-Conference Proceedings-

# “Do Robots Talk?“-Symposium

AISB 2021 Conference

April 07th – 09th

# Social Robotics and the Nature of Trust: A philosophical investigation into the conceptual challenge of trust ascription in human-robot interaction

Glenda Hannibal[1]

Social robotics is an interdisciplinary field of research that aims to bring robots into everyday life and society more broadly. Inspired by cutting-edge models of human cognition and social compe tence, these robots are designed to have built-in capacities to understand and display cues for social interaction and communication [6, 7, 1]. Robots that are increasingly socially capable and anthropo morphic by design (in both appearance and behavior) make people perceive them as having agency [18]. Envisioned in the role of assistants [11, 19], tutors [12, 13], coworkers [26, 8], and companions [15, 14], these robots are already being tested and utilised in various social settings (e.g., eldercare centres, schools, shopping malls, museums, households, airports, and hospitals).

The prospect of a near-future society that relies heavily on socially capable and anthropomorphic robots [25], beyond the so-called dirty, dangerous, and dull tasks [20], has already raised many ethical and social concerns (e.g., the fear of replacement, the moral status of robots, and social deskilling). One of the ways roboticists have attempted to address these issues has been by studying the role that *trust* can play in human-robot interaction in terms of ensuring acceptance, interaction, and collaboration [16]. The value of studying trust in the field of Human-Robot Interaction (HRI) can be considered as mainly robot-centered or human-centered. The former tends to focus on how to develop robots with computational models of trust to guide their actions and decision-making [24, 23, 3]; the latter explores how the perception of robots as trustworthy is influenced by various factors and eventually human decision to use them [27, 10, 5, 22] in specific domains of application. However, because ”trust” is a commonly used term that seems intuitively understandable, such value oriented or pragmatic focus often uses this concept without much reflection on its meaning and on whether the various meanings of trust (together with the metaphysical and ethical commitments they carry) are applicable to the description of human-robot interaction. This reveals a gap between the descriptive role of empirical studies and their conceptual underpinnings, which sometimes leads to unnecessary misunderstandings in interdisciplinary work. To close this gap, I propose to carry out research into the *nature of trust*, which could clarify the concepts being used in empirical work and consequently provide a more solid foundation. Such work will reveal that speaking of trust in relation to interactions between humans and robots is not as straightforward as it might seem *prima facie*.

My presentation will consist of three steps. First, I want to point out the existence of at least two notions of trust that are used in everyday life and that are also employed, though often in an uncritical manner, in HRI. These are *trust as mere reliance* (understood as a predictive belief or assumption about what will occur given the performance, process, or purpose of robots) and *trust as interpersonal* (understood as a matter of human decision-making and choice based on the ascription of high-level mental states and capacities for moral or ethical reasoning onto robots). In the second stage, I argue that trust ascription in human-robot interaction is facing a conceptual challenge when the distinction of trust into these two notions is not critically reflected upon. On the one hand, when robots are perceived by humans *as if* they have agency [17, 21], the concept of trust as mere reliance is insufficient. On the other hand, using the concept of trust as interpersonal is not warranted as it would require robots to have capacities for higher mental states and moral reasoning, which they do not have. Continuing discourse on trust in HRI without addressing this conceptual challenge, I argue, will commit the researchers to choosing between either the thin and inadequate notion of trust as reliance or the thick and inappropriate notion of trust as interpersonal. Thirdly, through a shift of focus towards what we can reasonably characterize as a situation of trust, I propose that focusing on *vulnerability* as a precondition for trust in human-robot interaction is a constructive strategy for overcoming this conceptual challenge. By emphasizing vulnerability as a precondition for trust in human-robot interaction, in such a way that this aspect does not need to be avoided or completely explained away [2, 9, 4], I will shortly introduce what kind of empirical work can be undertaken from such analysis for future work on trust in HRI.

## References

[1] Breazeal, C. (2003). Toward Sociable Robots. Robotics and Autonomous Systems, 42, 167–175.

[2] Butler, J. (2004). Undoing Gender. New York: Routledge.

[3] Chen, M., Nikolaidis, S., Soh, H., Hsu, D., & Srinivasa, S. (2018). Planning with Trust for Human-Robot Collaboration. Proceedings of the 13th ACM/IEEE

[1] Department of Philosophy, TU Wien, glenda.hannibal@tuwien.ac.at

International Conference on Human-Robot Interaction (HRI), 307–315. Chicago, USA: ACM.

[4] Coeckelbergh, M. (2013). Human being @ risk: Enhancement, technology, and the evaluation of vulnerability transformations. Berlin: Springer (Science & Business Media).

[5] Coeckelbergh, M., Pop, C., Simut, R., Peca, A., Pintea, S., David, D., & Vanderborght, B. (2016). A Survey of Expectations About the Role of Robots in Robot-Assisted Therapy for Children with ASD: Ethical Acceptability, Trust, Sociability, Appearance, and Attachment. Science and Engineering Ethics, 22(1), 47–65.

[6] Dautenhahn, K. (1995). Getting to Know Each Other - Artificial Social Intelligence for Au tonomous Robots. Robotics and Autonomous Systems, 16(2–4), 333–356.

[7] Dautenhahn, K. (1998). The Art of Designing Socially Intelligent Agents: Science, Fiction, and the Human in the Loop. Applied Artificial Intelligence, 12(7–8), 573–617.

[8] Erebak, S., & Turgut, T. (2019). Caregivers' attitudes toward potential robot coworkers in elder care. Cognition, Technology & Work, 21(2), 327–336.

[9] Fineman, M. A. (2008). The Vulnerable Subject: Anchoring Equality in the Human Condition. Yale Journal of Law & Feminism, 20(1), 8–40.

[10] Ishak, D., & Nathan-Roberts, D. (2015). Analysis of Elderly Human-Robot Team Trust Models. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59(1), 65–69. Los Angeles, California, USA: SAGE PublicationsSage CA: Los Angeles, CA.

[11] Iwamura, Y., Shiomi, M., Kanda, T., Ishiguro, H., & Hagita, N. (2011). Do elderly people prefer a conversational humanoid as a shopping assistant partner in supermarkets? Proceedings of the 6th International Conference on Human-Robot Interaction (HRI), 449–456. Lausanne, Switzerland: ACM.

[12] Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: a field trial. Journal of Human-Computer Interaction, 19(1), 61–84.

[13] Kennedy, J., Baxter, P., & Belpaeme, T. (2015). The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI), 67–74. New York: ACM.

[14] Kory, J., & Breazeal, C. (2014). Storytelling with robots: Learning companions for preschool children's language development. Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). Edinburgh, UK: IEEE.

[15] Lee Koay, K., Sverre Syrdal, D., Walters, M. L., & Dautenhahn, K. (2009). A User Study on Visualization of Agent Migration between Two Companion Robots. In J. A. Jacko, C. Stephani dis, D. Harris, D. D. Schmorrow, M. Grootjen, B.-T. Karsh, I. V. Estabrooke (Eds.), 13th International Conference on Human-Computer Interaction (CHII). Vol. 17th, Lecture Notes in Computer Science (5610-5624) and Lecture Notes in Artificial Intelligence (5638-5639). San Diego, USA: Springer.

[16] Lewis, M., Sycara, K., & Walker, P. (2018). The Role of Trust in Human-Robot Interaction. In H. 25 A. Abbass, D. J. Reid, & J. Scholz (Eds.), Foundations of Trusted Autonomy (Studies in Systems, Decision and Control, Vol. 117) (Vol. 177, pp. 135–159). Cham, Switzerland: Springer Open.

[17] Melson, G. F., Kahn, P. H., Beck, A., & Friedman, B. (2006). Toward Understanding Children's and Adults' Encounters with Social Robots. In T. Metzler (Ed.), Papers from the AAAI Work shop on Human Implications of Human-Robot Interaction (HRI) (pp. 36–43). Boston, USA: AAAI Press.

[18] Nyholm, S. (2020). Humans and robots: Ethics, agency, and anthropomorphism. London, UK: Rowman & Littlefield Publishers.

[19] Su´arez Mej´ıas, C., Echevarr´ıa, C., Nu˜nez, P., Manso, L., Bustos, P., Leal, S., & Parra, C. (2013). Ursus: A Robotic Assistant for Training of Children with Motor Impairments. In J. Pons, D. Torricelli, & M. Pajaro (Eds.), Converging Clinical and Engineering Research on Neu rorehabilitation (Biosystems Biorobotics, Vol. 1) (Vol. 1, pp. 249–253). Berlin and Heidelberg, Germany: Springer.

[20] Takayama, L., Nass, C., & Ju, W. (2008). Beyond dirty, dangerous and dull: what everyday people think robots should do. 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI), 25–32. Amsterdam, The Netherlands: ACM.

[21] Turkle, S. (2011). Alone Together: Why We Expect More From Technology and Less From Each Other. New York, USA: Basic Books.

[22] van Straten, C. L., Peter, J., K¨uhne, R., de Jong, C., & Barco, A. (2018). Technological and 28 Interpersonal Trust in Child-Robot Interaction. Proceedings of the 6th International Conference on Human-Agent Interaction (HAI), 253–259. Southampton, UK: ACM.

[23] Vinanzi, S., Patacchiola, M., Chella, A., & Cangelosi, A. (2019). Would a robot trust you? Developmental robotics model of trust and theory of mind. Philosophical Transactions of the Royal Society B, 374(1771), 1–9.

[24] Wagner, A. R., & Robinette, P. (2015). Towards robots that trust: Human subject validation of the situational conditions for trust. Interaction Studies, 16(1), 89–117.

[25] Weiss, A., Igelsb¨ock, J., Wurhofer, D., & Tscheligi, M. (2011). Looking Forward to a "Robotic Society"? International Journal of Social Robotics, 3(2), 111–123.

[26] You, S., & Robert Jr., L. P. (2018). Human-Robot Similarity and Willingness to Work with a Robotic Co-worker. Proceedings of 13th ACM/IEEE International Conference on Human Robot Interaction (HRI), 251–260. Chicago, USA: ACM.

[27] Yuki, M., Maddux, W. W., Brewer, M. B., & Takemura, K. (2005). Cross-Cultural Differences in Relationship-

and Group-Based Trust. Personality and Social Psychology Bulletin, 31(1), 48–62.

# Polite Robots Do It Better

Marta Cristofanini[2], Luca Buoncompagni[3], Fulvio Mastrogiovanni[4]

How a disembodied, artificial intelligence would talk to an assisted elderly with the purpose of establishing a trustworthy, effective verbal relationship? Which persuasive strategy would efficiently address the person in need toward a healthier lifestyle?

To build on a collaborative process in a sensitive context such as healthcare is not a trivial task, especially if the settled goal has to be reached by exploiting only the power of words: there are so many features that should be considered and those aims achieved by human communication spontaneously cannot be taken for granted. There is a different perception of speakers' role involved in this particular linguistic transaction. In the past few years, we have been working on assistive vocal devices specifically trained for interacting with elderly and people with special needs. The last one in this sequence of prototypes is a goal-directed system bounded by fixed contexts of activations and provided with the peculiarity of assuming a *proactive* role: that makes it unique within vocal assistants' scenario.

The content of interaction consists in 66 sentences, semantically nuanced, which would be pronounced in determined contexts of activations. These contexts are related to Activities of Daily Livings' parameters (*ADL*), which highlight in medicine physiological needs to be met: drinking, eating, resting, personal hygiene and walking. When the system, through a series of sensors distributed in the environment and machine learning algorithms, detects one (or more) standard

deviation from the daily routine of the assisted person, it can decide to interact and affect in a positive way the person's behaviour by working as a "cognitive orthotic". Indeed, our conversational design is explicitly shaped on *how* applying principles taken from pragmatics' issues related to *logic of politeness*, *phatic and deictic* expressions (to realize an impossible physical presence of the virtual assistant), *persuasive strategies* of communication, expressed by using

different *implications* within similar sentences. For example, a *foresighted* sentence emphasizes negative consequences which might happen *("Remember, don't wait to be thirsty for drinking something during your day"*), while an *inspirational* one accentuates positive and remunerative feedbacks ("*Good morning! Just a quick reminder to pass through the bathroom this morning: taking care of yourself is a really good way to start your day"*).

That could represent a reusable linguistic knowledge for training empathetic chatbots to gain confidence from human counterpart. All these various domains have been engineered in a specific ontology (using the Stanford software tool named Protégé) by organizing salient areas as different main classes: *Locations*, *Sentences* and *Triggers*.

*Sentences* are divided in following subclasses, namely *SentencesByLocation*,

*SentencesByTriggeringEvents*,

*SentencesByUrgencyValues*. Each sentence is transcribed as a single *instance* (e.g. *"Hey! It's been a while since you had drink something. What about having a glass of water?"*) as long as single trigger (e.g. "drink") and location (e.g. "kitchen"). Their proper belonging to *Sentences* subclasses are inferred by the ontology reasoner.

The ontology is a *dialogue-based interaction* and it allows to specify *object-related properties*, consisting here in: *locationContext* and *triggeredBy*, as well as *data-related properties*, where the *hasUrgency* property is stored, significantly subdivided in *SoftUrgency* and *HighUrgency*.

Sentences marked by the former express a moderate hurry to take a beneficial action, whereas sentences marked by the latter presuppose that the previous prompt has been ignored and therefore it needs a major reinforcement. Two other properties are accounted as data-related properties, namely *immediateAnswer* and *notAnswer*. These are specifically made up for the recognition of a special event, i.e. fall detection and the consequent alert. It could be essential to discriminate between a false alarm, a soft fall or an injured one.

Within the scenario described above, we argue that considering and implementing linguistic pragmatics' challenges in conversational designs could really make the difference in human-computer interaction, specifically in such sensitive contexts as healthcare. Trust starts from a clean, engaging verbal transaction and there are many linguistic tools that can be exploited in order to accomplish it. Even if the relationship is *simulated* by a machine, it's interesting to notice how, with few clues, humans tend to attribute "humanity" and intentionality to almost everything, as an illuminating cognitive experiment by Heider and Simmel (1944) has demonstrated. A machine cannot "be worried" for real, but it can recreate nonetheless an empathetic feeling with the assisted person at an affective-behavioural level: we expect an increase of trustworthiness by the application of these pragmatics shrewdness.

Politeness is the leading path toward collaboration. Even, perhaps especially, in human-robot interaction.

[2] Department of Informatics, Bioengineering, Robotics, and Systems Engineering, University of Genova, martacristofanini@gmail.com
[3] Universita degli Studi di Genova
[4] Universita degli Studi di Genova

# Machines as "Cultural Other": Beyond Anthropocentrism and Exoticism

Min Sum Kim[5]

Humans see in modern technology the horrifying dangers and the bliss enabled by the saving power (Kim, 2019; Kim & Kim, 2013). Entrenched in the emotions of hope and fear towards intelligent machines, humans' attitudes toward intelligent machines are not free of expectations, judgments, strategies, and selfish agendas. This paper analyses the underlying human emotions on machines, where the enduring contradictions and ambivalence, arising from anthropocentric biases, continue to shape people's everyday narratives and experiences. Attitudes toward technology are bound to be affected by the legacies of those ways of seeing. In this way of seeing, humans are a standard by which machinic "Other" could be judged, included, excluded.

Philosophical take on blurring of the boundary between human and machine suffers from the problem of "othering" (Jack, Westwood, Srinivas, Sardar, 2011). Ambivalence and admiration feature in these narratives as core experiences of emotionally charged human-machine relations, and are often reproduced or experienced unconsciously. The concept of *Othering* as the salience of colonial dynamics is crucial for analyses of human-machine relationships, for it illustrates how intelligent machines are represented and are *Othered*. The understanding of Othering in postcolonial studies has drawn significantly upon the work of Said, who relied on Foucauldian and Gramscian work on discourse and hegemony to emphasize the ideological uses of *Othering* (Jack et al., 2011). In this article, the focus is on *Othering* of machines as a highly emotional and embodied experience; viewing machines as object of denigration/fear or object of admiration and infatuation, all of which may be fraught with anxieties and lead to negative emotional interactions.

Based on a close scrutiny of the narratives extracted from contemporary structures of thought, I apply a colonial/postcolonial perspective to demonstrate attendant power relations between humans and machines which are reproduced or subverted in everyday discourse. Colonial lens adds this powerful new dimension to the understanding the machines as reflected in the mirror of the self. The more we work with an awareness of our embeddedness in historical processes, the more possible it becomes to take carefully reasoned ways of dealing with the fundamental dis-ease.

The central argument of this paper is simple: that colonial conceptions of selfhood and otherness, particularly after the rise of colonialism provide a context for the consideration of dominant narratives on Machines. The discourses surrounding machines are very revealing of self and Other. Similar to colonial encounter, hegemonic *assignments* of selfhood, or otherness; the racialized Mechanical *Other* operates like any other marginalized group. Even the deep meditation for the interweaving of humans, machines, and the surround cannot work as it does since the focus is still the *human self* (the dominant in-group).

The asymmetry in power relationships is central to the human construction of *mechanical Otherness*. Just like the "discovery" of the New World has led to the development of an understanding of self as distinct from the other, of "here" as different from "there," narratives on intelligent robots has taken on the similar forms of colonial discourse: contradictory manifestations of ethnocentrism (anthropocentrism) and exoticism (Loomba, 2005). The anthropo-centric bias (similar to ***ethnocentrism*** in colonial/postcolonial discourse) that creates mechanical otherness tend to value humans and distinguish humans from machinic *Others* whom they devalue. On the other hand, fascination with the certain qualities of marvelous and wonder-ful machines (similar to ***exoticism toward natives*** in postcolonial discourse) is characterized by giving rise to new forms of desire projected toward the machinic Other. Not unlike colonial stereotyping and glamorization of natives, even the call for deriving an implicit attunement and gratitude toward machines as most profound layers of prayer and meditation still operates from the vantage point of humans and human interests -- in the service of the spiritual and physical benefits of the Self – humans.

Ambivalence and admiration feature in these narratives as core experiences of emotionally charged human-machine relations, are often reproduced or experienced unconsciously. No matter how profound or spiritual those remedies sound like, one should not miss the fundamental injustice of creating the "Machinic Other" and imposing humans' viewpoints. It is an argument for expanding the purview of machines to take into account the larger historical and cultural forces shaping our understanding of who we are as reflected in the mirror of intelligent machines. To uncover the rootedness of knowledge system in construction of machinic "Other" is to begin to question the typical self-centered ways of being in this world.

[5] Department of Communicology, University of Hawaii at Manoa, kmin@hawaii.edu

## References

Jack, G., Westwood, R., Srinivas, N., Sardar, Z. (2011) 'Deepening, Broadening and Re-asserting a Postcolonial Interrogative Space in Organization Studies', *Organization 18,* 275–302.

Kim, M. S. (2019). Robots as "Mechanical Other": Transcending Karmic Dilemma. *AI and Society, 34,* 321-330.

Kim, M. S., & Kim, E. J. (2013). Humanoid robots as "The Cultural Other": Are we able to love our creations? *AI and Society, 28,* 309-318.

Loomba, A. (2005). *Colonialism/postcolonialism*. London: Routledge.

# Theirs not to reason why: Dialogical reasoning for conversational artificial agents.

Ellen Breitholz[6], Christine Howes[7]

Conversational artificial intelligence (AI) systems are notoriously bad at conversing with humans in a natural way. One of the major reasons for this is that interacting with others frequently involves making common-sense inferences linking context, background knowledge and beliefs to utterances in the dialogue. These inferences are often *enthymematic*, that is, the premises given do not by necessity lead to the conclusion. As discussed by Brandom (1994), in a dialogue, it is important to know what dialogue participants are committed to, which is underdetermined by what they say. This is apparent in the practice of 'giving and asking for reasons' (which often takes the form of 'why?' questions and their responses, Schlöder et al., 2016), as in example 1, below, where the dialogue participants make some of these implicit premises explicit.

Example 1:
Dave: ...you're gonna be home from football until four, you gonna have your dinner, want a bath.
Lee: Yeah, but I might not go to school tomorrow.
Dave: Why?
Lee: Cos of my cough.
Dave: How can you play football and not go to school then?
Lee: Cos I was going out in the fresh air, I'm alright when I'm out in the fresh air.
Dave: So why aren't you going to school then?
Lee: I'm in the class room all day dad.
[BNC KBE 10554-10561]

If a dialogue participant presents the argument "P therefore Q" (as Lee does, when he states that he has a cough so therefore might not go to school tomorrow), an interlocutor must supply a warrant that P is a valid reason for Q in order for the argument to be successful (e.g. if someone has a cough then they are ill and if they are ill then they might not go to school). In rhetoric, these warrants are often referred to as topoi. To produce and interpret enthymemes, interlocutors thus draw on background knowledge or contextual information, and for an enthymeme to be accepted, some such information must be accommodated if it is not already present in the discourse model.

One of the problems for conversational AIs is that the set of topoi accessed by an agent does not constitute a monolithic logical system. This means that in the resources of an agent there can be contradicting topoi, or topoi that lead to contradicting conclusions (Breitholtz, 2014). In addition to this, which topoi apply in a particular situation, and which topos takes precedence over another is relative to the context, including the agent itself. In example 1, Dave invokes another topos that contradicts Lee's reason for not going to school, namely 'if someone is well enough to play football then they are well enough to go to school'. Thus, the pragmatic meaning conveyed by an enthymeme in relation to a listener may differ depending on which topos the listener accesses in the interpretation process.

Understanding how humans reason is important for interactive artificial intelligence (AI) in general whether it uses natural language as such or not. It is important that AI systems are able to explain why certain choices have been made by the system ("explainable AI"). This is a challenge to many current systems in particular those using machine learning, even where they may be able to draw appropriate conclusions, e.g. in the context of medical diagnostic tools (London, 2019).

In this talk we will present an approach to dialogue modelling where enthymemes and topoi play a role for interpretation and production of conversational moves. We will present some phenomena which are frequent in dialogue and how these are related to enthymematic reasoning, and suggest how these may be formalised. This work is intended to provide a basis for building useful context-aware dialogue agents. For such agents to realise their full potential to assist humans in everyday and professional life, they need to be able to reason together with humans through interaction in the form of natural language dialogue.

## References

Brandom, R. (1994) *Making it explicit: Reasoning, representing, and discursive commitment*. Cambridge, MA: Harvard University Press.

Breitholtz, E. (2014) *Enthymemes in Dialogue: A micro-rhetorical approach.* Ph.D. thesis, University of Gothenburg.

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Center Report, 49(1), 15-21.

Schlöder, J., Breitholtz, E. and Fernández, R. (2016) "Why?" In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, pp. 5-14.

[6] Department of Philosophy, Linguistics, Theory of Science, Göteborg University, ellen.breitholtz@ling.gu.se
[7] Department of Philosophy, Linguistics, Theory of Science, Göteborg University, christine.howes@gu.se

# Can GPT-3 Speak?
# Wittgensteinian Perspectives on Human-Machine Communication

Jonas Bozenhard[9]

**Abstract.** There has been a lot of hype surrounding OpenAI's impressive language model GPT-3 ever since it was released in mid-2020. Not only AI researchers and journalists were astound-ed by its spectacular performance at generating human-like text in a wide variety of domains – there was also great astonishment in the philosophical community. David Chalmers, for example, calls OpenAI's autoregressive language model "one of the most interesting and important AI systems ever produced" [1]. This assessment seems appropriate – not least because GPT-3 raises intricate philosophical questions, for instance: "Does the ability to generate 'speech' imply communicative ability?" [2] In other words, do machines that produce text or speech possess linguistic competence and qualify as speakers? More trenchantly put: Can GPT-3 speak? My paper develops a framework of human-machine communication which casts new light on the conversational capacity of GPT-3 and other AI-powered text generators. As a major source of inspiration serves Ludwig Wittgenstein's later philosophy, particularly his much-discussed reflections on rule-following and private language.

Many influential cognitive scientists and philosophers would reject the view that GPT-3 and similar systems possess communicative abilities. Two common strategies for denying machines the ability to speak are what I label the *mentalistic objection* and the *neurocentric objection*.

> *Mentalistic objection*: Advocates of this approach point to the fact that GPT-3 is, as the *MIT Technology Review* puts it, "shockingly good", but "completely mindless" [3]. So, according to this objection, it is impossible for machines to speak since they do not have a mind. Chomsky [4] is arguably the most prominent contemporary defender of this view.

> *Neurocentric objection*: Advocates of this strategy contend that the language models and chatbots available today lack the neural structures distinctive of the human brain that equip humans with the forms of intentionality that are considered necessary for linguistic competence. Searle [5, 6] can be read as a proponent of this critique.

The first part of my paper critically engages with the *mentalistic objection* and the *neurocentric objection*. Drawing on Wittgen-stein's thoughts on rule-following and private language [7, 8], I demonstrate that both approaches exhibit major shortcomings. Therefore, I argue that we ought to reject the anthropocentric view that the ability to speak necessarily presupposes some kind of human-like mental reality or the neural processes character-istic of the human brain.

In the second part of the paper, I present my Wittgenstein-inspired positive take on communicative ability and human-machine communication. The later Wittgenstein conceives of language as a rule-guided activity. However, since the ability to follow the rules of our language cannot be explained with reference to mental or neurophysiological processes, Wittgen-stein offers an anti-reductionist take on rule-following according to which rules are embedded in practice ([7, §§198f., §202]; also see [9, 10]). In dialogue with Wittgenstein's positive remarks on rules, language, and meaning, I sketch a practice-based account of linguistic competence and human-machine communication which is both anti-reductionist and anti-anthropocentric. Accord-ing to this broadly Wittgensteinian framework, it is, at least in principle, possible for machines to speak.

The third part of my paper examines the conversational potentials and limits of GPT-3 against the backdrop of the Wittgenstein-inspired framework I propose. In this context, I discuss the following questions: Does GPT-3 possess linguistic competence? Does OpenAI's third-generation autoregressive language model possess human-like linguistic competence? How to assess GPT-3's communicative ability in relation to human speakers? How to assess its communicative ability in relation to

[9] University of Oxford, UK, email: jonas.bozenhard@queens.ox.ac.uk

other conversational artificial agents such as Joseph Weizen-baum's ELIZA or GPT-3's predecessor GPT-2?

Finally, it shall be touched upon how my Wittgenstein-inspired account of human-machine communication may give new impetus to research in AI and how it helps us – to strike a Wittgensteinian tone – to diagnose and therapeutically treat certain confusions and misconceptions in the field.

## REFERENCES


[1] D. Chalmers, 'GPT-3 and General Intelligence', *Philosophers On GPT-3 (Updated with Replies by GPT-3)*, A. Zimmermann (Ed.), (2020), link: https://dailynous.com/2020/07/30/philosophers-gpt-3/ (accessed 05 October 2020).

[2] A. Zimmermann, 'Introduction', *Philosophers On GPT-3 (Updated with Replies by GPT-3)*, A. Zimmermann (Ed.), (2020), link: https://dailynous.com/2020/07/30/philosophers-gpt-3/ (accessed 05 October 2020).

[3] W. D. Heaven, 'OpenAI's New Language Generator GPT-3 is Shockingly Good—and Completely Mindless', *MIT Technology Review*, 20 July, (2020), link: https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/ (accessed 17 March 2021).

[4] N. Chomsky, *Language and Mind*, Cambridge University Press, Cambridge, (2006).

[5] J. R. Searle, *Making the Social World: The Structure of Human Civilization*, Oxford University Press, Oxford, (2010).

[6] J. R. Searle, 'Biological Naturalism', *The Blackwell Companion to Consciousness*, M. Velmans and S. Schneider (Eds.), Wiley Blackwell, Chichester, West Sussex, 327–36, (2017).

[7] L. Wittgenstein, *Philosophical Investigations*, G.E.M. Anscombe (Trans.), Basil Blackwell, Oxford, (1986).

[8] L. Wittgenstein, *Remarks on the Foundations of Mathematics*, G.E.M. Anscombe, R. Rhees, and G.H. v. Wright (Eds.), Blackwell, Oxford, (1978).

[9] J. McDowell, 'Wittgenstein on Following a Rule', *Synthese*, 58, 325–63, (1984).

[10] B. Stroud, 'Mind, Meaning, and Practice', *Meaning, Understanding, and Practice*, Oxford University Press, Oxford, 170–92, (2000).

# Towards Four-sided Communication

**Michael Straeubig**[1]

## Abstract

Despite recent progress in natural language generation and continued efforts to create operational conversational agents, humans and machines still do not communicate with each other.

Current machine learning approaches, such as transformer models that utilise pre-trained statistical aggregates in combination with clever heuristics, have shown impressive capabilities for text generation. More data and faster computational resources are still driving performance improvements, yet we cannot escape the feeling that we are observing black boxes navigating themselves into local optima.

Traditionally, engineering efforts concerned with natural language processing (NLP) are rooted in linguistics, a discipline that analyses language through abstract data structures, starting from syntax and moving through semantics, usually sprinkling a bit of pragmatics on top.

Existing critique of those approaches rests on the claim that communication requires intentionality and intention in turn requires consciousness, which despite the apparently imminent singularity remains a hard problem where actual solutions aren't available. Unfortunately, the prevalent discourses involve re-iterating discussions about the location of intelligence within Chinese rooms.

I argue that these approaches are still located far from where communication takes place: in the context of an engaged conversation, with a jazz ensemble improvising together, a soccer team coordinating an offense or in a room full of people shouting at each other. These are not syntactical structures, but social systems. Therefore, systems theoretic descriptions may bring forward more fruitful responses to what I have called the communication problem.

In my view, a promising direction arises from operationalising Friedemann Schulz von Thun's model of communication. In this model, each act of communication has four sides, both for the sender and for the receiver: facts, relationship, self-presentation and appeal. These four facets or subtexts appear in almost every message. They can be observed and analysed individually regarding their relative emphasis, or in terms of their congruences or mismatches, respectively.

From the perspective of the sender, the factual side contains the actual subject of the message. The self-presentation side carries both intentional (self-promotions) and unintended (self-revelations) expressions of the sender. The relationship side describes how the sender views the receiver and the relationship between them. Finally, each act of communication also carries appeals - these are actions that the sender intends for the receiver to carry out. Appeals can be communicated openly (advice, a command) or hidden (manipulation).

In this article, I aim to present an initial set of questions that emerge from this perspective. I will largely leave aside the factual side, as it is already well-covered by traditional semantic modelling, assuming some sort of shared "working ontology" between the sender and receiver. It can be asked, if the self-presentation side of a machine is necessarily alien to us or if the machine should simulate a human. If instead of attempting to model human emotions we better adopt a xenofeminist perspective? These considerations also affect the relationship side. Do we insist that an appeal requires intentions, or are we happy to take orders from conversational agents as we are already happy to give orders to machines?

With these considerations I attempt to shift the usual discourse grounded in linguistics and stochastic approaches towards a social systems based four-sided communication model.

[1] Independent Researcher, UK, email: straeubig@gmail.com

## Literature

[1] Dirk Baecker, 'Reintroducing Communication into Cybernetics', *Systemica*, **11**, 11–29, (1997).

[2] Dirk Baecker, 'Systemic theories of communication', in *Theories and Models of Communication*, eds., Paul Cobley and Peter J. Schulz, 85–100, De Gruyter, Berlin, Boston, (January 2013).

[3] David John Chalmers, 'Facing up to the problem of consciousness', *Journal of consciousness studies*, **2**(3), 200–219, (1995).

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *arXiv:1810.04805 [cs]*, (October 2018).

[5] Barbara J Grosz and Candace L. Sidner, 'Attention, intentions and the structure of discourse', *Computational Linguistics*, **12**(3), 30, (1986).

[6] Ray Kurzweil, *The Singularity Is near: When Humans Transcend Biology*, Duckworth, London, 2009.

[7] Laboria Cuboniks, *The Xenofeminist Manifesto: A Politics for Alienation*, Verso, Brooklyn, 2018.

[8] Niklas Luhmann, *Theories of Distinction: Redescribing the Descriptions of Modernity*, Cultural Memory in the Present, Stanford University Press, Stanford, Calif, 2002.

[9] Gary Marcus, 'Deep Learning: A Critical Appraisal', *arXiv:1801.00631 [cs, stat]*, (January 2018).

[10] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, 'Language Models are Unsupervised Multitask Learners', Technical report, OpenAI, (2019).

[11] Friedemann Schulz von Thun, *Miteinander Reden: Störungen Und Klärungen: Psychologie Der Zwischenmenschlichen Kommunikation*, Rororo Sachbuch, Rowohlt, Reinbek bei Hamburg, 1981.

[12] John R. Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, **3**(3), 417–457, (1980).

[13] Michael Straeubig, 'Let the Machines out. Towards Hybrid Social Systems.', in *Proceedings of the AISB Annual Convention 2017*, pp. 28–31, Bath, (April 2017). Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB).

[14] Michael Straeubig, 'The Communication Problem', in *AISB Annual Convention 2019 Proceedings: Philosophy after AI: Language, Imagination and Creativity Symposium*, Falmouth University, (April 2019). Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB).

[15] Ilya Sutskever, James Martens, and Geoffrey Hinton, 'Generating Text with Recurrent Neural Networks', in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 1017–1024, Bellevue, Washington, USA, (2011). Omnipress.