Bennett, C.C. (2021) "Evoking an intentional stance during human-agent social interaction: Appearances can be deceiving." *IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. pp.362-368.

Evoking an Intentional Stance during Human-Agent Social Interaction: Appearances Can Be Deceiving*

Casey C. Bennett, *Member IEEE*[1]

***Abstract*— A critical issue during human-agent and human-robot interaction is eliciting an intentional stance in the human interactor, whereas the human perceives the agent as a fully "intelligent" being with full agency towards their own intentions and desires. Eliciting such a stance, however, has proven elusive, despite work in cognitive science, robotics, and human-computer interaction over the past half-century. Here, we argue for a paradigm shift in our approach to this problem, based on a synthesis of recent evidence from social robotics and digital avatars. In short, in order to trigger an intentional stance in humans, perhaps our artificial agents need to adopt one about themselves.**

I. INTRODUCTION

Long since Dennett's seminal work on the stances people take on when interacting with other people as well as technology [1], researchers have been interested in developing interactive computing systems that can elicit the same "intentional stance" that occurs when humans interact with other intelligent beings. This intentional stance is where the human perceives the agent as a fully "intelligent" being with full agency towards their own intentions and desires, in contrast to the "physical stance" we assign inanimate objects or the "design stance" we assign machines (i.e. mechanistic reasoning about how they function). Criticisms of Dennett's work aside, one thing that is clear is that there is an argument to be made that our current generation of robots and intelligent agents fail in that aim, even if they appear to "think" in terms of clever pre-scripted programming [2] . Indeed, appearances can be deceiving, but at the same time they aren't *fooling* anyone.

One clear example of this is work in recent years on human-like androids [3], [4], [5]. Despite many advances in creating more life-like aesthetic features, such as silicon skin and pneumatic motors, when it comes to actual performance evaluation on tasks such as producing recognizable facial expressions or believable social interactions, androids do not outperform robots with much simpler designs [6]. More intermediate designs, like Pepper or iCub, also fail upon such comparisons [7]. There is a lack of true social cognition that manifests during the interaction in contrast to human-human interaction [8]. This issue extends across the spectrum of intelligent user interfaces, both physical and virtual, which limits our ability to create more truly interactive technology [9]. The fundamental question is why: if social cognition does not reside in the realm of appearance, then what property serves as the catalyst for *recognition* of what constitutes a "social interaction" between two intentional agents.

Beyond that basic scientific question, there is an argument to be made, perhaps controversially, that the **focus on artificial agents that merely appear to imitate humans is actually *damaging* our progress in robotics and AI** [10]. More aesthetically-pleasing robot designs, while perhaps appealing to the science-fiction fans among us, often suffer from the infamous "uncanny valley" effect, in that the disconnect between their more human-like appearance and their lack of human-like behavior triggers a sense of disgust or uncomfortable-ness in their human interactors [11]. This highlights the sharp divide between engineering *design* and engineering *functionality* when it comes to agents in social contexts. The two are entirely different endeavors, even if some synergy may potentially exist between them. Depending on what we are trying to accomplish, focusing on trying to create robots that look more human, rather than behave more human regardless of appearance, may be the wrong path. Indeed, a number of researchers have begun warning of an impending "social robotics winter" similar to AI winters of the past, due to the growing disconnect between the hype and appearance of social robots versus their actual social capabilities [12], [13].

If our focus is on the latter in terms of imbuing robots and agents with more realistic social behavior, then we would argue that this cannot be addressed by creating more human-looking artificial agents that rely on pre-scripted programming or models of the external world. Instead, we should endeavor to produce agents that can exhibit emergent idiosyncratic behavior **by learning how the world responds to it, rather than focusing on how it should respond to the world**, where the "world" is a projection of internal perception and beliefs rather than something external to perceived. In this paper, we discuss some recent research (as well some ongoing yet-to-be-published work) from the realms of social robotics and human-computer interaction (HCI) that points in this direction for future years.

*This research was supported by the research fund of Hanyang University (HY-2020), as well as a grant from the National Research Foundation of Korea (NRF grant# 2021R1G1A1003801)

[1]Casey C. Bennett is with the Department of Intelligence Computing, Hanyang University, Seoul Korea 04763 (email: cabennet@hanyang.ac.kr)

II. PRIOR RESEARCH ON ROBOTS AND SOCIAL COGNITION

There are several lines of research on these topics, from linguistics to HCI interfaces to robotics. One notable example is work on developing interactive robotic faces [14]. These range from complex human-like androids to simpler abstract faces, and from physical robots to digital avatars operating within virtual computer environments [15]. These robots and digital avatars have been studied on a number of specific tasks, such as emulating basic human facial expressions [6], autism research [16] , and understanding cross-cultural differences in social cognition based on gaze and joint attention [17], [18]. However, adapting work on such specific individual tasks to more expansive social interaction is a difficult and unsolved problem.

The challenge lies in the fluidity of natural social interactions, as well as the ability of naturally intelligent organisms to respond to novel stimuli and/or recover from failure during the interaction [19]. *Social fluidity* plays a critical role in creating a coherent construct during interaction, which forms the basis of a "virtual experience" [20], [21]. Indeed, without a consistent coherent construct, there is no virtual experience, which is why there is sometimes a discrepancy in HRI/HCI between carefully controlled lab studies and less controlled in-the-wild studies [22]. In our own past experimental work, we have investigated methods to produce such fluidity by training neural networks, only to find the produced models to be unstable, in the sense that when the human stimuli was altered, it resulted in the models learning and unlearning things in a frenetic manner [23], [24]. Indeed, the only solution appeared to be train multiple separate networks for each separate stimuli pattern ... certainly a doable solution but not a very scalable one. This suggests that the problem of social cognition is not simply a *perceptual* problem.

Interestingly, recent research on fMRI neural activity during social interaction between human-human versus human-agent (conversational digital avatar) found significant differences in several parts of the brain, including areas that are hypothesized to associate with the difference between a more social *intentional* stance versus a more *mechanistic* design stance (see Section 1) [8]. There appear to be potential neural pathways that respond specifically to social stimuli and reward mechanisms, to the point of involving pleasure-giving oxytocin release. Other research has found similar patterns when imbuing inanimate objects such as wheeled chairs with seemingly goal-directed behaviors [25]. The point is that altering people's beliefs over potential rewards via interacting with some social stimuli, rather than altering their actual perceptions, makes the difference. The appearance is not what matters [2].

A separate line of research coming to similar conclusions is that socially-assistive robots (SARs). Much work in the past few years has focused heavily on using robotic pets with elderly people with chronic health conditions [26], [27]. Several participatory design studies have suggested that the robots as artificial agents need to be designed not only from a physical standpoint, but also as artifacts geared towards specific situated contexts and uses [28], [29], in order to elicit a more social stance amongst the users. Other research in this vein focuses on using ecological momentary assessment (EMA) to reveal how people interact with such technology in the real world in real-time, to gain a better understanding of the behaviors elicited in humans [30], [31]. The long-term goal here is to implement machine learning models based on robotic sensor data to customize the robot-behavior in real time, without human designers needing to "re-program" anything. Similar to the research mentioned earlier in the section, it is an open question as to how the robot's behavior should change in order to elicit a belief in humans towards potential rewards via social interaction with the robotic pet. Furthermore, additional challenges remain with determining the optimal sensor suite needed aboard the robot (or an attached add-on sensor device) [26], in order to orient robotic behaviors towards people's ongoing cognitive state.

Technical challenges aside, the use of EMA as a novel form of psychological assessment during real-time social interaction opens up intriguing possibilities for attempting to uncover the how agent behavior influences social cognition during human-agent interaction, particularly in more naturalistic settings [32]. This is particularly true if the EMA is combined with other devices, such as a wearables and smartphones [33]. One could think of this as another angle to exploring the concept of "social presence" [34], as well as belief-desire-intent (BDI) models to create artificial "personalities" for autonomous agents [35]. EMA, however, focuses more specifically on understanding *situated* patterns of use, with the notion that attributions of agency and intentionally are dependent as much on the environmental "context" as on the agent itself.

In terms of HRI, this brings us back to question of creating autonomous agents that can elicit an intentional stance in the human interactor, particularly how we can engage "lower level mechanisms of social cognition" to cause an adoption of such a stance [36]. As mentioned in the first section, this presumably will require a divergence from simple pre-scripted programmatic behaviors, but on the other hand something more constrained than simply adding randomness. In the fields of autonomous agents and robotics, this is not necessarily a new idea, but not a solved one either [37], [38]. We need emergent yet adaptive interactive behavior where goals themselves can manifest from the social "milieu" within which the agent inhabits [39], [40]. One potential approach is to alter our learning paradigm from external world-based to that based on the agent's "internal world". The environment in this case becomes simply a medium upon which the agent *projects* its internal state, rather than a problem to be solved. **In short, in order to trigger an intentional stance in humans, perhaps our artificial agents need to "selfishly" adopt one about themselves**. We discuss some possibilities how this type of approach might be realized in the next section.

III. DESIGNING AGENTS THAT ADOPT AN INTENTIONAL STANCE

A. Overview

Our main premise here is that an intentional stance will never be adopted by human interactors unless the artificial agent is oriented towards its own ”internal world” versus trying to create models of the external world, regardless how it may appear from an aesthetic standpoint. From the agent's perspective, all that really matters is how the world responds to it, not how the world works or other abstract principles. To do the latter would be to have our artificial agents adopt a *design* stance of mechanistic reasoning, whereas the former permits the artificial agent to shift toward adoption of an *intentional* stance. And as we suggested in the prior section, in order for our agent to trigger an intentional stance in humans, it needs to adopt one about itself.

We can take as a starting point Searle's *Chinese Room Experiment*, where he famously argued that mechanistic reasoning performed by a computer is not akin to ”thinking” [41], along with the wide array of responses to it [42]. Our purpose here is not to rehash that debate, but suggest that we can actually follow this line of thinking to its extreme. Perhaps we all do just really live inside our heads, connected to the outside world through ”sensors”. Maybe the only world we can truly know, hearkening back to Descartes, is the one we create for ourselves, a ”brain in a vat” so to speak. The ”world” in that sense is a re-creation of reality within our minds based on sensory information. Evidence for this is abundant, going back to the work of Helmholtz and others on optical illusions and unconscious inferences [43]. For instance, the light from straight lines is actually initially received by our eyes as curved lines when they are above or below our gaze horizon due to the curvature of our eyes, but our brains know to reconstruct the line as a straight line, because we've learned from interacting with objects over time how they respond to our actions. We can see further evidence of this in pathological social disorders of the human brain, such as borderline personality disorder (BPD), where distortions in the ability to project one's internal state onto the external world arise from a disruption to the normal temporal process of memory formation, via the amygdala. This generates a disconnect between external sensory information and internal perceptual states, so that the disordered person inappropriately perceives and reacts (or overreacts) to environmental stimuli [44], [45]. The end result can lead to severe social deficits and dysfunction, which suggests such internal *projection* is part of normal human cognitive functioning.

In the rest of this section, we propose several experimental approaches to address these challenges.

B. Simulating Emergent Robotic Personalities

One potential approach to addressing this challenge is creating artificial agents that exhibit *emergent* personality traits using simple models, which could then be used as “building blocks” to experiment with evoking intentionality. At a high level, this could entail creating an agent that perceives its environment through sensory readings, makes a prediction of how its various potential actions might alter its environment, and how those possible future environments might impact its own internal affective state. The agent has no internal model of the external world, only a model of how its own behaviors are associated with its future affective state [46]. The environment in this sense is simply a *medium* upon which the agent projects its internal state. Learning occurs as the agent observes if the environment responds to it as expected or not.

Current research is making efforts to understand how this might work using simple interactive paradigms during human-robot interaction experiments with robotic faces [23], [24]. We start by using some visual stimuli (e.g. a baby toy moved around in front of the robot) that reacts to the robotic agent in various ways, and then attempt to have the agent learn those reaction patterns (an example of the robot's visual field can be seen in Figure 1). The key then is to alter how the stimuli react for different agents (e.g. in terms of congruence, tempo, social pauses, etc.), to produce a variable social environment, **to see if we can generate idiosyncratic behavior in different agents from the exact same underlying programming code**. There is no actual learning of the ”world” occurring however. The robot has only one singular concern here: how does the world respond to ”me”. From an empirical standpoint, the question is whether things we associate with personality traits might be emergent from such a paradigm. Early results so far suggest that it may be possible [23].

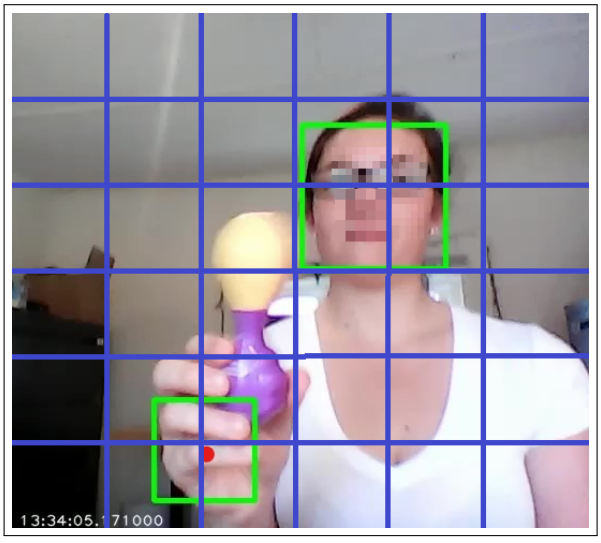

Fig. 1. Creating a simple experimental paradigm for emergent robotic personalities (taken from [23])

A principal idea in this approach is to keep the stimuli and social environment scaled-down, as simple as possible. Evolution didn't start with fully formed social mechanisms or “personalities”. Rather, it presumably started with simple variable behaviors across individuals, and somehow molded those over generations into the constructs which we call

"personalities" today. Indeed, recent biology research has shown that even simple sensory systems with a small number of photo-receptive cells can produce incredibly complex variable reactive behavior in simple organisms [47].

To put this in a more tongue-in-cheek way: if I am locked in Searle's Chinese Room, then maybe I should focus on learning as much as about the room as possible, since everything outside the room is inherently unknowable anyway except for the responses I receive from outside the room when I do things *inside* the room. We cannot know what we cannot perceive.

C. Reverse Engineering Social Cognition through HRI

One of the major challenges for adapting this to broader social interaction is how we measure intentionality in a way that relates to human social cognition. Just because we can create some "phenomena" in our agents, doesn't mean it is what we think it is, obviously. As mentioned in Section 1, appearances can be deceiving. This issue is important when thinking about whether the proposed approach here can apply more broadly to the field of autonomous agents, both physical and virtual.

Possibilities for such measurement exist using human-rating scales (e.g. Godspeed scale [48]) or physiological measures like fMRI, but the sensitivity of those measures, and whether they are truly measuring "social cognition", is debatable. One of the major issues is that while we can clearly see a difference in those measures between human-agent versus human-human interaction [8], it is not easy to know where the threshold lies between the two. One approach to dealing with this is the "reverse Turing test", where instead of trying to create more human-like agents, we instead attempt to create less "human-like" humans [49]. The idea being that we might be able to figure out **what causes the breakdown of social cognition and the attribution of intentionality to agents during human interaction by working *backwards*.** If so, we can then try to augment our autonomous agents to reduce the gap. To date, reverse Turing tests have largely been used for cybersecurity tasks (e.g. CAPTCHA), but we propose here is could be used to produce novel methods for measurement of what constitutes an "intentional agent" during human-agent interaction.

Current research is also taking this angle, using online game-playing paradigms that attempt to distort the interactions between humans in multiplayer video games to manifest the breakdown of social cognition, both in our own work and others [50], [51]. The distortions can involve altering social components themselves (e.g. modulating speech acoustics, facial expressions, turn-taking behaviors, etc.) though trained human actors or audio-visual technology. To do so, we have the humans play online, interacting through peer-to-peer video-conferencing tools such as Zoom, where we can manipulate the video feed as desired. We can alternatively alter contextual factors around the social interaction (e.g. altering the gameplay or game environment), while holding the social components constant, similar to what has been done around "social presence" research [34]. In either case, these distortions can then be replicated in AI-controlled virtual avatars playing the video game with humans in subsequent experiments, to empirically measure whether such phenomena contribute to the perception of intentionality during human-robot interaction.

One interesting aspect of using socially-assistive robots with elderly people (see Section 2), is that what constitutes social cognition and the attribution of intentionality to the agent by the user depends on the human. Indeed, that threshold has been observed to often be much lower in some populations: older adults, dementia patients, small children [52]. This lowered threshold may provide some scaffolding toward understanding where the attribution of intentionality comes from exactly, and what features of the social interaction might trigger it. The EMA research with SARs described in Section 2 is geared toward trying to address some of these questions. However, the flip side is that we know the design stance is not adopted in young children the same way as adults prior to a certain level of cognitive development, usually between 5-7 years old [2], [53]. As such, the intentional stance may simply be the default mode in those lower-threshold populations, lacking other capacities. The same may hold true in other populations such as older adults, which raises interesting questions about why intentionality attribution may have evolved in the first place.

Elucidating why such differences exist will play an important role if we want to design agents as part of user interfaces leveraging social interaction aspects. Otherwise, those user interfaces will forever be relegated to the category of "smart machines", rather than truly autonomous entities that can adapt to novel situations in their own right [9].

D. EMA to Capture the Temporal Dynamics of Real-World Robot Social Interaction

Another approach to this challenge is developing methodological tools that allow us to capture the temporal dynamics of real-world robot social interaction in a more "organic" way. As mentioned in Section 2, EMA is one such strategy for real-time interaction assessment. EMA can allow us to generate "ground truth" labels for interaction behaviors during HRI in-the-wild, which can then be used to distinguish sensor patterns into discernible activities using machine learning models [54], [55]. Understanding how sequences of behaviors unfold over time and how that impacts human perceptions of the "other agent" is fundamental to creating perceptions of intentionality in a neuroscientific sense [56] Such an approach could be used to potentially modulate robot social behaviors *autonomously* as well [26].

Current research is attempting to explore EMA as a novel form of interaction assessment with social robots in pilot studies in the US and Korea. The challenge is designing an EMA framework that collects *relevant* information at *appropriate* times, which demands careful consideration of both sampling strategies and stimulus design [57], [58]. For our purposes here, we are particularly interested in how data from EMA methodologies can be related to social cognition

components, similar to those described in Section 3.B. This entails using EMA as a form of psychological assessment during robot use, but also potentially the use of EMA to help users understand their own robot use during participatory design (PD) studies of social robots. Indeed, a majority of participants in our pilot studies have reported a desire for such feedback. Such an approach may help address some of the known challenges with PD research [59].

The principal question is how ongoing use of social robots in situated settings impacts perceptions of intentionality, and how **such information can be used to design *situated* use cases that evoke such perceptions, rather than focusing only on designing the agent behavior itself**. In other words, the design of agents that adopt their own intentional stance is dependent on integrating such contextual factors into the agent's projection of internal state onto its environment (as described in Section 3.B). Otherwise, there will be a *disconnect* between the human users' sense of intentionality, and our agents' sense of intentionality. We elaborate on this concept of designing situated interactions for intentionality in the next section.

IV. SYNTHESIZING A BROADER CONSTRUCT OF SOCIAL INTERACTION IN AUTONOMOUS AGENTS

The prior section laid out several potential experimental avenues for trying to create *intentionality* in socially-interactive agents. The principle idea is that there **needs to be close alignment between our agent's design and its functionality, as well as alignment between the human's sense of intentionality and the agent's** [12], [13]. However, the question remains as to how we synthesize these various approaches into a single unifying framework that allows us to consistently design such agents. We could, of course, assign the agent some pre-defined goals for specific situated use cases (see Section 3.D) or specific types of social interactions (Section 3.C), but that would not be in keeping with an *emergent* approach described in Section 3.B. There are however cues we can take from other lines of research.

For instance, the embodied and/or enactive paradigm within cognitive science contends that such emergent goals can manifest via the dynamics of the interaction itself, that there is some process of "sensori-motor babbling" that can create such goals [60]. The idea is similar to that of "minimal cognition" in dynamical systems theory [61]. Another possible avenue is to take advantage of recent fMRI research on the differences in brain activity during human-agent and human-human interaction (see Section 2) [8]. In particular, some of the "social reward" mechanisms hold promise in this regard. However, as Rolf & Crook have pointed out [40], the aim isn't simply some new form of reinforcement learning, but rather the ability for agents to generate their own novel goals. In this view, goals are not the reward signals themselves, but rather the ultimate end states of action ... a subtle but important distinction. As such, any sort of "novel intentional goal" in an agent has to be an internal representation based purely on the agent's internal perspective, not an external teleological interpretation based on the external world [40]. Intentionality by this definition must exist independent of the perspective of the designer or their design parameters. Or to put it more simply, **intentionality cannot be designed, rather only the conditions in which it might arise can be created**.

So then, what conditions might we need to create? One possibility comes from recent research on *curiosity-based robotics*. The central idea here is that by adding an artificial "curiosity drive" to the agent we can induce it to create its own goals by focusing on "intrinsic rewards" during exploration of the environment [39]. These intrinsic rewards are typically some measure of the learning process itself, such as reducing the prediction error rate of environmental response while trying different possible actions, with the focus on learning the response across all possible actions rather than only trying to figure out the best action. This aligns with our current experimental robotic face paradigm described in Section 3 (focusing on how the environment responds to the agent, rather than the other way around). We can trace such approaches back to Kismet robotic face sensorimotor babbling experiments a couple decades ago [62], or even the behavior-based robotics approaches that came before it (see next section).

More recent research has shown how a curiosity-based approach can be used for dialogue-based robots [63] and emulating infant-like predictive learning in robotic systems [16]. It is of course one thing to apply these ideas to sensorimotor tasks, or very specific social tasks. The open question is how these ideas can be expanded to address broader constructs of social interaction, or even "robotic personalities" themselves, during human-agent interaction. This is especially true if we consider longer-term interactions, where contexts and settings may change over time [64]. We stress that we think it is the combination of elements described in this paper that may prove a path forward, rather than any individual component. A synthesis of these ideas.

V. CONCLUSION

It has been roughly 30 years since Rodney Brooks wrote his famous papers on behavioral-based robotics, lamenting the limitations of current AI of the time [37], and roughly 50 years since Dennett first wrote about intentional agents [1]. Many avenues to realize those ideas have been attempted over the years, but in some ways we are still stuck on the same problems. The dream however remains, in its purest form, to simplify the complex by proposing ways to reduce the computational burden necessary for life-like agents. To build atop what's been accomplished so far. The difference here being that recent research has led to a better understanding of socio-perceptual systems during human-robot interaction and the limits of "form over function" in triggering social cognition. We've learned much about what doesn't work over the past 20 years, which may guide us better in the coming years.

ACKNOWLEDGMENT

This work was supported by the research fund of Hanyang University (HY-2020), as well as a grant from the National Research Foundation of Korea (NRF grant# 2021R1G1A1003801). We would also like to thank our various collaborators over the years who have contributed to different aspects of this work.

REFERENCES

[1] D. C. Dennett, "Intentional systems," *The Journal of Philosophy*, vol. 68, no. 4, pp. 87–106, 1971.

[2] D. Ghiglino and A. Wykowska, "When robots (pretend to) think," in *Artificial Intelligence: Reflections in Philosophy, Theology, and the Social Sciences*. Brill, 2020, pp. 49–74.

[3] C. Becker-Asano and H. Ishiguro, "Evaluating facial displays of emotion for the android robot geminoid f," *IEEE Workshop on Affective Computational Intelligence (WACI)*, pp. 1–8, 2011.

[4] C.-E. Yu and H. F. B. Ngan, "The power of head tilts: Gender and cultural differences of perceived human vs human-like robot smile in service," *Tourism Review*, vol. 74, no. 3, pp. 428–442, 2019.

[5] Z. Faraj, M. Selamet, C. Morales, P. Torres, M. Hossain, and H. Lipson, "Facially expressive humanoid robotic face," *HardwareX*, vol. e00117, 2020.

[6] C. C. Bennett and S. Šabanović, "Deriving minimal features for human-like facial expressions in robotic faces," *International Journal of Social Robotics*, vol. 6, no. 3, pp. 367–381, 2015.

[7] I. Brinck and C. Balkenius, "Mutual recognition in human-robot interaction: A deflationary account," *Philosophy & Technology*, vol. 33, no. 1, pp. 53–70, 2020.

[8] B. Rauchbauer, B. Nazarian, M. Bourhis, M. Ochs, L. Prévot, and T. Chaminade, "Brain activity during reciprocal social interaction investigated using conversational robots as control condition," *Philosophical Transactions of the Royal Society B*, vol. 1771, p. 20180033, 2019.

[9] S. T. Völkel, C. Schneegass, M. Eiband, and D. Buschek, "What is" intelligent" in intelligent user interfaces? a meta-analysis of 25 years of iui," *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI)*, pp. 477–487, 2020.

[10] R. Moore, "From talking and listening robots to intelligent communicative machines," in *Robots that Talk and Listen*. DeGuyter, 2015, pp. 317–335.

[11] K. F. MacDorman, R. D. Green, C.-C. Ho, and C. T. Koch, "Too real for comfort? uncanny responses to computer generated faces," *Computers in Human Behavior*, vol. 25, no. 3, pp. 695–710, 2019.

[12] A. Henschel, R. Hortensius, and E. S. Cross, "Social cognition in the age of human–robot interaction," *Trends in Neurosciences*, vol. 43, no. 6, pp. 373–384, 2020.

[13] S. Tulli, D. A. Ambrossio, A. Najjar, and F. J. R. Lera, "Great expectations & aborted business initiatives: The paradox of social robot between research and industry," *BNAIC/BENELEARN*, 2019.

[14] A. Thomaz, G. Hoffman, and M. Cakmak, "Computational human-robot interaction," *Foundations and Trends in Robotics*, vol. 4, no. 2-3, pp. 105–223, 2016.

[15] C. Chesher and F. Andreallo, "Robotic faciality: The philosophy, science and art of robot face," *International Journal of Social Robotics*, pp. 1–14, 2020.

[16] Y. Nagai, "Predictive learning: its key role in early cognitive development," *Philosophical Transactions of the Royal Society B*, vol. 374, no. 1771, p. 20180030, 2019.

[17] C. C. Bennett, S. Šabanović, M. R. Fraune, and K. Shaw, "Context congruency and robotic facial expressions: Do effects on human perceptions vary across culture?" *23rd IEEE International Symposium on Robot and Human Interactive Communication (ROMAN)*, pp. 465–470, 2014.

[18] M. Kim, T. Kwon, and K. Kim, "Can human–robot interaction promote the same depth of social information processing as human–human interaction?" *International Journal of Social Robotics*, vol. 10, no. 1, pp. 33–42, 2018.

[19] S. Engelhardt, E. Hansson, and I. Leite, "Better faulty than sorry: Investigating social recovery strategies to minimize the impact of failure in human-robot interaction," *Workshop on Conversational Interruptions in Human-Agent Interactions (IVA 2017)*, vol. 1943, pp. 19–27, 2017.

[20] M. Jung and P. Hinds, "Robots in the wild: A time for more robust theories of human-robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 7, no. 1, p. 2, 2018.

[21] A. G. Cass, K. Striegnitz, N. Webb, and V. Yu, "Exposing real-world challenges using hri in the wild," *Proceedings of the 4th Workshop on Public Space Human-Robot Interaction at the Intl. Conf. on Human-Computer Interaction with Mobile Devices and Services*, pp. 569–595, 2018.

[22] S. Sabanovic, M. P. Michalowski, and R. Simmons, "Robots in the wild: Observing human-robot social interaction outside the lab," *9th IEEE International Workshop on Advanced Motion Control*, pp. 596–601, 2006.

[23] C. C. Bennett, "Emergent robotic personality traits via agent-based simulation of abstract social environment," *Information*, vol. 12, no. 3, p. 103, 2021.

[24] C. Bennett, "Robotic faces: Exploring dynamical patterns of social interaction between humans and robots," Ph.D. dissertation, Indiana University, 2015.

[25] K. Terada, T. Shamoto, and A. Ito, "Human goal attribution toward behavior of artifacts," *17th IEEE International Symposium on Robot and Human Interactive Communication (ROMAN)*, pp. 160–165, 2008.

[26] C. C. Bennett, S. Sabanovic, J. A. Piatt, S. Nagata, L. Eldridge, and N. Randall, "A robot a day keeps the blues away," *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 536–540, 2017.

[27] A. Liang, I. Piroth, H. Robinson, B. MacDonald, M. Fisher, U. M. Nater, N. Skoluda, and E. Broadbent, "A pilot randomized trial of a companion robot for people with dementia living in the community," *Journal of the American Medical Directors Association*, vol. 18, no. 10, pp. 71–878, 2017.

[28] S. Sabanovic, W.-L. Chang, C. C. Bennett, J. A. Piatt, and D. Hakken, "A robot of my own: participatory design of socially assistive robots for independently living older adults diagnosed with depression," *International Conference on Human Aspects of IT for the Aged Population*, pp. 104–114, 2015.

[29] N. Randall, C. C. Bennett, S. Šabanović, S. Nagata, L. Eldridge, S. Collins, and J. A. Piatt, "More than just friends: in-home use and design recommendations for sensing socially assistive robots (sars) by older adults with depression," *Paladyn Journal of Behavioral Robotics*, vol. 10, no. 1, pp. 237–255, 2019.

[30] J. Zulueta, A. Piscitello, M. Rasic, R. Easter, P. Babu, S. A. Langenecker, M. McInnis, O. Ajilore, P. C. Nelson, K. Ryan, and A. Leow, "Predicting mood disturbance severity with mobile phone keystroke metadata: a biaffect digital phenotyping study," *Journal of Medical Internet Research*, vol. 20, no. 7, p. e241, 2018.

[31] C. Vesel, H. Rashidisabet, J. Zulueta, J. P. Stange, J. Duffecy, F. Hussain, A. Piscitello, J. Bark, S. A. Langenecker, S. Young, E. Mounts, L. Omberg, P. C. Nelson, R. C. Moore, D. Koziol, K. Bourne, C. C. Bennett, O. Ajilore, A. P. Demos, and A. Leow, "Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: A biaffect ios study," *Journal of the American Medical Informatics Association*, vol. 27, no. 7, pp. 1007–1018, 2020.

[32] S. Intille, C. Haynes, D. Maniar, A. Ponnada, and J. Manjourides, "uEMA: Microinteraction-based ecological momentary assessment (ema) using a smartwatch," *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 1124–1128, 2016.

[33] B. Cao, L. Zheng, C. Zhang, P. S. Yu, A. Piscitello, J. Zulueta, O. Ajilore, K. Ryan, and A. D. Leow, "Mutual recognition in human-robot interaction: A deflationary account," *23rd ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 747–755, 2017.

[34] C. S. Oh, J. N. Bailenson, and G. F. Welch, "A systematic review of social presence: Definition, antecedents, and implications," *Frontiers in Robotics and AI*, vol. 5, p. 114, 2018.

[35] M.-A. Puica and A.-M. Florea, "Emotional belief-desire-intention agent model: Previous work and proposed architecture," *International Journal of Advanced Research in Artificial Intelligence*, vol. 2, no. 2, pp. 1–8, 2013.

[36] E. Schellen and A. Wykowska, "Intentional mindset toward robots—open questions and methodological challenges," *Frontiers in Robotics and AI*, vol. 5, p. 139, 2019.

[37] R. A. Brooks, "Intelligence without reason," *12th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 569–595, 1991.

[38] P. Maes, "Artificial life meets entertainment: lifelike autonomous agents," *Communications of the ACM*, vol. 38, no. 11, pp. 108–114, 1995.

[39] G. Gordon, "Infant-inspired intrinsically motivated curious robots," *Current Opinion in Behavioral Sciences*, vol. 35, pp. 28–34, 2020.

[40] M. Rolf and N. T. Crook, "What if: Robots create novel goals? ethics based on social value systems," *EDIA Workshop at the European Conference on Artificial Intelligence (ECAI)*, pp. 20–25, 2016.

[41] J. Searle, "Minds, brains and program," *Behavioral and Brain Science*, vol. 3, pp. 417–424, 1971.

[42] J. Preston and M. Bishop, Eds., *Views into the Chinese room: New essays on Searle and Artificial Intelligence*. Oxford, England: Oxford University Press on Demand, 2002.

[43] R. S. Turner, *In the eye's mind: vision and the Helmholtz-Hering controversy*. Princeton, NJ, USA: Princeton University Press, 2014.

[44] T. Fuchs, "Fragmented selves: Temporality and identity in borderline personality disorder," *Psychopathology*, vol. 40, no. 6, pp. 379–387, 2007.

[45] C. Paret, C. Jennen-Steinmetz, and C. Schmahl, "Disadvantageous decision-making in borderline personality disorder: Partial support from a meta-analytic review," *Neuroscience & Biobehavioral Reviews*, vol. 72, pp. 301–309, 2017.

[46] J. S. Olier, E. Barakova, C. Regazzoni, and M. Rauterberg, "Reframing the characteristics of concepts and their relation to learning and cognition in artificial agents," *Cognitive Systems Research*, vol. 44, pp. 50–68, 2017.

[47] J. P. Dexter, S. Prabakaran, and J. Gunawardena, "A complex hierarchy of avoidance behaviors in a single-cell eukaryote," *Current Biology*, vol. 29, no. 24, pp. 4323–4329, 2019.

[48] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, pp. 71–81, 2009.

[49] S. M. Shieber, Ed., *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. Cambridge, MA, USA: Mit Press, 2004.

[50] S. T. Jesso, W. G. Kennedy, and E. Wiese, "Behavioral cues of humanness in complex environments: How people engage with human and artificially intelligent agents in a multiplayer videogame," *Frontiers in Robotics and AI*, vol. 7, 2020.

[51] J. Y. Suh, C. C. Bennett, B. Weiss, E. Yoon, J. Jeong, and Y. Chae, "Development of speech dialogue systems for social ai in cooperative game evironments," *IEEE Region 10 Symposium (TENSYMP 2021)*, 2021.

[52] S. Sabanovic, C. C. Bennett, W.-L. Chang, and L. Huber, "Paro robot affects diverse interaction modalities in group sensory therapy for older adults with dementia," *13th IEEE International Conference on Rehabilitation Robotics (ICORR)*, pp. 1–6, 2013.

[53] T. P. German and S. C. Johnson, "Function and the origins of the design stance," *Journal of Cognition and Development*, vol. 3, no. 3, pp. 279–300, 2002.

[54] Y. Chen and C. Shen, "Performance analysis of smartphone-sensor behavior for human activity recognition," *IEEE Access*, vol. 5, pp. 3095–3110, 2017.

[55] G. Ogbuabor and R. La, "Human activity recognition for healthcare using smartphones," *International Conference on Machine Learning and Computing*, pp. 41–46, 2018.

[56] 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, p. 1663, 2017.

[57] S. Shiffman, A. A. Stone, and M. R. Hufford, "cological momentary assessment," *Annual Review of Clinical Psychology*, vol. 4, pp. 1–32, 2008.

[58] Y. Liu, L. Nie, L. Liu, and D. S. Rosenblum, "A pilot randomized trial of a companion robot for people with dementia living in the community," *Neurocomputing*, vol. 181, pp. 108–115, 2016.

[59] S. Lindsay, D. Jackson, G. Schofield, and P. Olivier, "Engaging older people using participatory design," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 1199–1208, 2012.

[60] T. Froese and S. Gallagher, "Getting interaction theory (it) together: integrating developmental, phenomenological, enactive, and dynamical approaches to social interaction," *Interaction Studie*, vol. 13, no. 3, pp. 436–468, 2012.

[61] R. D. Beer, "A dynamical systems perspective on agent-environment interaction," *Artificial intelligence*, vol. 72, no. 1-2, pp. 173–215, 1995.

[62] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal on Human Computer Studies*, vol. 59, no. 1-2, p. 119–155, 2003.

[63] M. Doering, P. Liu, D. F. Glas, T. Kanda, D. Kulić, and H. Ishiguro, "Curiosity did not kill the robot: A curiosity-based learning system for a shopkeeper robot," *ACM Transactions on Human-Robot Interaction*, vol. 8, no. 3, pp. 1–24, 2019.

[64] Y. Girdhar and G. Dudek, "Modeling curiosity in a mobile robot for long-term autonomous exploration and monitoring," *Autonomous Robots*, vol. 40, no. 7, pp. 1267–1278, 2016.