Explanation through behavior: a human-inspired framework for explainable robotics

Marwen Belkaid[1]

Abstract— **Can robots' observable behavior give human observers insights into their hidden internal processes? This paper examines this idea with the aim of developing a framework for explainability in robotics. This framework constitutes a promising avenue for the development of intuitive human-robot interactions in which people would use the same brain mechanisms on which they rely to understand other humans during social interactions.**

I. Introduction

In recent years, research in artificial intelligence (AI) and robotics has witnessed the rise of increasingly sophisticated machine learning algorithms. Some of these techniques have been able to achieve impressive results and push the state of the art forward in a variety of applications. However, this leap in performance often came at the cost of a complexity rendering it hard to understand how and why the algorithm works. Consequently, concerns from both the scientific community and the general public have caused explainability to reemerge as a major topic in the field [1][2].

In this paper, we focus on the issue of explainability in robotics, and in human-robot interactions in particular. Indeed, while the need for explainability is relevant in many areas of applications of artificial intelligence, it is of outstanding importance in domains where these techniques are integrated in robots which operate in close vicinity to humans. In such contexts, it is crucial for users to understand why they behave as they do [3][4]. Because the issue of explainability is multifaceted, it should be addressed from various angles. Here, I focus on how the robot's observable behavior can provide users with insights into its hidden internal processes. Inspired by the central role of behavior in explaining brain mechanisms and by the importance of communicative behavior in human-human social understanding, *Explanation through behavior* is proposed as a framework for explainability in robotics and human-robot interactions in particular.

II. Explanation through behavior: how neuroscience and psychology deal with black-box cognitive systems

In the past years, the issue of explainability has been put in the spotlight to a great extent due to the success of deep learning techniques. Deep neural networks are designed to recognize patterns in data by iteratively processing and refining inputs through multiple layers of elementary operations. Trained on large amounts of data, they have been able to reach impressive results in various contexts, including robotics. Indeed, the strength of these methods lies in their capacity to build representations from high dimensional data, which makes them a good candidate for many robotic applications. However, the efficiency of these sophisticated machine learning techniques comes at the cost of explainability.

Compared to other methods such as decision trees or rule-based algorithms, it can be challenging to describe how deep networks generate their output. Consequently, they are often described as 'black-box' methods [1][2]. There is ongoing research aiming to develop more explainable models [2]. For instance, a large portion of research work in explainable artificial intelligence focuses on how to describe the processing steps which led to the system's output in human-interpretable terms. This approach can be very useful in many applications to enable tracing every decision step and identifying potentials errors or biases in order to address them. However, it is not necessarily the best candidate for robotic applications. To caricature, a robot verbally describing all its processing steps would overload the user with useless information.

Interestingly, the issue of explaining the inner workings of a black-box system is not specific to modern artificial intelligence. In fact, understanding why complex computing systems act as they do is essentially the challenge that psychologist and neuroscientists have been facing for decades. To study the human brain, experimental psychologists try to infer latent mental processes and computations from the behaviors they observe. While modern recording and manipulation techniques provide important insights into brain mechanisms, many researchers emphasize the need to refocus neuroscientific studies on behavior [5][6]. Because the fundamental function of the brain is to generate behavior, the best way to understand how the brain works is through a detailed analysis and characterization of the elicited behavior.

I argue that there is a lesson for roboticists to learn from this. Much like the brain's computations generate behavior, a robot's control algorithm determines the machine's actions. Crucially, unlike many other AI-powered systems, robots are embodied systems. Because they are physically present in the real world, their behavior is observable from the outside and can provide insights into the system' inner workings. This is a unique opportunity to tackle the issue of explainability in robotics. Others previously emphasized how behavior can be a medium for explainability [7]. Here I propose to ask how to make it a design principle. In other words,

[1] Marwen Belkaid is with ETIS UMR 8051, CY Paris Cergy Université, ENSEA, CNRS, F95000 Cergy, France marwen.belkaid@ensea.fr

how to design cognitive architectures for robots such that their output behavior gives the human observer insights into hidden internal processes.

III. Communicative behavior and social signals in human-human interaction

Human interactions rely on both verbal and non-verbal communication. While language is essential, extensive research in social psychology also highlighted the importance of non-verbal social signals, i.e. social cues which are exchanged through non-verbal communicative behavior[8].

Humans use a large variety of non-verbal signals, from facial expressions and gaze to postures, gestures and interpersonal distance. Body cues like posture and gestures can indicate personality traits such as dominance or trustworthiness [9]. In addition, they serve as co-speech cues which can be used to infer useful information such as whether one is being addressed during a social interaction [10]. Another example of the information humans extract from gestures is provided by a series of studies showing that people can predict from the onset of a reaching motion whether the intention is to place the object somewhere or to pass it to someone [11].

Furthermore, in human interactions, a lot of information can be communicated through the face. For instance, face cues have been extensively examined in the study of emotion expression, where patterns of configurations in face muscles are associated with certain internal states The role of face cues in communicating intention has also been emphasized [12][13]. Combined, gaze direction and facial expressions are an essential feature of interpersonal communication in social contexts [14]

In particular, because it often indicates the focus of attention, gaze can be very informative [14]. For example, when directed toward an object, gaze may indicate preference for the object or the intention to grasp it for example. On the other hand, gazing at someone may signal one's willingness to initiate interaction. One of the most commonly used experimental paradigms to study gaze in social contexts is the gaze cueing paradigm[15]: a face is presented on the screen, then its gaze is shifted toward the screen's left or right side. Subsequently, a target is shown either at the gazed-at location or at the opposite location. Typically, participants are faster in indicating validly cued targets (i.e. when gaze was directed to the target side) compared to invalidly cued targets (i.e. when gaze was directed to the opposite side). This gaze cueing effect highlights the facilitating role of others' gaze as well as the natural tendency of human to follow such a cue.

Overall, non-verbal signals can be highly informative about one's goals, intentions, emotions and other mental states. Non-verbal communication can be explicit or implicit, in the sense that individuals can voluntarily emit a signal to communicate with another person, but can also provide a cue without being aware of it. But most importantly, it provides a key communication channel enabling intuitive and often effortless social understanding.

IV. Communicative behavior and social signals in human-robot interaction

Given their importance in human-human interactions, several studies investigated the use of non-verbal social signals by robots. In particular, robots with human-like appearances may exploit features of their physical to exhibit communicative behaviors. Thus, humanoid robots can use bodily cues including head and arm movements in the form of postures and gestures, as well as interpersonal distance [16][17]. In addition, robot heads with anthropomorphic features enable the study of robot face cues [18][19]. Studies showed that even with minimalistic features mostly focusing on movements of the eyes, eyebrows and mouth, humans can recognize stereotypical facial expressions of emotion categories such as anger, joy or surprise [20], and report higher feeling of trust and empathy toward to machine [19].

Overall, adding co-speech behavioral cues can improve human-robot communication. Various combinations of social cues have been examined. For example, robot pointing gestures were found to help provide guidance for customers inside a shopping mall [17]. Bodily cues – including head orientation, gesture and distance – appear to be more effective than vocal cues to convince participants to change their mind and follow the robot's advice [16]. Such bodily signals also have an impact on subjective impressions people have about the robot, related to (negative) social reactance [21], perception of social presence [22] and trust [23].

Because of its key role in human interactions, several research works focused on gaze. In some cases, head orienting provided a coarse indication of gaze direction [16][24][25]. Others used more anthropomorphic heads allowing eye movements [26][27][28]. Thus, previous studies showed that robot gaze improves interactions in handover [29], conversational [24] and cooperative tasks [27]. In particular, eye contact plays an important role in these interactions [29][27] and increases the attribution of intentionality [30] and subjective feelings of engagement [30][31]. In the previous section, I mentioned the gaze cueing paradigm widely used in psychology. Kompatsiari and colleagues conducted a series of experiments recreating this paradigm with the humanoid robot iCub; i.e. replacing static stimuli shown on screen with embodied robot eye movement [28][31]. These studies successfully replicate the gaze cueing effect and further demonstrate the facilitating role of gaze in naturalistic, situated interactions with embodied agents.

Humans rely on robots' gaze direction in contexts where a social meaning can be extracted from the signals; for instance signaling turn-taking or intimacy [26], hinting toward a preferred object [26], or indicating the location of a future stimulus [28]. But robot gaze is also processed in situations where it is irrelevant. In a recent study, we created a task in which participants played a strategic game with the robot iCub [32]. Just before they chose their next move, participants looked at the robot which either established eye contact with them or avoided it with an averted gaze. Crucially, the nature of the gaze was totally irrelevant to the task: it was neither related to the robot's choice nor to the participant's. Yet, our results showed effects on both behavioral and neural levels, with robot gaze impacting response times and brain oscillation during decision-making. We also found that participants who interacted with a robot often establishing eye contact with them had a reduced tendency to use self-oriented which did not take the robot's actions into account. These results provide additional support for the notion that robot gaze carries a social meaning, and has the potential to reinforce the socialness of an interaction. Interestingly, gaze is such a powerful social signal that it can have an effect on human-robot interaction even when it is not programmed to communicate something in particular.

V. Explanation through behavior: a framework for human-robot interaction

This paper examines the idea that robot behavior can be a medium for explainability, providing human observers with insights into their inner workings. Because social understanding in human interactions relies heavily on non-verbal communicative behaviors, I argue that exploiting these communication channels could be the basis of a framework for *Explanation through behavior* in human-robot interactions. I provided various examples of studies investigating non-verbal robot signals. But are robot signals actually perceived as *social* signals? Based on previous research, it seems reasonable to consider that this is possible, at least to some extent. Indeed, various works highlighted humans' tendency to anthropomorphize robots, i.e. to attribute human-like qualities and characteristics to them despite the fact that they are not human [33][34]. Socialness is one of those qualities. In many cases, people have been found to see robots as social agents and expect them to behave in a socially intelligent way [35] [36]. This is shown in some studies through participants explicit responses to questionnaires. Concurring evidence with more implicit measures is provided by studies like the one mentioned above where we showed that the nature of robot's gaze modulated the socialness of participants' strategy in a joint decision-making task [32]. Altogether, it appears that robot signals can indeed be perceived by humans as social signals.

In a recent paper, Wallkotter and colleagues reviewed the existing literature about the use of social signals in contexts related to explainability in human-robot interactions [37]. For instance, one line of research focuses on the legibility of robot motion and the communication of intent though motion [38]. Other studies made use of specific gestures or gaze cues to provide feedback about what the robot has learned to a human tutor [39][40]. Promising results from these research works offer initial support for the approach advocated here. Nevertheless, how to ensure that communicative behaviors be meaningful and useful to users in complex scenarios remains an open question. As it has been pointed out, it is challenging for humans to build accurate mental models of sophisticated robots [3]. Therefore, I submit that developing the framework of *Explanation through behavior* requires progress in two challenging aspects: 1) how to design robot architectures that generate informative social signals, and 2) how to evaluate the ability of those architectures to improve users' understanding of robots' inner workings.

A. Generating grounded social signals for robots

Studies examining social signals in human-robot interaction often employ the Wizard of Oz technique, where an experimenter controls the robot's behavior to make sure it is triggered at the right moment [17], [21]. Other studies involve controlled interactions allow the robot's behavior to be preprogrammed by the task designer [32]. Another case is where behaviors are triggered by ad-hoc rules tailored for specific tasks [39]. Overall, it is a challenge to design robot architectures such that social signals are not teleoperated, preprogrammed, or scripted. Moving beyond the state of the art requires designing architectures capable of generating social signals more autonomously.

In recent years, reinforcement learning (RL) has been one of the most popular frameworks in the study of autonomous behaviors. It has proven to be a successful model of human (and animal) learning [41] and a powerful computational framework for robot learning [42]. The core idea in RL is that autonomous behavior is driven by the process of learning and selecting actions that maximize the agent's rewards. Rewards are typically obtained from the environment after performing an action. Intrinsically motivated RL goes a step further by using reward functions based not only on external signals but also on internal signals [43]. For instance, a decrease in a prediction error function can serve as an intrinsic reward indicating a progress in some task learning. Robots using these sophisticated techniques have the potential to develop skills with little to no human intervention. Yet, it has been pointed out that these systems also need to be equipped with the ability to interact intelligently with humans [44].

How can robot architectures connect autonomous and social behaviors? Examining natural cognition offers useful insight in this regard. Indeed, research in affective sciences emphasizes the dual role of emotion in regulating both autonomous behaviors and social interactions [45][46]. On the one hand, emotion determines how organisms perceive and respond to threats and reward opportunities in their environment. On the other hand, interpersonal communication is facilitated by emotional expressions through non-verbal face and body cues like facial expressions and gestures. Therefore, I posit that for robots' social behaviors to be truly communicative and meaningful, they must be grounded in the system's internal processes. To this end, affective states must serve as a bridge between internal processes and communicative, social behavior.

Previous studies provide good starting points on how to achieve this. Broekens and Chetouani proposed that emotion expression can rely on a definition of certain emotions as a function of temporal difference errors which are at the core of most RL models [4]. In this context, joy and distress are respectively elicited by situations where the outcome is better or worse than predicted by the system. Similarly, other models also derive robot affect from variables related to task learning. In previous works, we developed a model where repeated failure and long-lasting success respectively lead to states of frustration and boredom [47][48]. Related to the framework of intrinsic motivation, these affects are derived from prediction errors which enable the characterization of novelty and progress in task performance based on learned sensorimotor associations. These affective signals were then used to modulate visual attention, subsequently orient the robot's gaze [48]. Taken together, these examples show a promising approach for how to go from internal signals to social signals. Further research is needed to consolidate these models and build architecture capable of producing grounded communicative behaviors which can inform human observes about the system's internal states.

B. Assessing human response to robot social signals

Previous studies reported that participants sometimes misinterpreted robot non-verbal signals [39][40]. This observation stresses the importance of rigorously evaluating how people interpret and respond to those behaviors, and whether the latter ultimately help them understand the robot's functioning. I contend that to fulfill their communicative role, robot signals must be perceived as social signals by human users. In this regard, the literature suggests that robots' perceived socialness depends on a variety of factors such as the robot's appearance, the robot's behavior, but also the person's prior beliefs and expectations [35][49]. Regarding robot-related factors, Wiese and colleagues speculate that behavior probably outweighs appearance[49]. In line with this, we recently found that displaying communicative gestures was associated with higher trust ratings, a measure that correlated positively with both the performance-related scale and the social scale of the MDMT (Multi-Dimensional Measure of Trust) questionnaire [23].

The attribution of intentionality can also be considered a good index of perceived socialness. Indeed, one core process in social cognition is the adoption of intentional stance, a strategy that humans use to interpret the behavior of others with reference to mental states. The InStance Test (IST) was developed to assess the extent to which people adopted the intentional stance toward robots [50]. Watching a movie with a (teleoperated) robot with a rich social behavior was found to increase IST scores [51]. Similarly, interacting with a socially communicative robot in the context of a decision-making task resulted in higher IST scores compared to the mere observation of the same communicative behaviors with no interaction context [52].

In addition to subjective measures by means of explicit reporting, Wiese and colleagues proposed to make use of neuroscientific methods to assess how humans respond to robots [49]. This can be achieved by adapting psychological paradigms to human-robot interaction in order to obtain well-controlled behavioral measures. One example of this approach is the gaze cueing paradigm mentioned above [28]. A complementary approach is to measure brain activity in areas typically associated with social cognition [49]. Indeed, studies suggest that robots can activate similar brain networks as those involved in human-human interactions [35]. Interestingly, the resting state activity in social brain networks was found to predict the adoption of the intentional stance toward robots [53]. Employing these methods to evaluate real-time, naturalistic interactions remains a challenge [35]. Nevertheless, these examples show the benefit of combining subjective (explicit) and objective (implicit) measures. This applies to the general question of perceived socialness of robots, but also to assessing whether the use of communicative behavior is a viable solution to the issue of explainability in human-robot interaction more specifically .

VI. Conclusions

Social understanding strongly relies on communicative behavior. Specifically, humans are very good at guessing others' internal states from various non-verbal social signals. In addition to often being accurate, this process is also rather effortless. In this paper, I explored the idea of exploiting non-verbal communicative behavior as a medium for explainability in human-robot interactions, setting up the basis of *Explanation through behavior* as a framework for robotics. I identified two challenges facing the development of this framework: 1) designing robot architecture that generate social signals such that they can effectively inform humans about their internal processes; and 2) assessing whether humans indeed interpret those signals in a way that can help them better understand robots. Promising directions for future works have been put forward to address these challenges. Overall, the proposed framework constitutes a compelling avenue for the development of intuitive human-robot interactions in which people would use the same brain mechanisms they rely on to understand other humans.

Acknowledgment

I would like to thank Olivier Sigaud and Mohamed Chetouani for their feedback about many ideas presented in this paper, as well as Lorenzo Parenti, Kyveli Kompatsiari and Agnieszka Wykowska for fruitful discussions about social signals in human-human and human-robot interaction.

References

[1] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—Explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, p. eaay7120, Dec. 2019. [Online]. Available: https://www.science.org/doi/10.1126/scirobotics.aay7120

[2] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, June 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519308103

[3] S. Thellman and T. Ziemke, "The Perceptual Belief Problem: Why Explainability Is a Tough Challenge in Social Robotics," *ACM Transactions on Human-Robot Interaction*, vol. 10, no. 3, pp. 1–15, Sept. 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3461781

[4] J. Broekens and M. Chetouani, "Towards Transparent Robot Learning Through TDRL-Based Emotional Expressions," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 352–362, Apr. 2021, conference Name: IEEE Transactions on Affective Computing.

[5] J. W. Krakauer, A. A. Ghazanfar, A. Gomez-Marin, M. A. MacIver, and D. Poeppel, “Neuroscience Needs Behavior: Correcting a Reductionist Bias,” *Neuron*, vol. 93, no. 3, pp. 480–490, Feb. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0896627316310406

[6] Y. Niv, “The primacy of behavioral research for understanding the brain,” *Behavioral Neuroscience*, vol. 135, pp. 601–609, 2021, place: US Publisher: American Psychological Association.

[7] K. Chen, T. Hwu, H. J. Kashyap, J. L. Krichmar, K. Stewart, J. Xing, and X. Zou, “Neurorobots as a Means Toward Neuroethology and Explainable AI,” *Frontiers in Neurorobotics*, vol. 14, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2020.570308

[8] J. K. Burgoon, V. L. Manusov, and L. K. Guerrero, *Nonverbal communication*, second edition ed. New York, NY: Routledge, Taylor & Francis Group, 2022.

[9] M. Koppensteiner, P. Stephan, and J. P. M. Jäschke, “Moving speeches: Dominance, trustworthiness and competence in body motion,” *Personality and Individual Differences*, vol. 94, pp. 101–106, May 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191886916300137

[10] A. Nagels, T. Kircher, M. Steines, and B. Straube, “Feeling addressed! The role of body orientation and co-speech gesture in social communication,” *Human Brain Mapping*, vol. 36, no. 5, pp. 1925–1936, 2015, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.22746. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.22746

[11] C. Becchio, L. Sartori, M. Bulgheroni, and U. Castiello, “The case of Dr. Jekyll and Mr. Hyde: A kinematic study on social intention,” *Consciousness and Cognition*, vol. 17, no. 3, pp. 557–564, Sept. 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053810007000207

[12] G. Horstmann, “What do facial expressions convey: Feeling states, behavioral intentions, or actions requests?” *Emotion*, vol. 3, pp. 150–166, 2003, place: US Publisher: American Psychological Association.

[13] K. R. Scherer, M. Mortillaro, and M. Mehu, “Understanding the Mechanisms Underlying the Production of Facial Expression of Emotion: A Componential Perspective,” *Emotion Review*, vol. 5, no. 1, pp. 47–53, Jan. 2013, publisher: SAGE Publications. [Online]. Available: https://doi.org/10.1177/1754073912451504

[14] N. George and L. Conty, “Facing the gaze of others,” *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 38, no. 3, pp. 197–207, June 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0987705308000385

[15] A. Frischen, A. P. Bayliss, and S. P. Tipper, “Gaze cueing of attention: Visual attention, social cognition, and individual differences,” *Psychological Bulletin*, vol. 133, pp. 694–724, 2007, place: US Publisher: American Psychological Association.

[16] V. Chidambaram, Y.-H. Chiang, and B. Mutlu, “Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues,” in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. Boston Massachusetts USA: ACM, Mar. 2012, pp. 293–300. [Online]. Available: https://dl.acm.org/doi/10.1145/2157689.2157798

[17] P. Heikkilä, H. Lammi, M. Niemelä, K. Belhassein, G. Sarthou, A. Tammela, A. Clodic, and R. Alami, “Should a Robot Guide Like a Human? A Qualitative Four-Phase Study of a Shopping Mall Robot,” in *Social Robotics*, ser. Lecture Notes in Computer Science, M. A. Salichs, S. S. Ge, E. I. Barakova, J.-J. Cabibihan, A. R. Wagner, Á. Castro-González, and H. He, Eds. Cham: Springer International Publishing, 2019, pp. 548–557.

[18] S. Boucenna, P. Gaussier, P. Andry, and L. Hafemeister, “A Robot Learns the Facial Expressions Recognition and Face/Non-face Discrimination Through an Imitation Game,” *International Journal of Social Robotics*, vol. 6, no. 4, pp. 633–652, Nov. 2014. [Online]. Available: https://doi.org/10.1007/s12369-014-0245-z

[19] B. Gonsior, S. Sosnowski, C. Mayer, J. Blume, B. Radig, D. Wollherr, and K. Kuhnlenz, “Improving aspects of empathy and subjective performance for HRI through mirroring facial expressions,” in *2011 RO-MAN*. Atlanta, GA, USA: IEEE, July 2011, pp. 350–356. [Online]. Available: http://ieeexplore.ieee.org/document/6005294/

[20] C. Breazeal, “Emotion and sociable humanoid robots,” *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 119–155, July 2003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1071581903000181

[21] A. S. Ghazali, J. Ham, E. Barakova, and P. Markopoulos, “Assessing the effect of persuasive robots interactive social cues on users’ psychological reactance, liking, trusting beliefs and compliance,” *Advanced Robotics*, vol. 33, no. 7-8, pp. 325–337, Apr. 2019, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01691864.2019.1589570. [Online]. Available: https://doi.org/10.1080/01691864.2019.1589570

[22] S. Fiore, T. Wiltshire, E. Lobato, F. Jentsch, W. Huang, and B. Axelrod, “Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior,” *Frontiers in Psychology*, vol. 4, 2013. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00859

[23] L. Parenti, A. W. Lukomski, D. De Tommaso, M. Belkaid, and A. Wykowska, “Human-Likeness of Feedback Gestures Affects Decision Processes and Subjective Trust,” *International Journal of Social Robotics*, Nov. 2022. [Online]. Available: https://doi.org/10.1007/s12369-022-00927-5

[24] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu, “Conversational gaze aversion for humanlike robots,” in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. Bielefeld Germany: ACM, Mar. 2014, pp. 25–32. [Online]. Available: https://dl.acm.org/doi/10.1145/2559636.2559666

[25] H. Romat, M.-A. Williams, X. Wang, B. Johnston, and H. Bard, “Natural human-robot interaction using social cues,” in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2016, pp. 503–504, iSSN: 2167-2148.

[26] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, “Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior,” in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. La Jolla California USA: ACM, Mar. 2009, pp. 69–76. [Online]. Available: https://dl.acm.org/doi/10.1145/1514095.1514110

[27] J.-D. Boucher, U. Pattacini, A. Lelong, G. Bailly, F. Elisei, S. Fagel, P. Dominey, and J. Ventre-Dominey, “I Reach Faster When I See You Look: Gaze Effects in Human–Human and Human–Robot Face-to-Face Cooperation,” *Frontiers in Neurorobotics*, vol. 6, 2012. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2012.00003

[28] K. Kompatsiari, F. Ciardo, V. Tikhanoff, G. Metta, and A. Wykowska, “On the role of eye contact in gaze cueing,” *Scientific Reports*, vol. 8, no. 1, p. 17842, Dec. 2018, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-018-36136-2

[29] A. Moon, D. M. Troniak, B. Gleeson, M. K. Pan, M. Zheng, B. A. Blumer, K. MacLean, and E. A. Croft, “Meet me where i’m gazing: how shared attention gaze affects human-robot handover timing,” in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. Bielefeld Germany: ACM, Mar. 2014, pp. 334–341. [Online]. Available: https://dl.acm.org/doi/10.1145/2559636.2559656

[30] A. Ito, S. Hayakawa, and T. Terada, “Why robots need body for mind communication - an attempt of eye-contact between human and robot,” in *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, Sept. 2004, pp. 473–478.

[31] K. Kompatsiari, F. Ciardo, V. Tikhanoff, G. Metta, and A. Wykowska, “It’s in the Eyes: The Engaging Role of Eye Contact in HRI,” *International Journal of Social Robotics*, vol. 13, no. 3, pp. 525–535, June 2021. [Online]. Available: https://doi.org/10.1007/s12369-019-00565-4

[32] M. Belkaid, K. Kompatsiari, D. De Tommaso, I. Zablith, and A. Wykowska, “Mutual gaze with a robot affects human neural activity and delays decision-making processes,” *Science Robotics*, vol. 6, no. 58, p. eabc5044, Sept. 2021, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.abc5044

[33] F. Abell, F. Happé, and U. Frith, “Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development,” *Cognitive Development*,

vol. 15, no. 1, pp. 1–16, Jan. 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885201400000149

[34] N. Spatola, S. Marchesi, and A. Wykowska, "Different models of anthropomorphism across cultures and ontological limits in current frameworks the integrative framework of anthropomorphism," *Frontiers in Robotics and AI*, vol. 9, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2022.863319

[35] R. Hortensius and E. S. Cross, "From automata to animate beings: the scope and limits of attributing socialness to artificial agents," *Annals of the New York Academy of Sciences*, vol. 1426, no. 1, pp. 93–110, 2018, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nyas.13727. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/nyas.13727

[36] A. C. Horstmann and N. C. Krämer, "Great Expectations? Relation of Previous Experiences With Social Robots in Real Life or in the Media and Expectancies Based on Qualitative and Quantitative Assessment," *Frontiers in Psychology*, vol. 10, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00939

[37] S. Wallkötter, S. Tulli, G. Castellano, A. Paiva, and M. Chetouani, "Explainable Embodied Agents Through Social Cues: A Review," *ACM Transactions on Human-Robot Interaction*, vol. 10, no. 3, pp. 1–24, Sept. 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3457188

[38] M. Kwon, S. H. Huang, and A. D. Dragan, "Expressing Robot Incapability," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: Association for Computing Machinery, Feb. 2018, pp. 87–95. [Online]. Available: https://dl.acm.org/doi/10.1145/3171221.3171276

[39] C. Chao, M. Cakmak, and A. L. Thomaz, "Transparent active learning for robots," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2010, pp. 317–324, iSSN: 2167-2148.

[40] S. H. Huang, I. Huang, R. Pandya, and A. D. Dragan, "Nonverbal Robot Feedback for Human Teachers," Nov. 2019, arXiv:1911.02320 [cs]. [Online]. Available: http://arxiv.org/abs/1911.02320

[41] E. O. Neftci and B. B. Averbeck, "Reinforcement learning in artificial and biological systems," *Nature Machine Intelligence*, vol. 1, no. 3, pp. 133–143, Mar. 2019, number: 3 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s42256-019-0025-4

[42] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine, "How to train your robot with deep reinforcement learning: lessons we have learned," *The International Journal of Robotics Research*, vol. 40, no. 4-5, pp. 698–721, Apr. 2021. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0278364920987859

[43] G. Baldassarre, T. Stafford, M. Mirolli, P. Redgrave, R. M. Ryan, and A. Barto, "Intrinsic motivations and open-ended development in animals, humans, and robots: an overview," *Frontiers in Psychology*, vol. 5, 2014. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00985

[44] O. Sigaud, A. Akakzia, H. Caselles-Dupré, C. Colas, P.-Y. Oudeyer, and M. Chetouani, "Towards Teachable Autotelic Agents," *IEEE Transactions on Cognitive and Developmental Systems*, pp. 1–1, 2022, conference Name: IEEE Transactions on Cognitive and Developmental Systems.

[45] M. A. Arbib and J.-M. Fellous, "Emotions: from brain to robot," *Trends in Cognitive Sciences*, vol. 8, no. 12, pp. 554–561, Dec. 2004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364661304002669

[46] C. E. Izard, "The Many Meanings/Aspects of Emotion: Definitions, Functions, Activation, and Regulation," *Emotion Review*, vol. 2, no. 4, pp. 363–370, Oct. 2010. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1754073910374661

[47] A. Jauffret, M. Belkaid, N. Cuperlier, P. Gaussier, and P. Tarroux, "Frustration as a way toward autonomy and self-improvement in robotic navigation," in *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, Aug. 2013, pp. 1–7, iSSN: 2161-9476.

[48] M. Belkaid, N. Cuperlier, and P. Gaussier, "Emotional metacontrol of attention: Top-down modulation of sensorimotor processes in a robotic visual search task," *PLOS ONE*, vol. 12, no. 9, p. e0184960, Sept. 2017, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0184960

[49] E. Wiese, G. Metta, and A. Wykowska, "Robots As Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social," *Frontiers in Psychology*, vol. 8, 2017. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01663

[50] S. Marchesi, D. Ghiglino, F. Ciardo, J. Perez-Osorio, E. Baykara, and A. Wykowska, "Do We Adopt the Intentional Stance Toward Humanoid Robots?" *Frontiers in Psychology*, vol. 10, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00450

[51] S. Marchesi, D. De Tommaso, J. Perez-Osorio, and A. Wykowska, "Belief in sharing the same phenomenological experience increases the likelihood of adopting the intentional stance toward a humanoid robot." *Technology, Mind, and Behavior*, vol. 3, no. 3, pp. 11–11, June 2022. [Online]. Available: https://tmb.apaopen.org/pub/56dkj53d

[52] L. Parenti, S. Marchesi, M. Belkaid, and A. Wykowska, "Attributing Intentionality to Artificial Agents: Exposure Versus Interactive Scenarios," in *Social Robotics*, ser. Lecture Notes in Computer Science, F. Cavallo, J.-J. Cabibihan, L. Fiorini, A. Sorrentino, H. He, X. Liu, Y. Matsumoto, and S. S. Ge, Eds. Cham: Springer Nature Switzerland, 2022, pp. 347–356.

[53] F. Bossi, C. Willemse, J. Cavazza, S. Marchesi, V. Murino, and A. Wykowska, "The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots," *Science Robotics*, vol. 5, no. 46, p. eabb6652, Sept. 2020, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.abb6652