Intentions with actions:

The role of intentionality attribution on the vicarious sense of agency in human-robot interaction

Cecilia Roselli [a, b], Francesca Ciardo [a], and Agnieszka Wykowska [a]*

[a] Social Cognition in Human Robot Interaction Unit, Fondazione Istituto Italiano di Tecnologia, Center for Human Technologies, via Enrico Melen 83, 16152 Genoa, Italy

[b] DIBRIS, Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi, via all'Opera Pia 12, 16145 Genoa, Italy

* Corresponding author: Agnieszka Wykowska

Italian Institute of Technology

Center for Human Technologies

Via Enrico Melen 83

16152 Genova (Italy)

E-mail: Agnieszka.Wykowska@iit.it

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.

1. 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

[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.

they represent. When observing others' actions, although to a lesser extent than during self-action 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 &

[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.

Clark, 2003; Haggard, Clark, & Kalogeras, 2002; Tsakiris & Haggard, 2003). In line with that, 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. Aim of the study

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.

3. 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.

3.1. Experiment 1

3.1.1. 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 experiment 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 **Fig.1**).

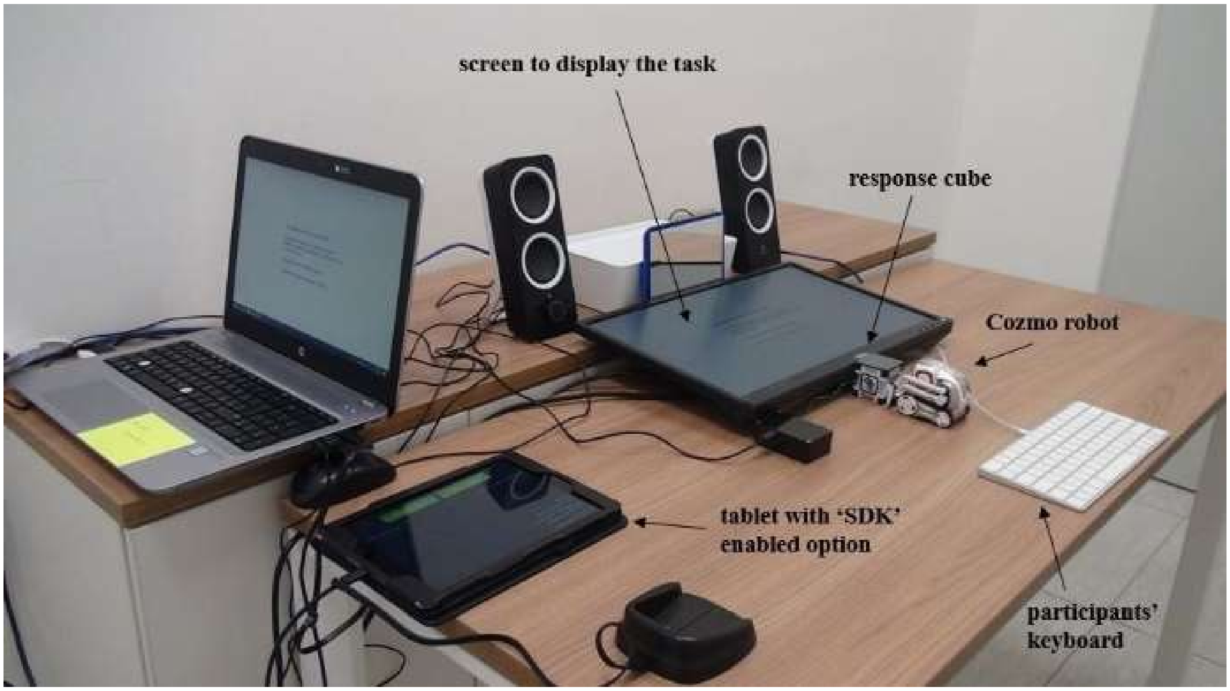

Fig.1. Experimental setup.

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. 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 (**Fig.2**).

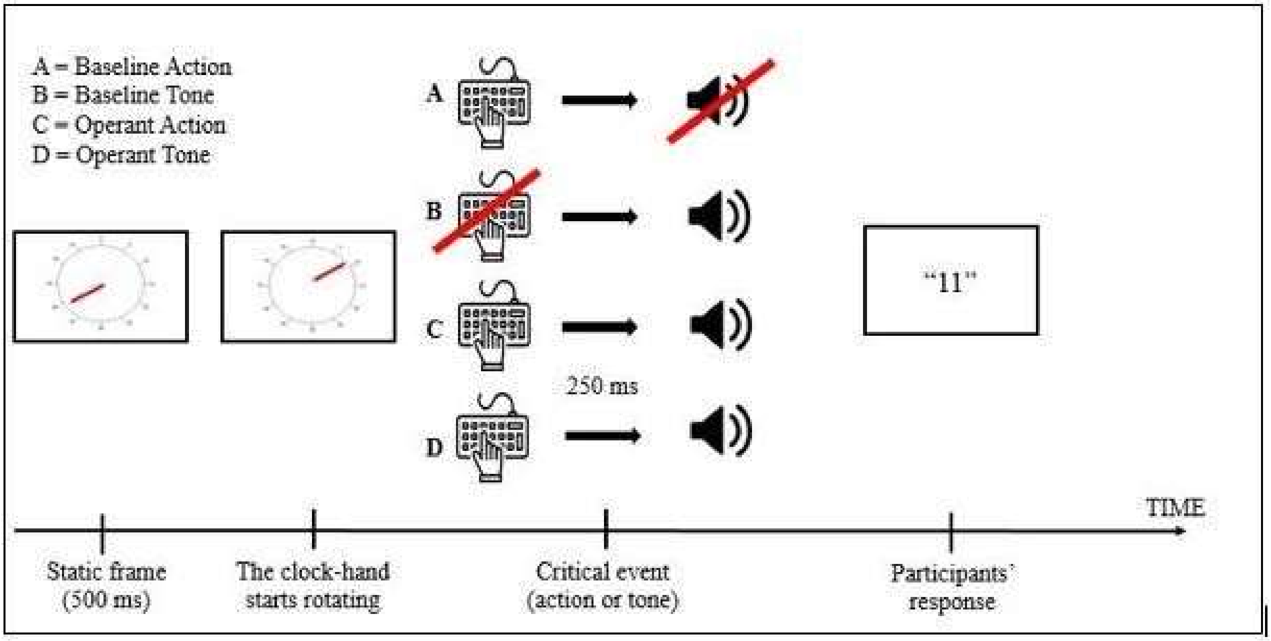

Fig.2. Trial sequence. Participants were instructed to observe the rotating clock hand displayed on the screen and to report its position at the occurrence of the critical event (action or tone). Note that, in Baseline blocks, only one event occurred, either action (**A**) or tone (**B**); in Operant 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.

3.1.2. 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 (**Fig. 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 **Fig. 3B).** On the contrary, when they are asked to judge the occurrence of the tone (**Fig. 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 **Fig. 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 **Fig. 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 (**Fig. 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).

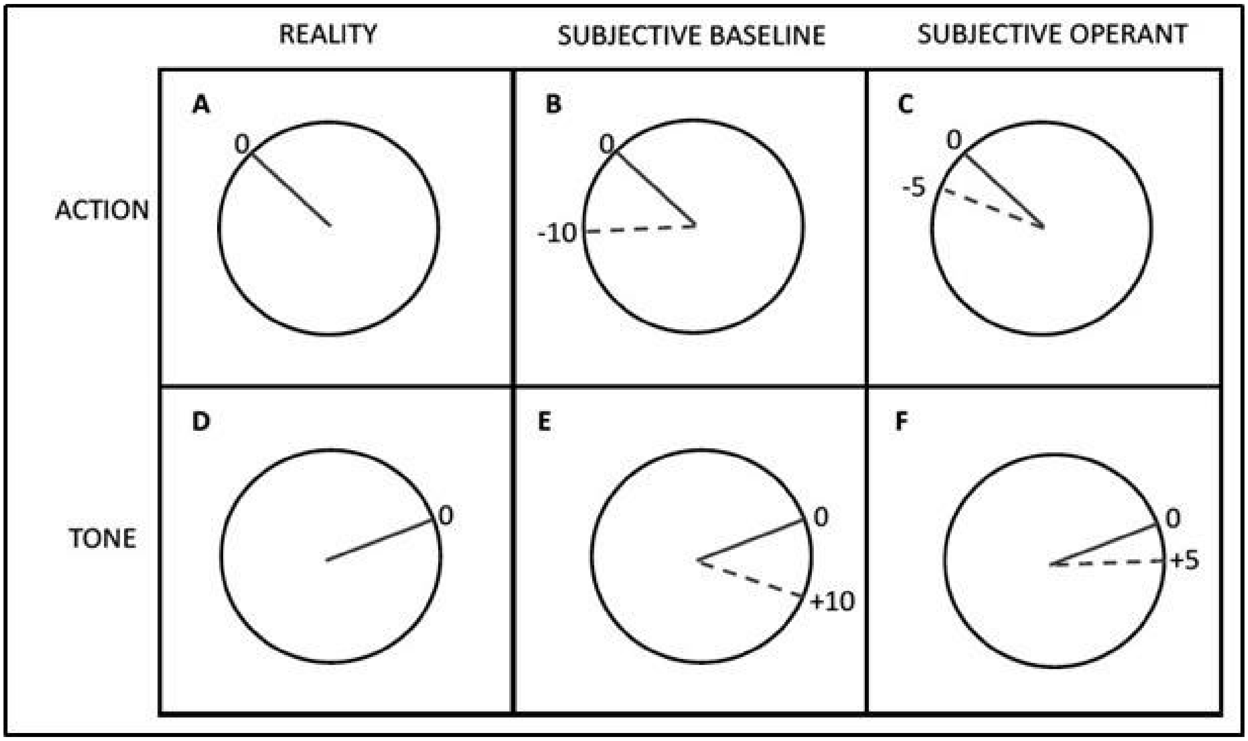

Fig.3. Visual representation of the over- and under-estimation of time of occurrence of tone and action events in an Intentional Binding task with a schematic representation of the clock stimulus, together with the clock hands. **Upper row**: action events. **Lower row**: tone events. **A**: time of 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 details specific directions of the IB effect.

3.1.3. Results

Action events. When the critical event was the action, results showed a main effect of Context [$\beta = -26.06$, 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 [$\beta = 39.1$, SE = 2.55, $t_{(5082)} = 15.29$, $p < 0.0001$, CI = (34; 44.1); ($M_{\text{Individual}} = -39.1$, $SE_{\text{Individual}} = 8.76$; $M_{\text{Social}} = -78.1$, $SE_{\text{Social}} = 8.79$)]. Moreover, a main effect of Block type emerged [$\beta = 27.69$, 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 [$\beta = -14.7$, SE = 2.55, $t_{(5081)} = -5.77$, $p < 0.0001$, CI = (-19.7; -9.71); ($M_{\text{Operant}}= -51.2$, $SE_{\text{Operant}} = 8.78$; $M_{\text{Baseline}}= -65.9$, $SE_{\text{Baseline}}= 8.77$)]. Finally, a significant Context * Block type interaction emerged [$\beta = -25.98$, 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 **Fig.4**).

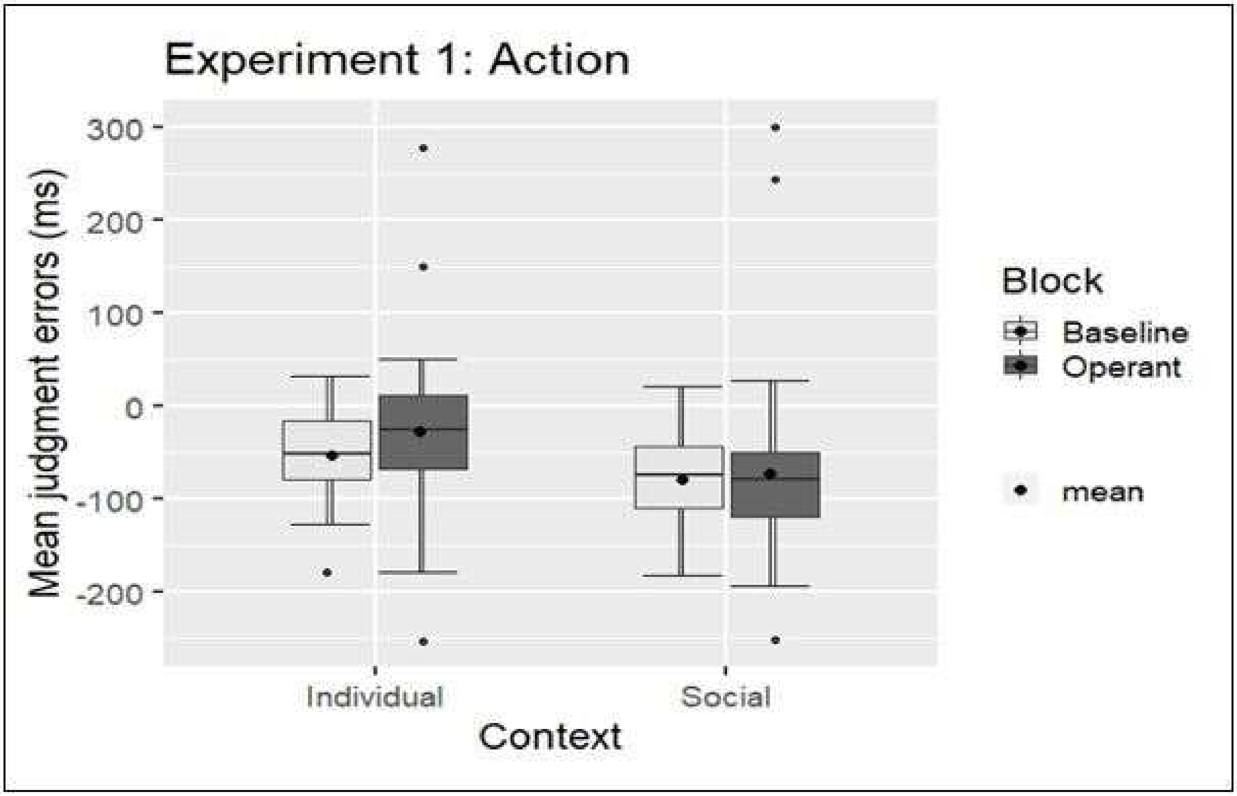

Fig.4. Experiment 1: Mean JEs plotted as a function of Context (Individual, Social) and Block type (Baseline, Operant) when the critical event to judge was the action. Error bars represent standard errors.

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 **Fig.5**).

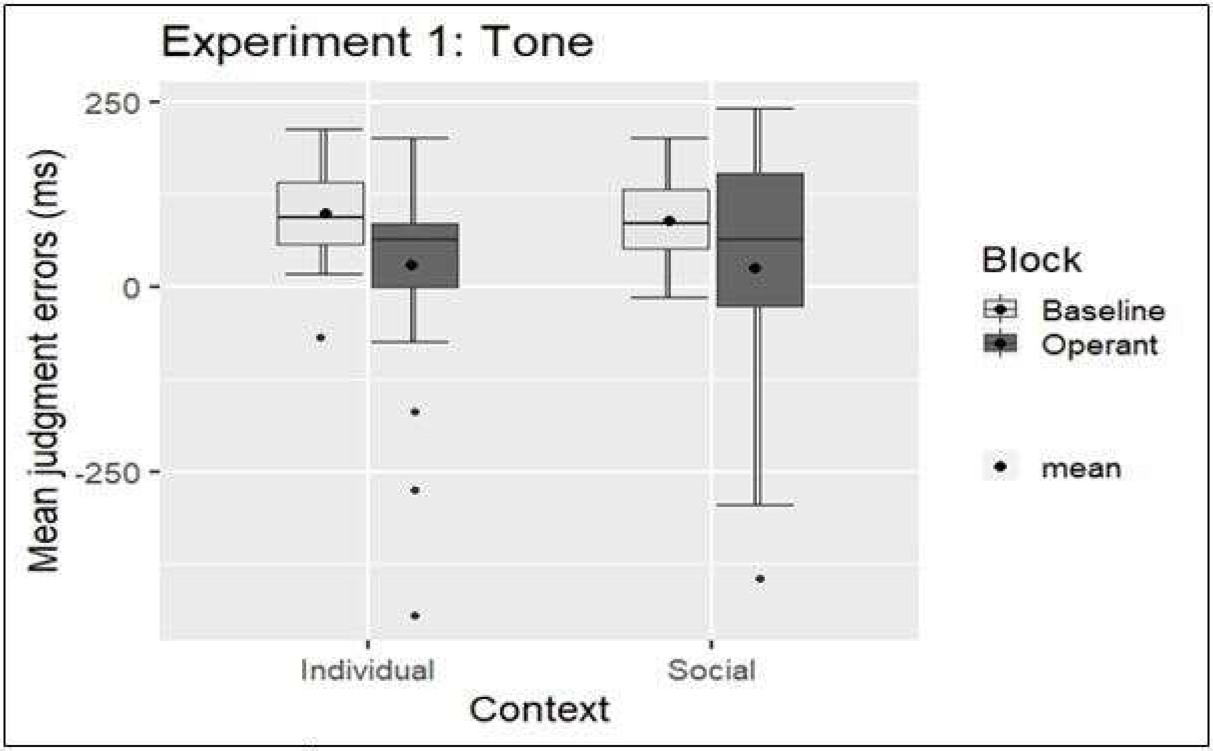

Fig.5. Experiment 1: Mean JEs plotted as a function of Context (Individual, Social) and Block type (Baseline, Operant) when the critical event to judge was the tone. Error bars represent standard errors.

3.2. Experiment 2

3.2.1. 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. 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.

3.2.2. 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).

3.2.3. 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

$_{\text{Baseline}}$ = 115, SE $_{\text{Baseline}}$ = 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 $_{\text{Operant}}$ = 0.17, SE $_{\text{Operant}}$ = 12; M $_{\text{Baseline}}$ = -46.8, SE $_{\text{Baseline}}$ = 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 $_{\text{Operant}}$ = 224.68, SE $_{\text{Operant}}$ = 12.1; M $_{\text{Baseline}}$ = 277.41, SE $_{\text{Baseline}}$ = 12.1)] (see **Fig.6**).

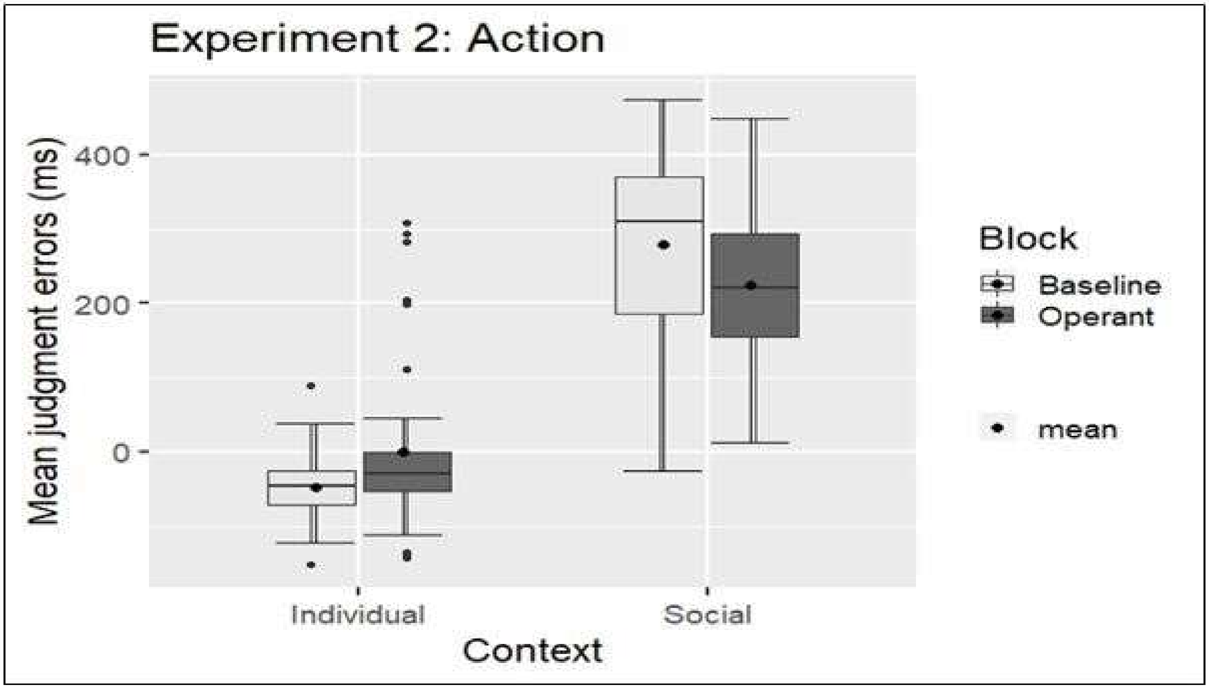

Fig.6. Experiment 2: Mean JEs plotted as a function of Context (Individual, Social) and Block type (Baseline, Operant) when the critical event to judge was the action. Error bars represent standard errors.

Tone events. When the critical event to judge was the tone, results showed a significant main effect of Context [$\beta = -10.41$, 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 [$\beta = -17.7$, SE = 2.72, $t_{(5322)} = -6.51$, $p < 0.0001$, CI = (-23; -12.4); ($M_{\text{Individual}} = 59.6$, $SE_{\text{Individual}} = 9.17$; $M_{\text{Social}} = 77.3$, $SE_{\text{Social}} = 9.18$)]. Moreover, a significant main effect of Block type emerged [$\beta = -49.33$, 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 [$\beta = 21.2$, SE = 2.72, $t_{(5322)} = 7.79$, $p < 0.0001$, CI = (15.9; -26.5); ($M_{\text{Operant}} = 57.8$, $SE_{\text{Operant}} = 9.18$; $M_{\text{Baseline}} = 79$, $SE_{\text{Baseline}} = 9.17$)]. Finally, the Context * Block type interaction was significant [$\beta = 56.26$, 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 [$\beta = 49.33$, SE = 3.78, $t_{(5322)} = 13.04$, $p < 0.0001$, CI = (39.35; 59.31); ($M_{\text{Operant}} = 34.9$, $SE_{\text{Operant}} = 9.36$; $M_{\text{Baseline}} = 84.2$, $SE_{\text{Baseline}} = 9.36$)]. However, this was not true for the Social Context, where no significant differences emerged between JEs in Operant and in the corresponding Baseline block [$\beta = -6.93$, SE = 3.91, $t_{(5322)} = -1.77$, $p = 0.45$, CI = (-17.2; 3.38); ($M_{\text{Operant}} = 80.7$, $SE_{\text{Operant}} = 9.41$; $M_{\text{Baseline}} = 73.8$, $SE_{\text{Baseline}} = 9.36$)] (see **Fig.7**).

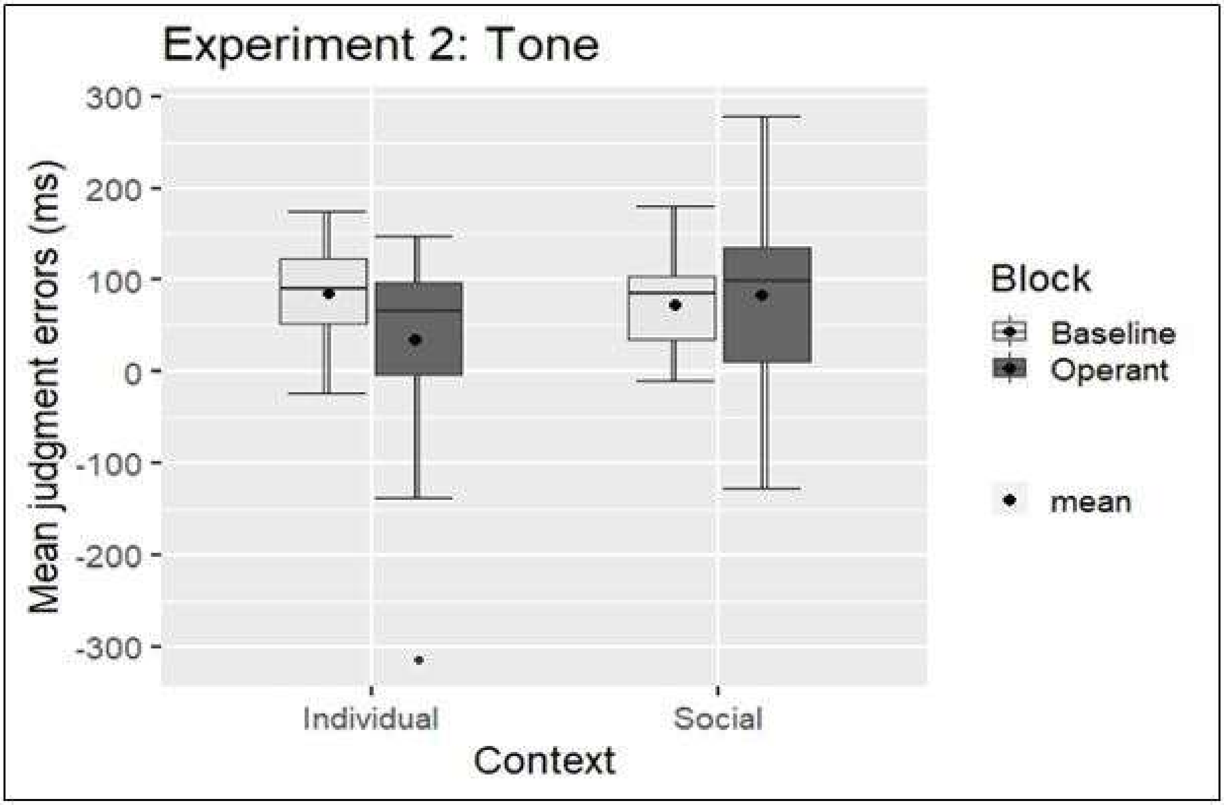

Fig.7. Experiment 2: Mean JEs plotted as a function of Context (Individual, Social) and Block type (Baseline, Operant) when the critical event to judge was the tone. Error bars represent standard errors.

3.3. Discussion of the role of action representation in 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, 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.

4. The role of adopting Intentional Stance in 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.

4.1. 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, 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.

4.2. 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); ($M_{\text{Experiment 1}}$= -79.5, $SE_{\text{Experiment 1}}$ = 15.2; $M_{\text{Experiment 2}}$ = 250, $SE_{\text{Experiment 2}}$ = 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 [$\beta = -60.38$, 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 [$\beta = 20.12$, 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 [$\beta = -12.85$, SE = 4.02, $t_{(2378)} = -3.19$, $p = 0.001$, CI = (-20.75; -4.95)] but only marginally in Baseline blocks [$\beta = 5.15$, 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 [$\beta = -38.2$, SE = 14.62, $t_{(2450.16)} = - 2.61$, $p = 0.009$, CI = (-66.87; -9.54)] but not a Block type * Waytz score interaction [$\beta = -5.12$, SE = 4.83, $t_{(2450.14)} = -1.05$, $p = 0.28$, CI = (-14.59; -4.35)] (**see Fig.8**).

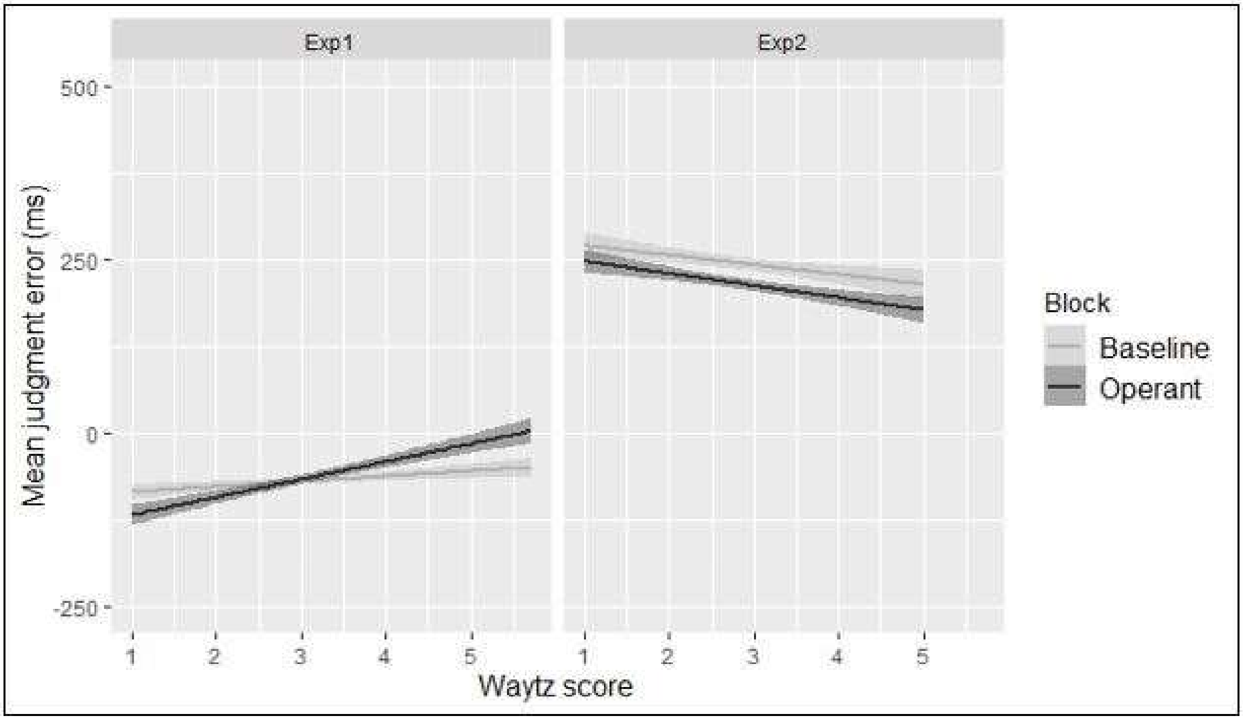

Fig.8. Mean JEs plotted as a function of Block type (Baseline, Operant), Waytz score, and Experiment (1, 2) when the critical event to be judged was the action.

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 Fig.9**).

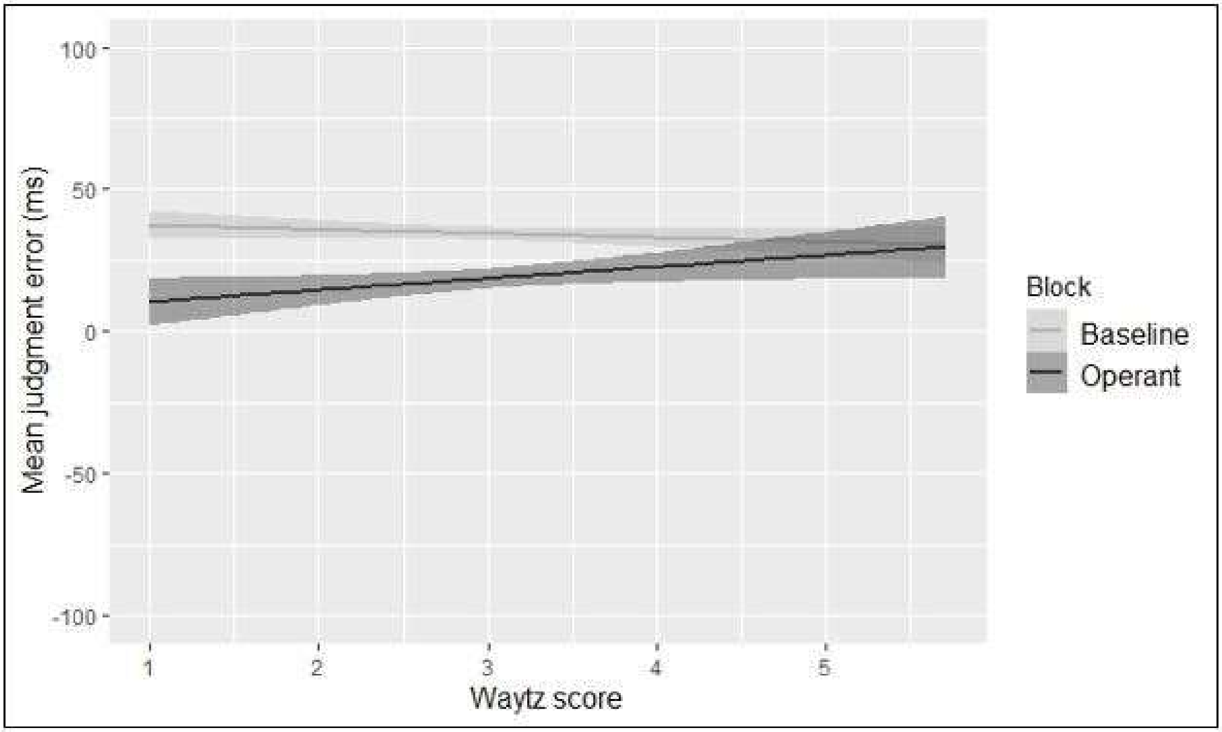

Fig.9. Mean JEs plotted as a function of Block type (Baseline, Operant) and Waytz score when the critical event was the tone.

4.3. Discussion of the role of the adoption of Intentional Stance in 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, Sidarus, Bonicalzi, & Haggard, 2017; Beyer, Sidarus, Fleming, & Haggard, 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.

5. 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 Wohschläger 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).

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.

Acknowledgments

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

Declaration of Conflicting Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

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.

Author 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.

Repository

Data, videos, and Supplementary Materials with Cozmo code are currently available via the Open Science Framework at the following link: https://osf.io/3scvw/ (project name: "Intentions with actions: the role of intentionality attribution on vicarious Sense of Agency in Human-Robot Interaction").

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.