local effect size, from PROC MIXED. *Frontiers in Psychology*, *3*, 111.

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

Strother, L., House, K. A., & Obhi, S. S. (2010). Subjective agency and awareness of shared actions. *Consciousness and Cognition*, *19*(1), 12-20.

Team, R. C. (2013). R: A language and environment for statistical computing. URL: https://www.R-project.org/. Last access: 5/8/2020.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., & Cacioppo, J. T. (2010). Making sense by making sentient: effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3), 410.

SECTION III- CONCLUSIONS

3.1. Synopsis of the results

The aim of the work described in the present thesis was twofold. First, I aimed to investigate whether humans can experience vicarious SoA over robot's actions and outcomes, and under what conditions. Second, I aimed to disentangle whether vicarious SoA may serve as an implicit measure of attribution of intentionality towards robots. To address these questions, I conducted a series of studies employing the Intentional Binding (IB) paradigm (Haggard et al., 2002; see Moore & Obhi, 2012 for a review) as a reliable and well-established measure of implicit SoA.

The study reported in **Publication I** aimed at adapting the IB task to an experimental HRI protocol, to be able to measure SoA at the implicit level. Furthermore, it aimed at replicating – with an implicit measure – the phenomenon of diffusion of responsibility during HRI reported by Ciardo and colleagues (2018; 2020). To this aim, participants performed an IB task both alone (Solo Condition) and with the non-anthropomorphic Cozmo robot (Social Condition). We focused our interest on judgment errors for self-generated actions and outcomes. Results showed that the mere social presence of the robot affected participants' SoA for their actions only. Specifically, when the robot was in charge to perform the task (Social Condition), participants were more accurate in reporting the time of occurrence of self-generated actions compared to when they performed the task alone (Individual Condition), thus resulting in a smaller IB effect relative to the Individual Condition. However, the social presence of the robot did not affect judgments of self-generated outcomes, as suggested by the comparable IB effect both when the robot was in charge to perform the task (Social Condition) and when it was not (Individual Condition). Taken together, the results of Publication I led to two main findings. First, they were in line with the results from Ciardo and colleagues' studies (2018; 2020), which demonstrated – using explicit measures - that the social presence of the robot affected participants' SoA. Second, we found a dissociation between actions

and outcomes, as in our study the social presence of the robot affected participants' SoA only for self-generated actions. Finally, and importantly, the IB paradigm applied to an HRI protocol resulted in eliciting the IB effect and can be used as a method to implicitly investigate humans' SoA during social interactions with robotic agents.

The study reported in **Publication II** aimed at investigating whether a non-anthropomorphic robot could elicit vicarious SoA in humans. In two experiments, we tested the contribution of two potential factors playing a role in the emergence of vicarious SoA, namely (1) the possibility to represent the robot's actions (physical vs. digital) at the sensorimotor level, and (2) the attribution of intentionality towards robots. Results showed that (1) the possibility to represent the robot's actions at the sensorimotor level affected the vicarious IB effect, and (2) the degree of attribution of intentionality predicted the magnitude of the vicarious SoA, but only for the robot's physical actions. Notably, this was not observed when the robot's actions were digital. When the critical event was the sensory outcome of the robot's action, the degree of intentionality attribution always predicted the magnitude of vicarious SoA. Taken together, these results suggested that both the possibility to represent the robot's actions and attribution of intentionality play a role in the emergence of vicarious SoA towards robots.

The study reported in **Publication III** assessed vicarious SoA over actions performed by the humanoid robot iCub (Metta et al., 2008). We employed an IB task and focused on the IB effect for action events only. Results showed that the vicarious IB effect emerged for iCub's actions. Moreover, individuals' tendency to attribute intentionality to robots positively predicted the magnitude of vicarious SoA, in such a way that the more participants tended to attribute intentionality to robots the more they experienced vicarious SoA. In conclusion, the results of

Publication III highlighted that the emergence of vicarious SoA over robot's actions results from the interplay between human-likeness of the robot's appearance and intentionality attribution.

3.2. The role of action representation in the emergence of vicarious SoA towards robots

The "motor" models of SoA, namely the *Comparator Model* (e.g., Wolpert et al., 1995; Frith et al., 2000; Frith, 2005) and the *Ideomotor Theory* (e.g., Prinz, 1997; Hommel et al., 2001; Hommel et al., 2003; Massen & Prinz, 2009) suggest that humans represent – at the sensorimotor level – the causal links between actions and their sensory outcomes. In other words, humans simulate the motor commands before action execution (self- or other-generated), which is then used to predict the outcomes of the action (Pacherie & Dokic, 2006). Accordingly, the possibility to create a sensorimotor representation of the action-outcome link is strictly connected to the experience of self-agency (e.g., Zapparoli et al., 2020). Following this reasoning, also vicarious SoA should rely on the same predictive mechanisms. Thus, the lack of vicarious SoA for computers, as reported by previous evidence (e.g., Obhi & Hall, 2011; Sahaï et al., 2019) might be explained by the disembodied nature of computer's actions. In other words, when the co-agent is a computer, no sensorimotor representation of the action-outcome link is activated (Ramnani & Miall, 2004), as its actions are not embodied as humans' actions. As a consequence, people do not experience vicarious SoA (e.g., Sahaï et al., 2019).

The findings of the present Ph.D. thesis showed that, when interacting with a robot, the possibility to form a sensorimotor representation of the causal and perceptual action-outcome link plays a crucial role. Indeed, results of Publication II showed that when the action of the robot was digital, i.e., it could not be represented in terms of action effects, humans did not experience vicarious

SoA over outcomes generated by robots. This suggests that to experience control over others' outcomes, humans need to have access to the action representation that generated that outcome.

Regarding vicarious SoA for robot's actions, results of Publication II and Publication III are consistent with the sensorimotor hypothesis, according to which humans are generally better at representing those actions that are part of their motor repertoire, and that they would be able to perform themselves (Springer, Hamilton, & Cross, 2012). Indeed, our results showed that vicarious SoA for action events emerged only for a humanoid robot (Publication III). This is in line with previous evidence reporting vicarious SoA for actions performed by a robotic hand resembling a human-like shape (Khalighinejad et al., 2016). Importantly, results of Publication II indicated that no vicarious SoA emerged over actions performed by a non-anthropomorphic robot (i.e., the Cozmo robot). Notably, the effector used by Cozmo to perform the physical action (i.e., the lift) differs, in terms of shape and motion, from the one that humans use to perform the same kind of action (i.e., the hand). As a result, the sensorimotor representation of Cozmo's actions might have been less accurate. When Cozmo's actions were digital, and thus they could not be represented either at the sensorimotor level or merely in terms of their action effects, vicarious SoA never occurred (also not with respect to the judgment regarding the outcome). This is in line with previous evidence reporting a lack of vicarious SoA over disembodied actions executed by a computer (Sahaï et al., 2019), which does not have visible effectors moving in the environment and thus it cannot activate the sensorimotor representation of the action.

Taken together, our results highlighted two main aspects. First, a dissociation between action and outcomes in the role of the action representation emerged when participants focused on outcomes generated by the robots, rather than on the actions that generated them. It was in line with the results of Publication I, which showed that the social presence of the robot reduced humans' SoA

only for self-generated actions, and not for outcomes. Along the same line, previous evidence suggested that the IB effect for actions and outcomes would rely on different mechanisms (e.g., Wolpe, Haggard, Siebner, & Rowe, 2013; Zapparoli et al., 2020).

Second, our findings showed that the possibility to form an accurate sensorimotor representation of the robot's actions plays a crucial role in the emergence of vicarious SoA towards robots. Interestingly, the more the robot resembles the human-like shape, in terms of effectors used to execute the action, the more accurate the representation of its action is, with a greater feeling of agency experienced by the observer (i.e., vicarious SoA), over the robot's actions.

3.3. The role of intentionality attribution in the emergence of vicarious SoA towards robots

When interacting with a robot, in some contexts people adopt the Intentional instead of the Design stance to explain its behavior. That is, they perceive the robot as an intentional system rather than as a pre-programmed artifact (Dennett, 1971, 1981). Interestingly, the attribution of intentionality towards artificial systems, such as computers and robots, seems to relate with vicarious SoA, as previous evidence suggested that people do not experience vicarious SoA when they do not perceive these systems as intentional (e.g. Wohlschläger et al., 2003; Sahaï et al., 2019; Ciardo et al., 2020). It is in line with previous evidence showing a relationship between intentionality and SoA also at the individual level, since people are able to experience control over their actions only when they are voluntary and intentional (e.g., Haggard et al., 2003).

In this thesis, the role of attribution of intentionality in the emergence of vicarious SoA has been investigated in Publication II and Publication III. Results of Publication II showed that the attribution of intentionality towards robots acts differently when participants focused on the robot's outcome, rather than on the action that produced it. Specifically, individuals' tendency to

attribute intentionality to robots reduced vicarious SoA over robot's outcomes, regardless of the nature of the causing action (i.e., physical vs. digital). Conversely, it seems to act as a reinforcement of vicarious SoA over the robot actions, when the sensorimotor representation is possible. In other words, if robots are perceived as intentional agents, it may be that robot-generated actions are represented similarly to human-generated ones. When the robot is non-anthropomorphic (i.e., the Cozmo robot), the sensorimotor representation of its actions was not sufficiently accurate to elicit vicarious SoA by itself, due to the different types of effectors that the robot used to perform the physical action. In this context, the degree of attributed intentionality might boost the perceived similarity between the human and the robot. Consequently, vicarious SoA over robot physical actions may emerge. For digital actions, since their disembodied nature does not allow forming a sensorimotor representation, vicarious SoA never emerged, and the attributed intentionality could not help in this case. Importantly, however, when a robot has a human-like shape (e.g., the iCub robot), the sensorimotor representation of its action is sufficiently accurate to elicit vicarious SoA by itself. However, the results of Publication III showed that intentionality attribution positively predicted the magnitude of vicarious SoA over iCub's actions, suggesting that the effect can be magnified by the attribution of intentionality.

The role of attributed intentionality towards robots in the emergence of SoA would be in line with previous evidence showing that a robot – presumably perceived as an intentional agent – was able to affect one's own SoA, in contrast to a non-agentic, passive device (Ciardo et al., 2020). In a similar vein, Wohlschläger and colleagues (2003) interpreted the lack of vicarious SoA over actions executed by the mechanical lever as the result of a lack of intentionality attributed to the system (Wohlschläger et al., 2003). These findings may be explained in the light of the *Theory of Apparent Mental Causation* (Wegner & Wheatley, 1999; Wegner, 2002), which emphasizes the

role of high-level conceptual beliefs for the emergence of SoA. According to this theory, humans may use an *a priori* attribution of intentionality towards robots to explain the causal origin of robot actions. In other words, by attributing intentionality to robots, humans would infer that the "intention to act" of the robot caused its actions and the subsequent outcomes.

Taken together, our findings suggest that vicarious SoA can be used an as implicit measure of attribution of intentionality towards robots, in such a way that the more people ascribe intentional traits to robots the more they experience vicarious SoA.

3.4. The interplay of action representation and intentionality attribution in vicarious SoA towards robots

The findings of the present Ph.D. project highlighted that the phenomenon of vicarious SoA is the result of the interplay between low-level motor processes (i.e., the possibility to represent the robot's actions at the sensorimotor level), and high-level conceptual beliefs (i.e., individuals' tendency to attribute intentionality to robots). To allow for an accurate sensorimotor representation, the robot's actions need to be embodied, similar to actions that humans would be able to perform themselves. Notably, the more the effectors of the robot resemble a human-like shape, the more the representation of the robot's actions will be accurate, i.e., allowing humans to use the same predictive mechanism underlying their own motor control. In this context, the degree of attribution of intentionality towards robots acts as a reinforcement of the perceived similarity between humans and the robot.

These findings would support the view proposed by the *Cue Integration Theory* (Synofzik et al., 2008; Moore et al., 2009; Moore & Fletcher, 2012) according to which the brain integrates information from multiple internal and external cues to produce the experience of agency (see **Figure 6**).

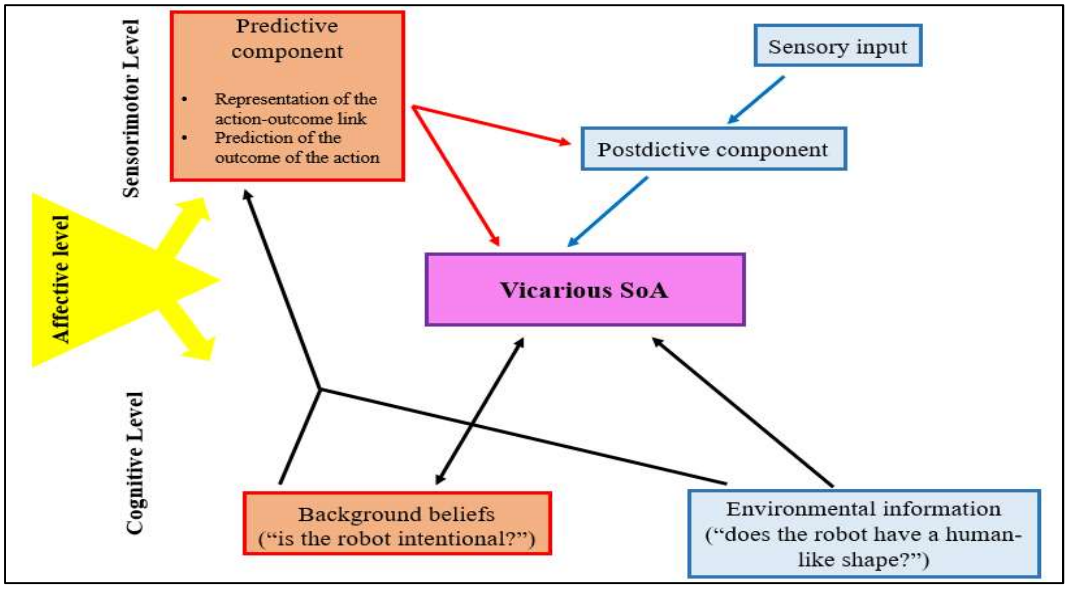

Figure 6. Schematic representation of the emergence of the vicarious SoA in HRI, in the light of the *Cue Integration Theory* (Synofzik et al., 2008; Moore et al., 2009; Moore & Fletcher, 2012). At the sensorimotor level, the predictive component comprises the sensorimotor representation of the action-outcome link, which is used to predict the outcome of the action. It can be integrated with external cues, such as the sensory outcomes- that represent the postdictive component. In presence of relevant cognitive information, such as the human-like shape of the robot, the sensorimotor presentation can be enhanced by increasing the perceived similarity between the humans and the robot. As a result, vicarious SoA can arise. If it is not the case, the degree of intentionality attributed to the robot can lead to the emergence of vicarious SoA only as a function of intentionality attribution.

The weightings of these cues are determined by their availability and reliability so that if one or more are lacking, vicarious SoA might not occur. For example, in the case of digital actions, which are lacking the necessary embodiment to allow people to represent them in sensorimotor terms, vicarious SoA does not occur.

In conclusion, these findings shed new light on the complex nature of the vicarious SoA towards artificial agents, suggesting that the availability and reliability of both internal (i.e., representation of robot's actions, degree of intentionality attribution to it) and external information (i.e., human-likeness of the robot) affect the presence and the magnitude of the vicarious SoA over robot's actions and outcomes.

3.5. Implications for the investigation of vicarious SoA towards robots

Vicarious SoA over robot's actions and outcomes is an important phenomenon in situations of interaction with those agents. For example, if we think about driving automation, the degree of control that the human driver experiences over the vehicle is crucial for the monitoring of the driving process and possible intervention when needed. If the human driver does not feel in control over the vehicle, it may have negative consequences. For instance, the driver may not be able to actively engage in driving, or excessively rely on the driving automation system (Wen, Kuroki, & Asama, 2019). In a medical context, if for example a failure occurs in a robotic surgery, causing permanent injuries to a patient, who should be blamed? The robot who physically injured the patient, or the medical doctor who was teleoperating the robot during the surgery? Acquiring knowledge about how people can feel in control over robot's actions and outcomes, and to what extent, can be beneficial to the construction of a safer technology potentially accounting for all possible abnormal conditions of use (Vilaza, Haselager, Campos, & Vuurpijl, 2014). Moreover, it could help to design robots while considering the impact that the presence of an embodied artificial agent may have on humans' decision-making. For example, it would be good if, in emergencies, robots were able to efficiently detect an emergency signal and act upon it, as the human counterpart may be not efficient and fast enough to react (Ciardo et al., 2020).

In a nutshell, a better comprehension of how human's SoA works in a social context with artificial agents may help to develop systems that can successfully adapt to humans.

3.6. Limitations and future directions

Results reported in this thesis highlighted the relevance of the vicarious SoA phenomenon for the HRI field. Some limitations, however, emerged in the studies reported here. One potential

limitation may be that our results refer only to vicarious SoA investigated at the implicit level, i.e., using the IB paradigm (Haggard et al., 2002; see Moore & Obhi, 2012 for a review). Although focusing on the implicit dimension of SoA was a deliberate decision, in order to avoid cognitive biases of explicit measures (see paragraph 1.2., p. 5), the combination of both implicit and explicit measures of SoA could have allowed having a better overview of the vicarious SoA phenomenon in HRI. Another potential limitation of the studies reported in the thesis is the behavioral nature of the task. Although using only behavioral measures was a choice motivated by the fact that, at present, little is still known about the vicarious SoA phenomenon in HRI, it is a matter of fact that our evidence is limited to the processing of temporal intervals. Thus, it does not allow us to conclude how people process both actions and outcomes at the neural level. Further research involving electrophysiological measures (EEG) should confirm our behavioral evidence, for example by investigating whether ERPs for action planning and outcome processing (e.g., Readiness Potential and N100 components; e.g., Libet et al., 1983; Näätänen & Picton, 1987) are similarly affected for self-generated and robot's actions in an IB task (e.g., Hinz, Ciardo, & Wykowska, 2021).

Finally, even though for all the studies included in the thesis we used a social interaction protocol, the nature of the task may lack ecological validity for two reasons. First, the IB task (i.e., reporting the position of the clock hand at the occurrence of a critical event) is not a common task in everyday life, especially if we have to imagine tasks that will be shared with robots. Although our choice of using the IB task was motivated by the need to compare our results with previous evidence collected in cognitive neuroscience studies on human-human and human-computer interactions, future studies might consider conducting experiments on vicarious SoA in setups of higher ecological validity. Second, our results are so far limited to the type of robots that we used, i.e.,

the Cozmo and the iCub robots. However, robots are a wide category of agents which can have various shapes and can be able to perform different types of actions according to the task they are developed for. Thus, it will be crucial for future studies to investigate vicarious SoA in HRI by employing more interactive and ecologically valid scenarios and across different types of robots. Until then, I hope I have offered some useful hints to understand the relevance of the vicarious SoA phenomenon in HRI, which may be potentially employed to design robots able to be successfully integrated into human societies.

3.4. Conclusions

In the near future, the presence of robots in social spaces shared with humans will rapidly increase. In the attempt to design well-tailored robots that can successfully interact with humans, the work described in this thesis focused on the investigation of the vicarious SoA. To this aim, we identified factors potentially relevant for the emergence of vicarious SoA in HRI: (1) the possibility to represent the robot's actions at the sensorimotor level, (2), the human-like shape of the robot, and (3) the degree of attribution of intentionality towards robots. Results presented in this thesis highlighted that the interplay among these factors plays a role in vicarious SoA over robot actions and their outcomes. In more detail, these findings suggested that a possibility to represent robot actions at the sensorimotor level leads to the emergence of vicarious SoA, and this is strengthened by a robot's resemblance to a human-like shape. In addition, individuals' tendency to attribute intentionality to robots enhances the vicarious SoA. Hence, vicarious SoA can serve as an implicit measure of attribution of intentionality to robots.

In conclusion, the evidence collected in this thesis extended knowledge about the vicarious SoA phenomenon and gave potential hints to design robots well-tailored to humans' attitudes and needs.

SUPPLEMENTARY MATERIALS

Supplementary Materials

Publication II: Intentions with actions: the role of intentionality attribution on the vicarious sense of agency in Human-Robot Interaction

Cecilia Roselli [1,2], Francesca Ciardo [1], and Agnieszka Wykowska [1]

[1] Social Cognition in Human Robot Interaction, Fondazione Istituto Italiano di Tecnologia, Center for Human Technologies, via Enrico Melen 83, Genova, Italy

[2] DIBRIS, Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi, Via all'Opera Pia 13, Genova, Italy

This section includes:

SM.1. Comparisons across experiments

Results showed that, when the to-be-judged event was the tone, participants experienced vicarious SoA only in Experiment 1, when the tone was the outcome of the robot's physical action, and not in Experiment 2, when the robot-generated action that caused the tone was digital. In action blocks, for the Social Context participants did not experience vicarious SoA (i.e., Social IB effect) for robot's physical actions in Experiment 1; interestingly, in Experiment 2 the Social IB effect had a reversed direction, with JEs resulted to be significantly more positive in Baseline compared to Operant block (i.e., larger overestimation).

To address directly the combinatorial effect of the type of action, physical or digital, and the following outcome in the emergence of vicarious SoA for robot's actions, we performed additional analyses to compare the results of Social contexts across the two experiments.

Statistical analyses

These analyses were based solely on data collected in the Social Context (i.e., when participants judged the occurrence of Cozmo's actions and outcomes). To investigate whether the nature of the robot's action (physical in Experiment 1, digital in Experiment 2) was predictive of the magnitude of IB effect in Social context, we run two identical mixed models, one for action and one for tone separately. JEs in Social Context were modeled as a function of Block type (Baseline, Operant), and Experiment (1, 2) as fixed effects, and participants as a random effect. Analyses were conducted by using the lme4 package (Bates et al., 2015) in R v.3.0.6. (R Core Team, 2014). Parameters estimated (β) and their associated t-tests (t, p-value) were calculated using the Satterthwaite approximation method for degrees of freedom (Kuznetsova et al., 2017); they were reported with the corresponding bootstrapped 95% confidence intervals (Tibshirani and Efron,

1993). Following two-way significant interaction, pairwise comparisons were performed with the 'emmeans' package in R studio (Lenth, 2019).

Action event. Results showed a significant main effect of Experiment, with a larger underestimation (i.e., more negative JEs) in Experiment 1 compared to Experiment 2 [β = 356.92, SE = 21.64, t (73.01) = -16.48, p < .0001, CI = (314.53; 399.31); (M Experiment 1 = -77.7, SE Experiment 1 = 15.2; M Experiment 2 = 251.8, SE Experiment 2 = 15.1)]. Moreover, the Block type * Experiment interaction was significant [β = -54.78, SE = 6.32, t (4744.64) = -8.66, p < .0001, CI = (-67.16; -42.38)]. Pairwise comparisons (Tukey's HSD correction for multiple comparisons) were run to investigate the significant two-way interaction. In line with previous results reported in the main text, a lack of vicarious IB effect was observed in Experiment 1, with no significant differences between JEs in Operant compared to Baseline block [β = -1.94, SE = 4.55, t (4744.9) = -0.42, p = 0.97, CI = (-13.6; 9.74); (M Operant = -76.7, SE Operant = 15.3; M Baseline = -78.7, SE Baseline = 15.3)]. In Experiment 2, a significant vicarious IB effect appeared to be in the opposite direction, with a smaller overestimation (i.e., less positive JEs) in Operant compared to the Baseline block [β = 52.84, SE = 4.39, t (4744.3) = 12.07, p < .0001 CI = (41.6; 64.1); (M Operant = 225.4, SE Operant = 15.3; M Baseline = 278.3, SE Baseline = 15.3)] (see Figure **SM1, Panel A**).

Tone event. Results showed a significant main effect of Block type, with a smaller overestimation (i.e., less positive JEs) in Operant compared to the Baseline block [β = -65.99, SE = 4.46, t (5039.32) = -14.77, p < .0001, CI = (-74.74; -57.23); (M Operant = 52.9, SE Operant = 8.75; M Baseline = 82.1, SE Baseline = 8.7)]. Moreover, the Block type * Experiment interaction resulted to be significant [β = 73.52, SE = 6.22, t (5037.07) = 11.81, p < .0001, CI = (61.8; 85.7)]. In order to explore the two-way significant interaction, we further run pairwise comparisons (Tukey's HSD

correction for multiple comparisons). They revealed that, in Experiment 1, overestimation was smaller (i.e., less positive JEs) in Operant compared to the Baseline block, thereby indicating a vicarious IB effect [β = 65.99, SE = 4.46, t = 14.77, p < .0001, CI = (54.5; 77.4); (M Operant = 24.5, SE Operant = 12.3; M Baseline = 90.5, SE Baseline = 12.2)]. However, this was not true in Experiment 2, where the absence of significant differences between JEs in Operant block and in the corresponding Baseline signaled the lack of the IB effect [β = -7.53, SE = 4.33, t (5034.6) = -1.73, p = 0.3, CI = (-18.7; 3.61); (M Operant = 81.3, SE Operant = 12.4; M Baseline = 73.8, SE Baseline = 12.4)] (see Figure **SM1, Panel B**).

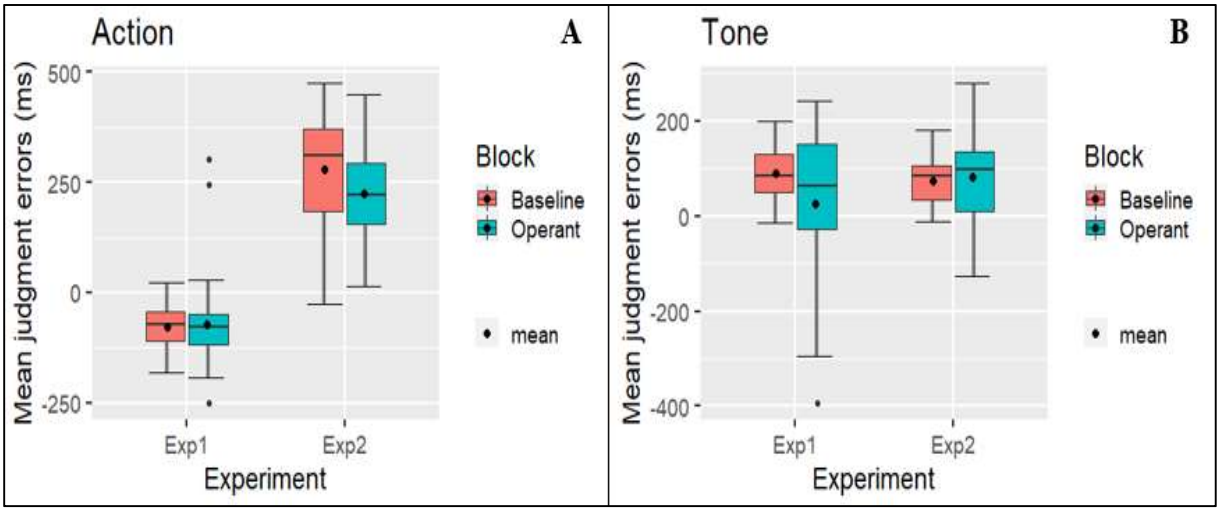

Figure SM.1. Mean judgment errors (JEs) are plotted as a function of Block type (Baseline, Operant), and Experiment (Exp1, Exp2), separately for Action (Panel A, on the left side) and Tone (Panel B, on the right side).

SM.2. Additional and exploratory analyses

SM.2.1. Regression models

When evaluating the role of attribution of intentionality for the emergence of vicarious SoA (paragraph 2.2.5, p. 79 onwards), we firstly run two regression models for the two experiments separately, with IB effect for trials of the Social Context as the dependent variable and Waytz score as a predictor variable (model: Social IB ~ Waytz). We used the coefficient of determination R^2 as a conventional goodness-of-fit measure (Cohen, 2013), as it represents a sample estimate of the proportion of variance in the outcome explained by the predictor (Miles, 2014).

For Action Events, results showed that, in Experiment 1, the regression equation did not fit the data well. Indeed only 9% of variance was explained [$\beta = 25.08$, $t = 1.88$, $p = 0.07$, $R^2 = 0.09$, CI = (-2; 52.18)]. This was true also for Experiment 2, with the model explaining only the 5% [$\beta = -4.58$, $t = -0.4$, $p = 0.69$, $R^2 = 0.005$, CI = (-27.65; 18.49)]. A similar pattern emerged for Tone Events, both in Experiment 1 [$\beta = 9.37$, $t = 0.55$, $p = 0.59$, $R^2 = 0.006$, CI = (-37.79; 58.97)] and Experiment 2 [$\beta = 10.59$, $t = 0.45$, $p = 0.67$, $R^2 = 0.009$, CI = (-25.48; 44.22)]. Therefore, we did not report these models and we used linear mixed-effects models with JEs as dependent variable (for results, please see paragraph 2.2.5.2, p.80).

SM.2.2. Models comparison

When evaluating the role of the attribution of intentionality for the emergence of vicarious SoA (paragraph 2.2.5, p. 79 onwards), we first decided to run a model comparison to evaluate the contribution of the "Experiment" (1, 2) factor, both in terms of the main effect (m2) and interactions (m3) to the model including only Block type and Waytz (m1).

m1: JEs ~ Block * Waytz

m2: JEs ~ Block * Waytz + Experiment

m3: JEs ~ Block * Waytz * Experiment

Results showed that m3 best fitted the data ($m1_{logLik}$ = -29680, $m2_{logLik}$ = -29628, $m3_{logLik}$ = -2583; χ^2 (3) = 88.29, p < 0.0001). Therefore, we report m3 (for results, please see please see paragraph 2.2.5.2, p.80).

SM.3. Cozmo functions

SM.3.1. Experiment 1

In Experiment 1, the Cozmo robot was programmed to use its lift to tap the adapted one-key button on the top of cube's surface. The functions we used to program the robot were the following:

```
import cozmo
from cozmo.util import degrees, distance_mm, radians, speed_mmps
from cozmo.objects import LightCube1Id
class TapGame:

    def __init__(self, robot: cozmo.robot.Robot):
        self.robot = robot
        self.close_to_tapping_pos = False
        self.sleeping = False

    def cozmo_go_away(self):
        self.robot.drive_straight(distance_mm(-40), speed_mmps(40)).wait_for_completed()
        self.robot.set_lift_height(0.0).wait_for_completed()

    def cozmo_go_init_pose(self):
        self.robot.set_lift_height(0.0).wait_for_completed()
        self.robot.set_head_angle(degrees(0.0)).wait_for_completed()

    def cozmo_ready_to_tap(self):
        global cozmocube
        if self.close_to_tapping_pos == False:
            if self.sleeping == True:
                self.cozmo_wakeup()
                self.sleeping = False
            self.robot.set_lift_height(1.0).wait_for_completed()
```

```
            self.robot.go_to_object(cozmocube, distance_mm(50)).wait_for_completed()
            self.close_to_tapping_pos = True

    def cozmo_tap(self):
        self.robot.play_anim_trigger(cozmo.anim.Triggers.OnSpeedtapTap, in_parallel = True)

    def cozmo_lights_on(self):
        self.robot.set_all_backpack_lights(cozmo.lights.red_light)
    def cozmo_lights_off(self):
        self.robot.set_all_backpack_lights(cozmo.lights.off_light)

    def cozmo_sleep(self):
        if self.sleeping == True:
            return
        if self.close_to_tapping_pos == True:
            self.cozmo_go_away()
            self.close_to_tapping_pos = False

self.robot.play_anim_trigger(cozmo.anim.Triggers.GoToSleepSleeping).wait_for_completed()
        self.sleeping = True

    def cozmo_wakeup(self):

self.robot.play_anim_trigger(cozmo.anim.Triggers.GoToSleepGetOut).wait_for_completed()

    def run(self):
        global cozmocube
        cozmocube = self.robot.world.get_light_cube(LightCube1Id)
        print(cozmocube)
        pass

def cozmo_program(robot: cozmo.robot.Robot):
    global game
    global cozmocube
    game = TapGame(robot)
    game.run()
    done.wait()

def cozmo_thread():
    cozmo.run_program(cozmo_program)
    global game
    global cozmocube
```

```
global game
global cozmocube
game = None
cozmocube = None
cozmo_mode = None
done = threading.Event()
cozmothread = threading.Thread(target = cozmo_thread, name = "cozmothread").start()
cozmo_mode = 'awake'
```

SM3.2. Experiment 2

In Experiment 2, the Cozmo robot was programmed to make a squeaking sound as a marker of the digital action. Different from Experiment 1, the robot did not raise the lift. The functions we used to program the robot were the following:

```
import cozmo
from cozmo.util import degrees, distance_mm, radians, speed_mmps
from cozmo.objects import LightCube1Id
class TapGame:
```

```
    class TapGame:

        def __init__(self, robot: cozmo.robot.Robot):
            self.robot = robot
            self.close_to_tapping_pos = False
            self.sleeping = False

        def cozmo_go_away(self):
            self.robot.drive_straight(distance_mm(-40), speed_mmps(50)).wait_for_completed()
            self.robot.set_lift_height(0.0).wait_for_completed()

        def cozmo_go_init_pose(self):
            self.robot.set_lift_height(0.0).wait_for_completed()
            self.robot.set_head_angle(degrees(0.0)).wait_for_completed()

        def cozmo_ready_to_tap(self):
            global cozmocube
            if self.close_to_tapping_pos == False:
                if self.sleeping == True:
                    self.cozmo_wakeup()
                    self.sleeping = False
```

```
            self.robot.go_to_object(cozmocube, distance_mm(40)).wait_for_completed()
            self.close_to_tapping_pos = True

    def cozmo_tap(self):
        self.robot.play_anim_trigger(cozmo.anim.Triggers.CodeLabWhee2, in_parallel =
True)
        pass

    def cozmo_lights_on(self):
        self.robot.set_all_backpack_lights(cozmo.lights.red_light)
    def cozmo_lights_off(self):
        self.robot.set_all_backpack_lights(cozmo.lights.off_light)

    def cozmo_sleep(self):
        if self.sleeping == True:
            return
        if self.close_to_tapping_pos == True:
            self.cozmo_go_away()
            self.close_to_tapping_pos = False

self.robot.play_anim_trigger(cozmo.anim.Triggers.GoToSleepSleeping).wait_for_compl
eted()
        self.sleeping = True

    def cozmo_wakeup(self):

self.robot.play_anim_trigger(cozmo.anim.Triggers.GoToSleepGetOut).wait_for_complet
ed()

    def run(self):
        global cozmocube
        cozmocube = self.robot.world.get_light_cube(LightCube1Id)
        print(cozmocube)
        pass

def cozmo_program(robot: cozmo.robot.Robot):
    global game
    global cozmocube
    game = TapGame(robot)
    game.run()
    done.wait()

def cozmo_thread():
    cozmo.run_program(cozmo_program)
    global game
    global cozmocube
```

```
        global game
        global cozmocube
        game = None
        cozmocube = None
        cozmo_mode = None
        done = threading.Event()
        cozmothread = threading.Thread(target = cozmo_thread, name = "cozmothread").start()
cozmo_mode = 'awake'
```

References

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Miles, J. (2014). R squared, adjusted R squared. *Wiley StatsRef: Statistics Reference Online*.

Supplementary Materials

Publication III: Human-likeness and attribution of intentionality predict vicarious sense of agency over humanoid robot actions

Cecilia Roselli[1,2], Francesca Ciardo[1], Davide De Tommaso[1], and Agnieszka Wykowska[1]

[1] Social Cognition in Human Robot Interaction, Fondazione Istituto Italiano di Tecnologia, Center for Human Technologies, via Enrico Melen 83, Genova, Italy

[2] DIBRIS, Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi, Via all'Opera Pia 13, Genova, Italy

This section includes:

SM.1. Latency measurement

Aim. The aim was to determine whether the onset of the auditory tone outcome, as signaled by the PTB library in Psychopy v.2021.2.0 (Peirce, 2007), matched the actual physical event produced by the audio speakers. This way, we wanted to ensure that the action-tone interval was set to 250 ms as in the classical IB paradigm based on the Libet clock method (e.g., Haggard, Clark, & Kalogeras, 2002). In order to obtain this latency, we measured, over many trials, time intervals between when the tone was supposed to start playing (in the code) and the onset (i.e., the start of rising edge) of the tone signal recorded by a dynamic microphone.

Equipment. Our target system subject to measure consists of the pc running code of the experiment (i.e., a workstation equipped with a 27' inches display, resolution 1920x1200) and a set of audio speakers. Our measuring system was a BioSemi EEG system used for collecting two input signals. The first one is the analog signal from a dynamic microphone placed in front of the audio speakers. The second one is a TTL signal from a Brain Product Trigger Box connected to the pc through a USB port.

Procedure. The two input signals are recorded for the entire duration of the test using the BioSemi EEG system and later analyzed as time series using MATLAB (MATLAB version R2021b). The first input is the audio signal that contains information about the actual onset/offset of the auditory stimuli played during the experiment. The second input is a TTL signal generated inside the code for triggering the event where the auditory tone was supposed to start (onset) and end (offset). The onset and offset of the auditory tone were marked with triggers S100 and S200, respectively. The test was performed both in the Solo (i.e., when participants performed the task alone) and in the Social Context (i.e., when participants performed the task with the iCub robot). Then, the signal related to the auditory tone was imported in EEGLab v.2020.0 (Delorme & Makeig, 2004). All

channels were removed but ERGO1, i.e., the microphone channel. We also added a second channel, i.e., MIC-RECT, obtained by applying (1) a Z-transformation (subtracting the mean and dividing by the SD of ERGO1), and (2) a signal rectification, (calculating the absolute value of ERGO1).

The subsequent step was to identify the time points in which the auditory tone on the MIC-RECT channel overcame its background noise, i.e., *significant activations*. To this aim, with the signal digitized at a 2048 Hz sampling rate, as a baseline we selected a period of 1000 ms, i.e., from 0 to 1000 ms after the recording of the auditory tone started. Then, we calculated Mean and SD of the MIC-RECT signal in this 1000 ms period.

Notably, time points that deviated Mean ± 3 SD were considered as *significant activations*. They were observed in the time interval between S100 and S200 triggers, i.e., between the onset and the offset of the auditory tone. *Significant activations* occurring outside the S100-S200 time interval were not considered for further analyses. Then, for each couple of S100-S200 time interval, we considered the latency of both S100 and S200 triggers. Specifically, we considered all the latencies (1) having *significant activations* (vt), and (2) occurring in this time interval. Then, the minimum latency as the start of the rising edge (tmin = min(vt)), and then the difference (dt = t100 – tmin) representing the processed latency, for both the Solo and the Social Context.

Notably, in raw data we found 1 outlier due to spurious *significant activations*. Thus, we applied a 20 ms threshold on the differences to exclude these outliers and to obtain more reliable estimations.

Results. In the Solo Context, i.e., when participants performed the task alone, results showed that the auditory tone was slightly delayed compared to the onset of the tone [M = -36.47; SD = 0.28;

CI_{Mean} = (-37.11; -36.13)] . Similar results were found in the Social Context, i.e., when participants performed the task with iCub [M = -36.55; SD = 0.33; CI_{Mean} = (-37.11; -36.13)].

Conclusions. Our results showed that, on average, there is a 36 ms delay between when the command to play the auditory tone is sent and the actual onset of the tone. Therefore, considering the little variability of the latencies, the action-tone time interval was modified accordingly in Psychopy to ensure that it corresponded to 250 ms.

SM.2. Robot integration

The iCub humanoid robot iCub (Metta et al., 2010) was connected to the experimental pc using a peer-to-peer Ethernet connection. This way, we created a network shared between the experimental pc and the iCub robot. In the experimental machine was installed all the software needed for controlling the robot, namely YARP [y] and all its basic modules. Therefore, for controlling the robot from the experimental script in PsychoPy, we used the YARP Python wrappers in Psychopy (see Metta, Fitzpatrick, & Natale, 2006 for more information). Then, we used predefined postures of the robot in joint space and the standard YARP position controller (*IPositionController*) to make the robot tapping. In such a way, we ensured a high accuracy on repeatability of the same movements across trials rather than using a kinematic controller in task space.

References

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods*, *134*(1), 9-21.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience, 5*(4), 382-385.

MATLAB and Statistics Toolbox Release 2012b, The MathWorks, Inc., Natick, Massachusetts, United States.

Metta, G., Fitzpatrick, P., & Natale, L. (2006). YARP: yet another robot platform. *International Journal of Advanced Robotic Systems*, 3(1), 8.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8-13.

REFERENCES

References

Balconi, M. (2010). The sense of agency in psychology and neuropsychology. In *Neuropsychology of the Sense of Agency*, 3-22. Springer, Milano.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37(2), 122.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50*(2), 248-287.

Barlas, Z., & Kopp, S. (2018). Action choice and outcome congruency independently affect intentional binding and feeling of control judgments. *Frontiers in Human Neuroscience*, 12, 137.

Beyer, F., Sidarus, N., Bonicalzi, S., & Haggard, P. (2017). Beyond self-serving bias: diffusion of responsibility reduces sense of agency and outcome monitoring. *Social Cognitive and Affective Neuroscience, 12*(1), 138-145.

Beyer, F., Sidarus, N., Fleming, S., & Haggard, P. (2018). Losing control in social situations: how the presence of others affects neural processes related to sense of agency. *eneuro*, *5*(1).

Blakemore, S. J., Wolpert, D. M., & Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nature Neuroscience*, 1(7), 635-640.

Blakemore, S. J., Wolpert, D. M., & Frith, C. D. (1999). The cerebellum contributes to somatosensory cortical activity during self-produced tactile stimulation. *Neuroimage*, 10(4), 448-459.

Blakemore, S. J., Wolpert, D., & Frith, C. (2000). Why can't you tickle yourself?. *Neuroreport*, 11(11), R11-R16.

References

Blakemore, S. J., Wolpert, D. M., & Frith, C. D. (2002). Abnormalities in the awareness of action. *Trends in Cognitive Sciences*, 6(6), 237-242.

Buehner, M. J. (2015). Awareness of voluntary and involuntary causal actions and their outcomes. *Psychology of Consciousness: Theory, Research, and Practice*, 2(3), 237.

Chambon, V., & Haggard, P. (2012). Sense of control depends on fluency of action selection, not motor performance. *Cognition*, 125(3), 441-451.

Chambon, V., & Haggard, P. (2013). Premotor or Ideomotor: How Does the Experience of Action Come About?. In W. Prinz, M. Beisert, & A. Herwig (Eds.), *Action science: Foundations of an emerging discipline*, 359 –380.

Christensen, J. F., Yoshie, M., Di Costa, S., & Haggard, P. (2016). Emotional valence, sense of agency and responsibility: A study using intentional binding. *Consciousness and Cognition*, 43, 1-10.

Ciardo F., De Tommaso D., Beyer F., Wykowska A. (2018) Reduced Sense of Agency in Human-Robot Interaction. In: Ge S. et al. (eds) Social Robotics. ICSR 2018. *Lecture Notes in Computer Science*, 11357, 441-450. Springer, Cham.

Ciardo, F., Beyer, F., De Tommaso, D., & Wykowska, A. (2020). Attribution of intentional agency towards robots reduces one's own sense of agency. *Cognition*, 194, 104109.

Ciardo, F., & Wykowska, A. (2018). Response coordination emerges in cooperative but not competitive joint task. *Frontiers in Psychology*, 9, 1919.

References

Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J., & Jeannerod, M. (1997). Looking for the agent: an investigation into consciousness of action and self-consciousness in schizophrenic patients. *Cognition*, 65(1), 71-86.

David, N., Newen, A., & Vogeley, K. (2008). The "sense of agency" and its underlying cognitive and neural mechanisms. *Consciousness and Cognition*, 17(2), 523-534.

De Vignemont, F., & Fourneret, P. (2004). The sense of agency: A philosophical and empirical review of the "Who" system. *Consciousness and Cognition*, 13(1), 1-19.

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*. MIT press, Cambridge, MA.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research*, 91(1), 176-180.

Dolk, T., Hommel, B., Colzato, L. S., Schütz-Bosbach, S., Prinz, W., & Liepelt, R. (2014). The joint Simon effect: a review and theoretical integration. *Frontiers in Psychology*, 5, 974.

Engbert, K., Wohlschläger, A., Thomas, R., & Haggard, P. (2007). Agency, subjective time, and other minds. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1261.

Farrer, C., Franck, N., Georgieff, N., Frith, C. D., Decety, J., & Jeannerod, M. (2003). Modulating the experience of agency: a positron emission tomography study. *Neuroimage*, 18(2), 324-333.

Farrer, C., & Frith, C. D. (2002). Experiencing oneself vs another person as being the cause of an action: the neural correlates of the experience of agency. *Neuroimage*, 15(3), 596-603.

References

Ferraro, L., Iani, C., Mariani, M., Milanese, N., & Rubichi, S. (2011). Facilitation and interference components in the joint Simon task. *Experimental Brain Research*, 211(3-4), 337.

Fink, G. R., Marshall, J. C., Halligan, P. W., Frith, C. D., Driver, J., Frackowiak, R. S., & Dolan, R. J. (1999). The neural consequences of conflict between intention and the senses. *Brain*, 122(3), 497-512.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science*, 308(5722), 662-667.

Fotopoulou, A., Tsakiris, M., Haggard, P., Vagopoulou, A., Rudd, A., & Kopelman, M. (2008). The role of motor intention in motor awareness: an experimental study on anosognosia for hemiplegia. *Brain*, 131(12), 3432-3442.

Franck, N., Farrer, C., Georgieff, N., Marie-Cardine, M., Daléry, J., d'Amato, T., & Jeannerod, M. (2001). Defective recognition of one's own actions in patients with schizophrenia. American *Journal of Psychiatry*, 158(3), 454-459.

Fried, I., Katz, A., McCarthy, G., Sass, K. J., Williamson, P., Spencer, S. S., & Spencer, D. D. (1991). Functional organization of human supplementary motor cortex studied by electrical stimulation. *Journal of Neuroscience*, 11(11), 3656-3666.

Frith, C. (2002). Attention to action and awareness of other minds. *Consciousness and Cognition*, 11(4), 481-487.

Frith, C. (2005). The self in action: Lessons from delusions of control. *Consciousness and Cognition*, 14(4), 752-770.

Frith, C. D. (2014). Action, agency and responsibility. *Neuropsychologia*, 55, 137-142.

References

Frith, C. D., Blakemore, S. J., & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1404), 1771-1788.

Frith, C. D., & Done, D. J. (1989). Experiences of alien control in schizophrenia reflect a disorder in the central monitoring of action. *Psychological Medicine*, 19(2), 359-363.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Sciences*, 4(1), 14-21.

Gallagher, S. (2004). Neurocognitive models of schizophrenia: a neurophenomenological critique. *Psychopathology*, 37(1), 8-19.

Georgieff, N., & Jeannerod, M. (1998). Beyond consciousness of external reality: a "who" system for consciousness of action and self-consciousness. *Consciousness and Cognition*, 7(3), 465-477.

Gentsch, A., & Schütz-Bosbach, S. (2011). I did it: unconscious expectation of sensory consequences modulates the experience of self-agency and its functional signature. *Journal of Cognitive Neuroscience*, 23(12), 3817-3828.

Gentsch, A., Schütz-Bosbach, S., Endrass, T., & Kathmann, N. (2012). Dysfunctional forward model mechanisms and aberrant sense of agency in obsessive-compulsive disorder. *Biological Psychiatry*, 71(7), 652-659.

Haering, C., & Kiesel, A. (2014). Intentional Binding is independent of the validity of the action effect's identity. *Acta Psychologica*, 152, 109-119.

Haggard, P. (2008). Human volition: towards a neuroscience of will. *Nature Reviews Neuroscience*, 9(12), 934-946.

References

Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews Neuroscience*, 18(4), 196-207.

Haggard, P., & Clark, S. (2003). Intentional action: Conscious experience and neural prediction. *Consciousness and Cognition*, 12(4), 695-707.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, 5(4), 382-385.

Haggard, P., & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research*, 126(1), 128-133.

Haggard, P., Martin, F., Taylor-Clarke, M., Jeannerod, M., & Franck, N. (2003). Awareness of action in schizophrenia. *Neuroreport*, 14(7), 1081-1085.

Hinz, N. A., Ciardo, F., & Wykowska, A. (2021). ERP markers of action planning and outcome monitoring in human–robot interaction. *Acta Psychologica*, 212, 103216.

Hoffman, R. E. (1986). Verbal hallucinations and language production processes in schizophrenia. *Behavioral and Brain Sciences*, 9(3), 503-517.

Hommel, B. (2015). Action control and the sense of agency. In *The Sense of Agency*, eds Haggard, P., and Eitam, B. New York: Oxford University Press, 307–326.

Hommel, B., Alonso, D., & Fuentes, L. (2003). Acquisition and generalization of action effects. *Visual Cognition*, 10(8), 965-986.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, 24(5), 849-878.

References

Hume, D. (1739). *A treatise of human nature*. Oxford: Clarendon Press.

Jeannerod, M. (2009). The sense of agency and its disturbances in schizophrenia: a reappraisal. *Experimental Brain Research*, 192(3), 527-532.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9(6), 718-727.

Khalighinejad, N., Bahrami, B., Caspar, E. A., & Haggard, P. (2016). Social transmission of experience of agency: an experimental study. *Frontiers in Psychology*, 7, 1315.

Kühn, S., Brass, M., & Haggard, P. (2013). Feeling in control: Neural correlates of experience of agency. *Cortex*, 49(7), 1935-19429.

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1993). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). In *Neurophysiology of consciousness*, 249-268. Birkhäuser, Boston, MA.

Limerick, H., Coyle, D., & Moore, J. W. (2014). The experience of agency in human-computer interactions: a review. *Frontiers in Human Neuroscience*, 8, 643.

MacDonald, P. A., & Paus, T. (2003). The role of parietal cortex in awareness of self-generated movements: a transcranial magnetic stimulation study. *Cerebral Cortex*, 13(9), 962-967.

Marchesi, S., Ghiglino, D., Ciardo, F., Pérez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots?. *Frontiers in Psychology*, 10, 450.

Massen, C., & Prinz, W. (2009). Movements, actions and tool-use actions: an ideomotor approach to imitation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1528), 2349-2358.

References

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*, 50-56.

Miele, D. B., Wager, T. D., Mitchell, J. P., & Metcalfe, J. (2011). Dissociating neural correlates of action monitoring and metacognition of agency. *Journal of Cognitive Neuroscience*, 23(11), 3620-36369.

Moore, J. W. (2016). What is the sense of agency and why does it matter?. *Frontiers in Psychology*, 7, 1272.

Moore, J. W., & Fletcher, P. C. (2012). Sense of agency in health and disease: a review of cue integration approaches. *Consciousness and Cognition*, 21(1), 59-68.

Moore, J., & Haggard, P. (2008). Awareness of action: Inference and prediction. *Consciousness and Cognition*, 17(1), 136-144.

Moore, J. W., & Obhi, S. S. (2012). Intentional binding and the sense of agency: a review. *Consciousness and Cognition*, 21(1), 546-561.

Moore, J. W., Ruge, D., Wenke, D., Rothwell, J., & Haggard, P. (2010). Disrupting the experience of control in the human brain: pre-supplementary motor area contributes to the sense of agency. *Proceedings of the Royal Society B: Biological Sciences*, 277(1693), 2503-2509.

Moore, J. W., Wegner, D. M., & Haggard, P. (2009). Modulating the sense of agency with external cues. *Consciousness and Cognition*, 18(4), 1056-1064.

Muth, F. V., Wirth, R., & Kunde, W. (2021). Temporal binding past the Libet clock: Testing design factors for an auditory timer. *Behavior Research Methods*, 53(3), 1322-13419.

References

Nahab, F. B., Kundu, P., Gallea, C., Kakareka, J., Pursley, R., Pohida, T., ... & Hallett, M. (2011). The neural processes underlying self-agency. *Cerebral Cortex*, 21(1), 48-55.

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, 24(4), 375-425.

Nichols, S. (2011). Experimental philosophy and the problem of free will. *Science*, 331(6023), 1401-1403.

Nielsen, T. I. (1963). Volition: A new experimental approach. *Scandinavian Journal of Psychology*, 4(1), 225-230.

Obhi, S. S., & Hall, P. (2011). Sense of agency in joint action: Influence of human and computer co-actors. *Experimental brain research*, *211*(3-4), 663-670.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107(1), 179-217.

Pacherie, E. (2012). The Phenomenology of Joint Action: Self-Agency vs. Joint-Agency. In A. Seemann (Ed.), *Joint attention: new developments*, 343–389. Cambridge MA: MIT Press.

Pacherie, E., & Dokic, J. (2006). From mirror neurons to joint actions. *Cognitive Systems Research*, 7(2-3), 101-112.

Pérez-Osorio, J., & Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, 33(3), 369-395.

Picard, N., & Strick, P. L. (2001). Imaging the premotor areas. *Current Opinion in Neurobiology*, 11(6), 663-672.

References

Pockett, S., & Miller, A. (2007). The rotating spot method of timing subjective events. *Consciousness and Cognition*, 16, 241–254.

Poonian, S. K., & Cunnington, R. (2013). Intentional binding in self-made and observed actions. *Experimental Brain Research*, 229(3), 419-427.

Poonian, S. K., McFadyen, J., Ogden, J., & Cunnington, R. (2015). Implicit agency in observed actions: evidence for N1 suppression of tones caused by self-made and observed actions. *Journal of Cognitive Neuroscience*, 27(4), 752-764.

Preston, C., & Newport, R. (2010). Self-denial and the role of intentions in the attribution of agency. *Consciousness and Cognition*, 19(4), 986-998.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology*, 9(2), 129-154.

Ramnani, N., & Miall, R. C. (2004). A system in the human brain for predicting the actions of others. *Nature neuroscience*, *7*(1), 85-90.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169-192.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2(9), 661-670.

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nature reviews neuroscience*, *11*(4), 264-274.

References

Sahaï, A., Desantis, A., Grynszpan, O., Pacherie, E., & Berberian, B. (2019). Action co-representation and the sense of agency during a joint Simon task: Comparing human and machine co-agents. *Consciousness and Cognition*, 67, 44-55.

Sahaï, A., Pacherie, E., Grynszpan, O., & Berberian, B. (2017). Predictive mechanisms are not involved the same way during human-human vs. human-machine interactions: a review. *Frontiers in Neurorobotics*, 11, 52.

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: just like one's own?. *Cognition*, 88(3), B11-B21.

Shneiderman, B., & Plaisant, C. (2010). *Designing the user interface: Strategies for effective human-computer interaction*. Boston: Addison-Wesley.

Sirigu, A., Daprati, E., Pradat-Diehl, P., Franck, N., & Jeannerod, M. (1999). Perception of self-generated movement following left parietal lesion. *Brain*, 122(10), 1867-1874.

Slachevsky, A., Pillon, B., Fourneret, P., Pradat-Diehl, P., Jeannerod, M., & Dubois, B. (2001). Preserved adjustment but impaired awareness in a sensory-motor conflict following prefrontal lesions. *Journal of Cognitive Neuroscience*, 13(3), 332-340.

Springer, A., Hamilton, A.F.C., & Cross, E.S. (2012), Simulating and predicting others' actions. *Psychological Research*, 76, 383–387.

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

References

Synofzik, M., Thier, P., Leube, D. T., Schlotterbeck, P., & Lindner, A. (2010). Misattributions of agency in schizophrenia are based on imprecise predictions about the sensory consequences of one's actions. *Brain*, 133(1), 262-271.

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Consciousness and Cognition*, 17(1), 219-239.

Synofzik, M., Vosgerau, G., & Voss, M. (2013). The experience of agency: an interplay between prediction and postdiction. *Frontiers in Psychology*, 4, 127.

Tanaka, T., Matsumoto, T., Hayashi, S., Takagi, S., & Kawabata, H. (2019). What makes action and outcome temporally close to each other: A systematic review and meta-analysis of temporal binding. *Timing and Time Perception*, 7(3), 189-218.

Tsai, C. C., Kuo, W. J., Hung, D. L., & Tzeng, O. J. (2008). Action co-representation is tuned to other humans. *Journal of Cognitive Neuroscience*, 20(11), 2015-2024.

Tsakiris, M., Haggard, P., Franck, N., Mainy, N., & Sirigu, A. (2005). A specific role for efferent information in self-recognition. *Cognition*, 96(3), 215-231.

Tsakiris, M., Hesse, M. D., Boy, C., Haggard, P., & Fink, G. R. (2007). Neural signatures of body ownership: a sensory network for bodily self-consciousness. *Cerebral Cortex*, 17(10), 2235-2244.

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001). I know what you are doing: A neurophysiological study. *Neuron*, 31(1), 155-165.

Vilaza, G. N., Haselager, W. F. F., Campos, A. M., & Vuurpijl, L. (2014). Using games to investigate sense of agency and attribution of responsibility. *Proceedings of the 2014 SBGames* (SBgames 2014), SBC, Porte Alegre.

References

Voss, M., Moore, J., Hauser, M., Gallinat, J., Heinz, A., & Haggard, P. (2010). Altered awareness of action in schizophrenia: a specific deficit in predicting action consequences. *Brain*, 133(10), 3104-3112.

Wegner, D. M. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Wegner, D. M., Sparrow, B., & Winerman, L. (2004). Vicarious agency: experiencing control over the movements of others. *Journal of Personality and Social Psychology*, 86(6), 838.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, 54(7), 480.

Weiller, C., Jüptner, M., Fellows, S., Rijntjes, M., Leonhardt, G., Kiebel, S., ... & Thilmann, A. F. (1996). Brain representation of active and passive movements. *Neuroimage*, 4(2), 105-110.

Wen, T., & Hsieh, S. (2015). Neuroimaging of the joint Simon effect with believed biological and non-biological co-actors. *Frontiers in Human Neuroscience*, 9, 483.

Wohlschläger, A., Haggard, P., Gesierich, B., & Prinz, W. (2003). The perceived onset time of self-and other-generated actions. *Psychological Science*, 14(6), 586-591.

Wolpe, N., Haggard, P., Siebner, H. R., & Rowe, J. B. (2013). Cue integration and the perception of action in intentional binding. *Experimental brain research*, *229*(3), 467-474.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, 269(5232), 1880-1882.

Yamaguchi, M., Wall, H. J., & Hommel, B. (2018). Sharing tasks or sharing actions? Evidence from the joint Simon task. *Psychological Research*, 82(2), 385-394.

References

Zapparoli, L., Seghezzi, S., Zirone, E., Guidali, G., Tettamanti, M., Banfi, G., ... & Paulesu, E. (2020). How the effects of actions become our own. *Science Advances*, 6(27), eaay8301.