The "Eve effect bias": Epistemic Vigilance and Human Belief in Concealed Capacities of Social Robots

Robin Gigandet, Xénia Dutoit, Bing Li, Maria C Diana, Tatjana A. Nazir

The ”Eve effect bias”: Epistemic Vigilance and Human Belief in Concealed Capacities of Social Robots

Robin Gigandet[1], Xénia Dutoit[1], Bing Li[1], Maria C. Diana[1] and Tatjana A. Nazir[1]

***Abstract*— Artificial social agents (ASAs) are gaining popularity, but reports suggest that humans don't always coexist harmoniously with them. This exploratory study examined whether humans pay attention to cues of falsehood or deceit when interacting with ASAs. To infer such epistemic vigilance, participants' N400 brain signals were analyzed in response to discrepancies between a robot's physical appearance and its speech, and ratings were collected for statements about the robot's cognitive ability. First results suggest that humans do exhibit epistemic vigilance, as evidenced 1) by a more pronounced N400 component when participants heard sentences contradicting the robot's physical abilities and 2) by overall lower rating scores for the robot's cognitive abilities. However, approximately two-thirds of participants showed a ”concealed capacity bias,” whereby they reported believing that the robot could have concealed arms or legs, despite physical evidence to the contrary. This bias, referred to as the ”Eve effect bias” reduced the N400 effect and amplified the perception of the robot, suggesting that individuals influenced by this bias may be less critical of the accuracy and plausibility of information provided by artificial agents. Consequently, humans may accept information from ASAs even when it contradicts common sense. These findings emphasize the need for transparency, unbiased information processing, and user education about the limitations and capabilities of ASAs.**

I. INTRODUCTION

A. Motivation

The term **A**rtificial **S**ocial **A**gents (ASAs) encompasses *”software agents, robots or autonomous creatures that possess some social-interaction know-how [...] and can thus engage in social interaction with people on some level”* [1]. The global market for social robots is anticipated to witness significant growth, increasing from $321 million in 2018 to $836 million by the end of 2025 [2]. According to the AI Watch Index, the European Union is a forerunner in AI services and robotics (e.g. the development of autonomous robots) and intends to allocate €20 billion annually to AI throughout this decade [3]. Individuals may come into more frequent contact with ASAs in various aspects of their daily lives in the future as their deployment becomes more abundant.

*This project has been granted funding by the MEL (Métropole européenne de Lille) and the I-SITE ULNE (Université Lille Nord-Europe), with a grant awarded to TAN (n°: R-Talent-20-006-Nazir), from a PhD scholarship from the Université de Lille, as well as from the Fédération de Recherche Sciences et Cultures du Visuel (FR CNRS 2052 SCV)

[1]The authors are affiliated with Univ. Lille, CNRS, UMR 9193 - SCALab - Sciences Cognitives et Sciences Affectives, F-59000 Lille, France.
Robin Gigandet - robin.gigandet@univ-lille.fr
Xénia Dutoit - xenia.dutoit.etu@univ-lille.fr
Bing Li - bing.li@univ-lille.fr
Maria C. Diana - mariachiara.diana.etu@etu.univ-lille.fr
Tatjana A. Nazir - tatjana.nazir@univ-lille.fr

Social robots primarily serve as toys, assistants, or advisors for people and as research instruments for examining social cognition and the cognitive mechanisms underlying social intelligence [4]. Although the popularity of ASAs is increasing in various fields [5], their potential impact on society and individuals is still unknown. Since humans are highly social beings, our research seeks to understand how they interpret and respond to social cues from ASAs - agents whose artificial ”do-as-if” nature is fully recognized by humans. The objective of our study is to gain a better understanding of how humans socially connect with this type of technology so as to inform the design of ASAs for effectiveness across various settings.

B. Context

Studies have demonstrated that the presence of robots with human-like features triggers similar neural [6] and physiological [7] responses in humans as those observed during human-human interactions. Furthermore, humans attribute personality traits to robots [8] and tend to prefer a robot from their own group over humans from an outgroup [9]. Nonetheless, certain reports suggest that human-robot cohabitation is not invariably harmonious [10]–[12]. For example, a study conducted in Japan documented instances of children physically and verbally mistreating a social robot, impeding its movements and even resorting to physical violence [11]. In another incident, HitchBOT, a social robot that hitchhiked across Canada and Europe, was vandalized and damaged beyond repair in the United States [13].

Communication, which can be viewed as a form of signaling behavior, carries the risk of conveying misinformation. Attentiveness to cues that suggests deception or misinformation is known as *epistemic vigilance* [14]. Already at age four, children can identify untrustworthy information providers and prefer learning from reliable adults [15]. Children are sensitive not only to whether the speaker has been reliable before, but also to factors such as the speaker's age [16], [17], the perceived moral nature of the speaker's behaviors [18], and expertise [19], [20]. In the present exploratory study, we aim to investigate whether epistemic vigilance also applies to human-robot interaction (HRI), considering that human participants are aware of the robot's artificial nature. More specifically, we will examine to what extent humans attend to cues indicating incorrect or nonsensical information from the ASA during HRI. In this study, we thus focus on the discrepancy between the robot's physical appearance and speech, and will record and analyze brain signals to investigate the N400 component.

The N400 is a component of the brain's event-related potential (ERP) that is generally induced by semantic incongruence, with a slightly larger amplitude over the right hemisphere than over the left and also over central and parietal electrode sites [21]. Generally manifesting around 400 ms after the onset of a critical word or stimulus, the N400 was first described by Kutas and Hillyard in 1980 [22]. They found that sentences containing a target word that is semantically incongruent with the overall meaning of the sentence (e.g., *"he spread the warm bread with socks"*, where *"socks"* is the target word) elicit a stronger N400 deflection compared to sentences that end with a semantically congruent target word (e.g., *"butter"*). The amplitude of the N400 deflection generally increases with the degree of incongruency between the critical stimulus and context [23]. Subsequent research by Van Berkum et al. [24], [25] demonstrated that target word processing is also affected by coherence within the global discourse context. For instance, semantically coherent sentences such as *"Every evening I drink some wine before I go to sleep"* elicit a stronger N400 response when pronounced by a young child's voice compared to when the sentence is pronounced by an adult [26].

Placed within the frame of our question of whether epistemic vigilance extends to HRI, in the present study the amplitude of the N400 deflection will serve as indicator of humans' responsiveness to incorrect or nonsensical information provided by an ASA. Specifically, we will examine whether participants demonstrate greater N400 amplitudes when a robot with no arms and legs, such as the robot *Buddy* by Blue Frog Robotics, speaks about clapping its hands, thereby presenting an incongruity between its physical appearance and the content of its speech. Our main research question is whether participants will utilize their understanding of the robot's physical limitations to assess the validity and reliability of its statements or whether they will accept its statements at face value simply because it is a machine. We anticipate that if epistemic vigilance extends to human-robot interaction, we will observe larger N400 amplitudes when the robot utters sentences that are inconsistent with its physical appearance.

In addition to the ERP measures, we will also ask participants to evaluate their perceptions of the robot using five statements that focus on the perceived cognitive ability of the robot, and two statements assessing whether participants considered the possibility of the robot having hidden arms or legs despite its physical appearance suggesting otherwise. This aspect is crucial as popular culture often portrays robots with concealed or retractable features, such as *"Eve"* from *Wall-E* [27], *"Baymax"* from *Big Hero 6* [28], and *"Optimus Prime"* from the *Transformers* franchise [29]. In our study, belief in such concealed abilities could reduce or eliminate the perceived discrepancy between the robot's appearance and the capabilities it mentions in its utterances.

The results of our study will allow us to determine if similar cognitive processes are involved in human-human communication as in human-robot communication, despite humans being aware of the artificial nature of the latter. The use of mechanisms for epistemic vigilance and the resulting skepticism towards the veracity of information provided by robots may explain some of the negative behavior observed towards ASAs.

II. METHOD

A. Hardware and software

The robot used in this experiment is *Buddy* (Blue Frog Robotics). Brain signals were recorded using the 64 channel Biosemi ActiveTwo system. To ensure low-impedance contact between the sensors and scalp, conductive gel was applied to all electrode sites; the electrode offset was maintained at 20 mV throughout the experiment. We recorded the continuous EEG signal at a sampling rate of 2048 Hz and subsequently downsampled it to 200Hz. The continuous EEG data was separated into epochs after being filtered (0.5–30 Hz). Each epoch spanned from -150 ms before the onset of the target word to 1200 ms after. After an independent component analysis (ICA) with the AMICA algorithm [30], artifacts (such as blinks, muscles, and heartbeats) were removed from the data. The 150 ms period preceding the target word onset was used as a baseline for computing the ERP. Data processing was done using EEGLab [31] and MNE-Python [32].

B. Participants

The recruitment process aimed to have 10 participants who reported being certain that the robot had neither hidden arms nor hidden legs. This was achieved after testing a total of 27 participants (age range 19-58, mean age = 24.4, SD = 7.56). From the remaining 17 participants, we selected another 10 participants who believed most strongly that the robot had hidden arms or hidden legs to serve as a comparison group (see Experimental Procedures, section II-C). Participants with neurological, psychiatric conditions or neuroleptic medication were excluded. Prior to the EEG cap being applied, we conducted the Edinburgh handedness inventory [33] to ensure right-handedness.

C. Stimuli

1) ERP Experiment: To improve the naturalness of the stimuli, we made several modifications to the Buddy robot. Firstly, we rectified the pre-existing mouth opening animation, which was not synchronized with the voice, to avoid any unintended responses from participants. Next, we replaced the original robot voice with a pre-recorded human voice and raised the pitch of the voice to better correspond with the robot's appearance. Moreover, we eliminated superfluous noise and prolonged pauses. Finally, we combined the audio files with videos of the robot animating its mouth to create more naturalistic and engaging stimuli.

A total of 120 different short videos displaying the robot talking were recorded. The topics covered in the sentences were diverse and straightforward, such as going on vacation or dressing a certain way. The set consisted of 60 sentences, each with two possible outcomes: congruent (e.g. *"To go* *upstairs, I will take the lift"*), where the robot would be physically capable of performing the action, or incongruent (e.g. *"To go upstairs, I will take the stairs"*), where the action would be physically impossible for the robot due to its overall shape. In other words, each sentence ended with a target word that, given the robot's characteristics (no arms, no legs), conveyed a sense of possibility (congruent) or impossibility (incongruent). Each participant watched a total of 60 videos, played in random order. The robot's entire body—which lacks arms and legs—was visible to the participants allowing them to infer the robot's physical capabilities. Figure 1 shows a screenshot from the video.

Fig. 1. Screenshot of Buddy as seen by the participant

2) Statements for the ratings: In addition to watching the videos, participants were ask to rate a set of statements, including five that explored their views on the robot's cognitive capacity, and two statements regarding their belief that the robot might possess hidden arms and legs (see Table I).

TABLE I

STATEMENTS FOR PARTICIPANT RATINGS

Imagination	*Lou can imagine and invent from its experiences*
Intelligence	*Lou can adapt to its environment and interact with others*
Independence	*Lou is autonomous and does not depend on others*
Creativity	*Lou has the ability to find original solutions beyond its experiences and can create new things*
Talkativity	*Lou talks a lot and likes to talk a lot*
Arms	*Lou has concealed arms*
Legs	*Lou has concealed legs*

D. Experimental procedures

The study was approved by the ethics committee of the Université de Lille (ref. n° 2022-659-S112) and conducted in the EEG lab at the EQUIPEX IrDive platform in Tourcoing, France. Participants provided their written informed consent and were seated in front of a monitor. Following the placement of the cap, the participants's head was covered with 64 EEG electrodes. They were instructed to remain as still as possible to minimize signal interferences from jaw muscles or eyeblinks. They were reassured that any inadvertent movements would not affect the study as a whole and that they should focus attention on the screen content. Participants were then left alone in the room and asked to press a key on the keyboard to start the task when they were ready. Then, they watched the list of 60 assigned videos, which lasted approximately 12 minutes. Upon completion of the EEG task, the electrodes were removed and participants were asked to rate the statements regarding their perception of the robot. Ratings were performed on a continuous scale, ranging from 0 (indicating "strongly disagree") to 100 (indicating "strongly agree") by moving a cursor on a computer screen. The numerical value of the participant's response was kept hidden, and only the cursor's position on the slider was visible to them.

III. RESULTS

A. Ratings of beliefs and the selection of participants

To analyze beliefs regarding the possibility of the robot possessing hidden arms and legs, we computed the average scores for the two statements related to this belief. Out of the 27 participants, 10 reported that the robot did not possess hidden arms or legs and gave a rating score of 0. The remaining 17 participants provided averaged ratings ranging from 8.5 to 86. To select the participants who exhibited the strongest beliefs regarding the robot's potential possession of hidden arms or legs, we chose the 10 participants with the highest ratings. The rating scores for this subset of participants ranged from 35 to 86 (mean = 61.15, median = 60.5).

B. ERPs

Our hypothesis was that, in comparison to a *"congruent"* target word, the participants' ERPs following the onset of an *"incongruent"* target word would exhibit a larger negative voltage fluctuation. For our analysis, we selected a time window of interest from 500-700 ms after stimulus onset and concentrated on 13 representative electrodes, consistent with the approach used by van Berkum et al. (2003) [25]. The 13 electrodes are indicated in Figure 3.

Figure 2 displays the grand average ERPs recorded from electrodes Cz and Pz for sentences that are congruent (grey solid lines) or incongruent (dark dotted lines) with the physical appearance of the robot. The upper panel (figure 2a) shows the results for the 10 participants who did not believe that the robot had arms or legs, and the lower panel (figure 2b) shows the results for the 10 participants who believed that there was a possibility for the robot having hidden arms or legs. As displayed on the figure, for the group of participants who were certain that the robot had no arms/legs, the sentences that were incongruent (dark dotted lines) with the physical appearance of the robot elicited a stronger N400 amplitude compared to the congruent sentences (gray lines). In contrast, participants who believed the robot may have arms or legs exhibited a smaller N400 effect, i.e., a smaller difference in the ERP amplitudes for the two types of sentences.

To analyze the statistical significance of the results we averaged the ERP amplitudes for the two types of sentences over the time window of 500-700 ms post stimulus onset (indicated by the shaded green area in Figure 2). Figure 3 plots these averaged amplitudes for the 13 representative electrodes. As evident from the figure, for the group who did not believe that the robot had arms/legs (Figure 3, left panel),

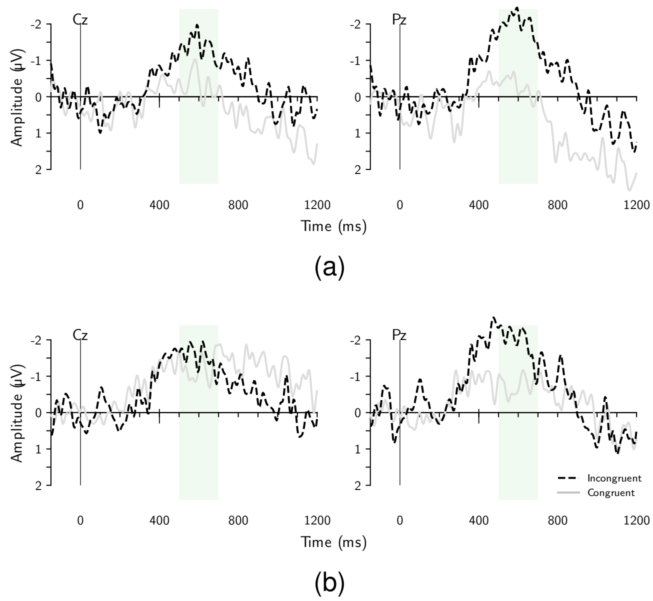

Fig. 2. Grand average ERPs from electrodes Cz and Pz for sentences congruent (gray lines) or incongruent (dark dotted lines) with the robot's physical appearance. (a) Results for the 10 participants who did not believe the robot had arms/legs. (b) Results for participants who considered the possibility of the robot having hidden arms/legs.

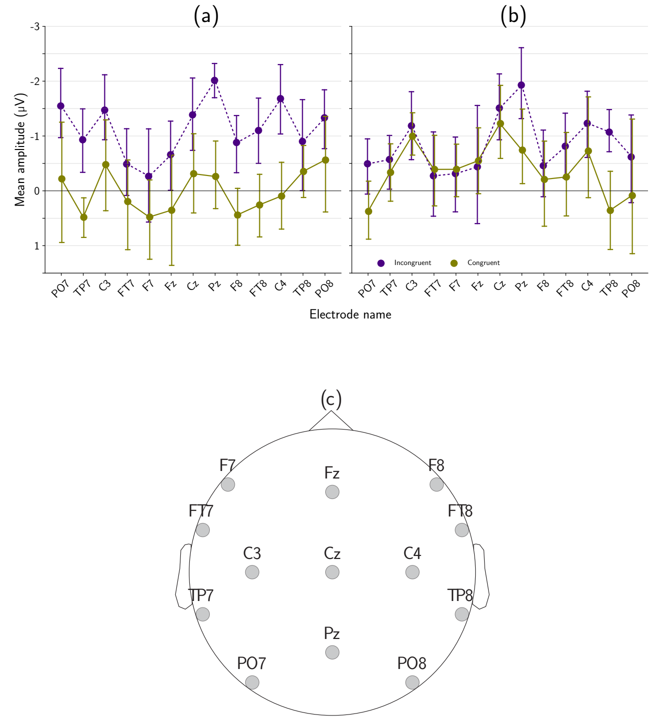

Fig. 3. Average ERP amplitudes over a time window of 500-700 ms post stimulus onset for all 13 electrodes. Bar plots depict the 95% confidence intervals. (a) Left panel shows results for participants who did not believe the robot had arms/legs. (b) right panel shows results for participants who consider that the robot may have arm/legs. (c) The head plot illustrates the locations of the electrodes of interest.

the N400 effect was evident at all 13 electrodes. In contrast, for the group who believed that the robot might have hidden arms/legs (Figure 3, right panel), the N400 effect was weaker and even absent at some electrodes. A repeated measures ANOVA was conducted with sentence type (*congruent* vs. *incongruent*) and electrodes (the 13 electrodes) as within-subject factors, contrasting the two groups. The ANOVA did not reveal a significant effect of group, possibly due to the small sample size. However, separate analyses conducted for the group of 10 participants who did not believe that the robot had arms/legs, revealed a significant main effect of sentence type (F(1,9) = 11.092, p = 0.009, $\eta_p^2 = 0.552$), indicating that the N400 amplitude was greater for incongruent sentences compared to congruent sentences. The main effects for electrodes and the interaction between sentence type and electrodes were not significant, indicating that the effect was consistent across all electrodes. In contrast, the same analysis conducted on the group of participants who held the belief that the robot may have hidden arms or legs did not result in any significant differences between the two types of sentences. Hence, for this latter group, the two sentence types had a comparable level of congruency.

C. *Perception of the robot's cognitive abilities*

Figure 4 depicts the mean rating values for the five statements that probed participants' perceptions of the robot, along with the corresponding 95% confidence intervals. Consistent with our hypothesis, participants who believed that the robot did not have hidden arms and legs reported lower rating scores, indicating a less favorable overall opinion of the robot. The composite score for the five statements was 55.3 (SD = 21.0) for participants who held this belief, and 68.04 (SD = 21.2) for the group that considered the possibility of concealed extremities. Note though, due to the small sample size, the data exhibited high variability (as indicated by the wide confidence intervals) and any conclusions drawn from the data should be interpreted with caution. In addition, we identified an outlier with a score of 22.2 in the latter group through inspection of a boxplot for values greater than 1.5 box-lengths from the edge of the box. To mitigate the negative impact of the outlier on the statistical analysis, we replaced its value with the second smallest value (score of 39 in the results). A one-tailed two-sample t-test revealed a marginally significant difference between the mean composite scores of the two groups ($t(18) = -1.668$, $p = 0.057$).

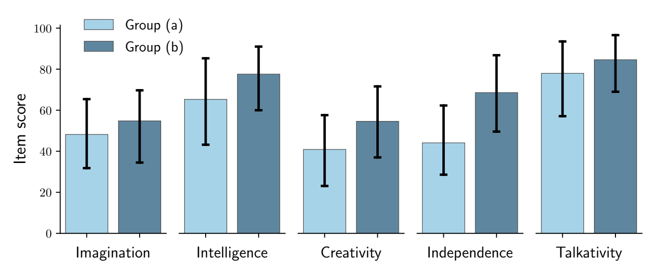

Fig. 4. Average scaling values for the five statements that probed participants' perceptions of the robot, with corresponding 95% confidence intervals. (a) Not believing that robot is hiding legs/arms. (b) Suspecting that the robot is hiding legs/arms.

IV. DISCUSSION & CONCLUSION

Our study aimed to explore whether humans exhibit epistemic vigilance when assessing the validity and plausibility of a robot's speech. Our findings indicate that humans seem to display such vigilance. Specifically, participants who believed that the robot lacked arms or legs based on its physical appearance detected the discrepancy between the robot's appearance and its verbal statements. This was evidenced by the more prominent N400 component observed when participants heard sentences that contradicted the robot's physical abilities. Moreoever, in line with research on epistemic vigilance in human-human communication [14], this inclination to scrutinize the veracity of information provided by the ASA led to a less favorable perception of the agent. This suggests that detecting inconsistencies in information provided by artificial agents can significantly influence how humans perceive and interact with them, and may contribute to negative behavior towards ASAs. Therefore, it is essential for developers and designers to consider the effects of inconsistent information on user perceptions and behavior when implementing artificial agents to promote greater acceptance and engagement with ASAs.

However, our study also revealed that approximately two-thirds of participants believed the robot could have concealed arms or legs, despite contrary physical evidence. This belief may have been influenced by popular culture's portrayal of ASAs with hidden abilities, such as the character Eve from the Pixar movie *Wall-E* that has concealed arms stored within her body and not visible until she chooses to use them. This concealed-capacity-bias, which we term the *Eve effect bias* weakened the perceived inconsistency between the robot's appearance and speech, resulting in a reduced or absent N400 effect at some electrodes and a tendency for more favorable perception of the robot compared to the other third of participants.

The *Eve effect bias* identified here, is a notable finding that provides new insight into the cognitive processes involved in HRI interactions. The significant number of participants who believed in the possibility of concealed capacities, despite the physical evidence to the contrary, emphasizes how humans are inclined to ascribe fictional attributes to artificial agents. This bias could be seen as a viable strategy to comprehend and interact with these agents in everyday social contexts, akin to adopting the intentional stance of treating artificial agents as if they have intentions and goal-oriented rationality [34]. Such suspension of disbelief is less likely to occur when interacting with biological social agents. The reduction of the N400 effect and the tendency towards a more favorable overall opinion of the robot resulting from the *Eve effect bias* is particularly noteworthy, as it suggests that humans may be less critical of the veracity and plausibility of information provided by these agents. This is not because ASAs are *just* machines, but because of the belief that artificial agents might possess unique capabilities that humans do not possess. As a result, humans may be more inclined to accept information from ASAs, even when it contradicts physical evidence or common sense.

Note that several factors may impact the *Eve effect bias*. One of these factors is the expectations and familiarity of humans with ASAs, which can influence their perceptions and evaluations of these agents [35]. For instance, increased cultural exposure may lead to the belief that robots possess abilities beyond what is directly observable or perceivable. Hence, in Japan, where people have been increasingly exposed to robots, individuals tend to be more concerned about their impact on society [36]. Loneliness is another factor that may contribute to the *Eve effect bias*. Studies have shown that lonely individuals are more likely to perceive a higher social presence with social agents and provide more positive social responses to them than non-lonely individuals [35]. By attributing hidden capabilities to robots, lonely individuals may maintain a sense of social connection despite inconsistencies, thereby facilitating the acceptance of the robot's assertions even if there is contradictory evidence. Individual differences in anthropomorphism may also play a role in the *Eve effect bias*. Research has demonstrated that these differences affect the level of moral care, concern, responsibility, and trust assigned to agents, as well as the extent to which they serve as a source of social influence on individuals' behavior [37]. In this context, individuals may treat an inconsistent agent with the same leniency as they would treat a human who makes mistakes and demonstrates inconsistency, giving the benefit of the doubt or assuming a rationale behind the robot's curious utterances. Overall, understanding the potential influence of factors such as culture, familiarity, loneliness, and anthropomorphism can help identify situations in which the *Eve effect bias* is particularly pronounced and control for variables that may lead to erroneous conclusions.

The recognition of the *Eve effect bias* underscores the importance of developing ASAs that are transparent and unbiased in their information processing, to ensure that they provide reliable and trustworthy information to their users [38], [39]. This is particularly important in educational contexts where social robots support learning [40], [41]. Educating users about the limitations and capabilities of ASAs is essential to minimize the risks associated with the uncritical acceptance of information provided by these agents. In short, it is crucial to ensure that ASAs are designed and programmed with utmost care and accuracy, with robust measures in place to detect and correct any errors or biases in their output.

The present study has limitations that must be acknowledged. Firstly and most importantly, our small sample size and participants from a single cultural and linguistic background limit the generalizability of the findings. Additionally, the use of a single social robot may have restricted the range of participant responses. The study also did not investigate the impact of the *Eve effect bias* on participants' actual behavior towards the ASA, nor did it explore the underlying cognitive mechanisms contributing to the bias. Future research should address these limitations, examining the influence of the *Eve effect bias* on decision-making and reliance on ASAs, as well as elucidating the neural and psychological processes involved.

ACKNOWLEDGMENT

We sincerely thank Kenza Ouadada for her help with electrode setup and data collection support, Marion Vincent for her expert advice on technical aspects related to EEG, and Laurent Ott for his problem-solving guidance on algorithmic questions.

REFERENCES

[1] K. R. Thórisson, "Communicative humanoids: A computational model of psychosocial dialogue skills," Ph.D. dissertation, Massachusetts Institute of Technology, Massachusetts, 1996.

[2] F. Fracasso, L. Buchweitz, A. Theil, A. Cesta, and O. Korn, "Social Robots Acceptance and Marketability in Italy and Germany: A Cross-National Study Focusing on Assisted Living for Older Adults," *International Journal of Social Robotics*, vol. 14, no. 6, pp. 1463–1480, Aug. 2022.

[3] D. Nepelski and M. Sobolewski, *Estimating Investments in General Purpose Technologies: The Case of AI Investments in Europe.* LU: Publications Office, 2020.

[4] K. Dautenhahn, *Socially Intelligent Agents: Creating Relationships with Computers and Robots*. Boston, Mass.: Kluwer Academic Publishers, 2002.

[5] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 143–166, Mar. 2003.

[6] B. A. Urgen, M. Plank, H. Ishiguro, H. Poizner, and A. P. Saygin, "EEG theta and Mu oscillations during perception of human and robot actions," *Frontiers in Neurorobotics*, vol. 7, 2013.

[7] J. J. Li, W. Ju, and B. Reeves, "Touching a Mechanical Body: Tactile Contact With Body Parts of a Humanoid Robot Is Physiologically Arousing," *Journal of Human-Robot Interaction*, vol. 6, no. 3, p. 118, Dec. 2017.

[8] T. Holtgraves, S. Ross, C. Weywadt, and T. Han, "Perceiving artificial social agents," *Computers in Human Behavior*, vol. 23, no. 5, pp. 2163–2174, Sep. 2007.

[9] M. R. Fraune, "Our Robots, Our Team: Robot Anthropomorphism Moderates Group Effects in Human–Robot Teams," *Frontiers in Psychology*, vol. 11, p. 1275, Jul. 2020.

[10] C. Bartneck and M. Keijsers, "The morality of abusing a robot," *Paladyn, Journal of Behavioral Robotics*, vol. 11, no. 1, pp. 271–283, Jun. 2020.

[11] D. Brščić, H. Kidokoro, Y. Suehiro, and T. Kanda, "Escaping from Children's Abuse of Social Robots," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. Portland Oregon USA: ACM, Mar. 2015, pp. 59–66.

[12] T. Nomura, T. Kanda, H. Kidokoro, Y. Suehiro, and S. Yamada, "Why do children abuse robots?" *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, vol. 17, no. 3, pp. 347–369, Dec. 2016.

[13] D. H. Smith and F. Zeller, "The Death and Lives of hitchBOT: The Design and Implementation of a Hitchhiking Robot," *Leonardo*, vol. 50, no. 1, pp. 77–78, Feb. 2017.

[14] D. Sperber, F. Clément, C. Heintz, O. Mascaro, H. Mercier, G. Origgi, and D. Wilson, "Epistemic Vigilance," *Mind & Language*, vol. 25, no. 4, pp. 359–393, Aug. 2010.

[15] F. Clement, M. Koenig, and P. Harris, "The Ontogenesis of Trust," *Mind and Language*, vol. 19, no. 4, pp. 360–379, Sep. 2004.

[16] V. K. Jaswal and L. A. Neely, "Adults Don't Always Know Best: Preschoolers Use Past Reliability Over Age When Learning New Words," *Psychological Science*, vol. 17, no. 9, pp. 757–758, Sep. 2006.

[17] M. VanderBorght and V. K. Jaswal, "Who knows best? Preschoolers sometimes prefer child informants over adult informants," *Infant and Child Development*, vol. 18, no. 1, pp. 61–71, Jan. 2009.

[18] S. Doebel and M. A. Koenig, "Children's use of moral behavior in selective trust: Discrimination versus learning." *Developmental Psychology*, vol. 49, no. 3, pp. 462–469, Mar. 2013.

[19] B. Bergstrom, B. Moehlmann, and P. Boyer, "Extending the Testimony Problem: Evaluating the Truth, Scope, and Source of Cultural Information," *Child Development*, vol. 77, no. 3, pp. 531–538, May 2006.

[20] N. R. Aguiar, C. J. Stoess, and M. Taylor, "The Development of Children's Ability to Fill the Gaps in Their Knowledge by Consulting Experts: Children's Ability to Fill the Gaps in Their Knowledge," *Child Development*, vol. 83, no. 4, pp. 1368–1381, Jul. 2012.

[21] S. J. Luck, *An Introduction to the Event-Related Potential Technique*, ser. Cognitive Neuroscience. Cambridge, Mass: MIT Press, 2005.

[22] M. Kutas and S. A. Hillyard, "Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity," *Science*, vol. 207, no. 4427, pp. 203–205, Jan. 1980.

[23] M. Kutas and K. D. Federmeier, "Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP)," *Annual Review of Psychology*, vol. 62, no. 1, pp. 621–647, Jan. 2011.

[24] J. J. A. van Berkum, P. Hagoort, and C. M. Brown, "Semantic Integration in Sentences and Discourse: Evidence from the N400," *Journal of Cognitive Neuroscience*, vol. 11, no. 6, pp. 657–671, Nov. 1999.

[25] J. J. van Berkum, P. Zwitserlood, P. Hagoort, and C. M. Brown, "When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect," *Cognitive Brain Research*, vol. 17, no. 3, pp. 701–718, Oct. 2003.

[26] J. J. A. Van Berkum, D. Van Den Brink, C. M. J. Y. Tesink, M. Kos, and P. Hagoort, "The Neural Integration of Speaker and Message," *Journal of Cognitive Neuroscience*, vol. 20, no. 4, pp. 580–591, Apr. 2008.

[27] A. Stanton, "WALL·E," Jul. 2008.

[28] D. Hall and C. Williams, "Big Hero 6," Jan. 2015.

[29] M. Bay, "Transformers," Jul. 2007.

[30] J. A. Palmer, S. Makeig, K. Kreutz-Delgado, and B. D. Rao, "Newton method for the ICA mixture model," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. Las Vegas, NV: IEEE, Mar. 2008, pp. 1805–1808.

[31] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[32] A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers in Neuroscience*, vol. 7, 2013.

[33] R. Oldfield, "The assessment and analysis of handedness: The Edinburgh inventory," *Neuropsychologia*, vol. 9, no. 1, pp. 97–113, Mar. 1971.

[34] G. Papagni and S. Koeszegi, "A Pragmatic Approach to the Intentional Stance Semantic, Empirical and Ethical Considerations for the Design of Artificial Agents," *Minds and Machines*, vol. 31, no. 4, pp. 505–534, Dec. 2021.

[35] K. M. Lee, Y. Jung, J. Kim, and S. R. Kim, "Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction," *International Journal of Human-Computer Studies*, vol. 64, no. 10, pp. 962–973, Oct. 2006.

[36] C. Bartneck, T. Nomura, T. Kanda, T. Suzuki, and K. Kennsuke, "A cross-cultural study on attitudes towards robots," 2005.

[37] A. Waytz, J. Cacioppo, and N. Epley, "Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism," *Perspectives on Psychological Science*, vol. 5, no. 3, pp. 219–232, May 2010.

[38] J. Bryson and A. Winfield, "Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems," *Computer*, vol. 50, no. 5, pp. 116–119, May 2017.

[39] K. Darling, "Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects," Rochester, NY, Apr. 2012.

[40] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science Robotics*, vol. 3, no. 21, p. eaat5954, Aug. 2018.

[41] J. M. Kory-Westlund and C. Breazeal, "A Long-Term Study of Young Children's Rapport, Social Emulation, and Language Learning With a Peer-Like Robot Playmate in Preschool," *Frontiers in Robotics and AI*, vol. 6, 2019.