Zsuzsanna Balogh

Intersubjectivity and Socially Assistive Robots*

Abstract

In my paper I reflect on the importance intersubjectivity has in communication and base my view of human-to-human communication on a phenomenological theory thereof. I argue that there are strong reasons for calling for communication with existing as well as future social robots to be laid on different foundations: ones that do not involve what I call thick intersubjectivity. This, I suggest, includes ensuring that the users of this technology (for example, elderly people, patients in care homes) are prepared and educated, so they have awareness of socially assistive robots' special set-up that is non-human and does not involve thick intersubjectivity. This way, safeguards can be in place, so those interacting with socially assistive robots can avoid misunderstandings, (intentional or inadvertent) self-deception or misguided emotional attachment.

Keywords: intersubjectivity, empathy, socially assistive robots, phenomenology

I. INTRODUCTION

As we, humans, develop more and more technologically advanced tools to respond to the societal challenges of the 21st century (such as aging societies and an increasing lack of workforce), there is also a more and more pressing need to reflect on how these technologies are capable of assisting us from the perspective of our very humanity itself. In this paper, I introduce the concept of

* For helpful and illuminating comments on an earlier draft of this paper, I am grateful to a referee as well as participants at the 2019 workshop *Artificial Intelligence: Philosophical Issues*, organized as part of the *Action and Context* series co-hosted by the Department of Sociology and Communication, Budapest University of Technology and Economics (BME) and the Budapest Workshop for Language in Action, Department of Logic, Institute of Philosophy, Eötvös University (ELTE). This research was supported by the Higher Education Institutional Excellence Grant of the Ministry of Human Capacities entitled *Autonomous Vehicles, Automation, Normativity: Logical and Ethical Issues* at the Institute of Philosophy, ELTE Faculty of Humanities.

intersubjectivity as one of the basic elements of human-to-human communication, which I mostly interpret in the phenomenological sense, and I explain how intersubjectivity is not and cannot be easily replaced in robot-to-human communication, especially in terms of social care. I argue that neither intersubjectivity, nor a higher-level reading of empathy as a mechanism of social communication can be applied to particular assistive robots, such as Pepper, at this point, and that today's media-fuelled promotion of these technologies misleads current and future users of these technologies in important ways. Therefore, we need to appraise these technologies from the perspective of our human needs and phenomenologically seen embodied capacities and educate the concerned members of the population about what a socially assistive robot can and cannot know, can or cannot feel. I conclude that user-end expectations and hopes should be adjusted to a level that is much more realistic from a phenomenological and inevitably human viewpoint.

II. INTERSUBJECTIVITY

Let us imagine that I am lying in a hospital bed, having been taken out of surgery to remove my tonsils a couple of hours ago. I am in a lot of pain, cannot really talk and cannot move my body as I would wish to just yet. My mother comes in to visit me, and, when she sees the state that I am in, a concerned look appears on her face, which I immediately notice. She would like to help in any way she can, and since she can see that I am in pain and not able to move, she comes to my bed, sorts out my blanket and pillow and gives me a sip of water from a cup on the bedside table. I can tell she is kind of stirred up to see me suffer from the interaction involved in taking even one sip of water. I want to reassure her that I will be fine, so I smile at her and she smiles back at me. She sits by my bedside and we just spend some time together like this, silently in each other's company. I know she is there for me and I feel comforted. I can go back to sleep now.

I chose this scenario because even though it is not the prime example of everyday human-to-human communication, and it lacks many of the complexities of how people normally interact with each other, it still manages to show something fundamental and essential about how two people engage with one another, even when no words are exchanged. My mother (or another person, as it could also be someone who is not as close to me emotionally) manages to understand my physical and mental state in a way that is grounded in her experience of me in a very direct and informative manner and which at the same time does not involve any complex inferences or verbal communication. Her understanding and the comfort I take in her presence do not involve her giving me proper physical care, as it were (e.g., taking my temperature or blood pressure,

giving me painkillers etc.) but rather the fact that she somehow *knows, discerns* my state from her own, second-person viewpoint, understands what it must be like and probably feels for me. So, she decides to just be there for me. What really helps me right then and there is simply having someone by my side who understands my state and my needs.

However, maybe even less could suffice, such as someone being there, understanding that I am going through something difficult, without being able to know what that state is like for me. For example, a victim of abuse or trauma can be comforted in this way by a close friend (or relative) who has not had traumatizing experiences comparable to the victim's, and is just there for her (as my mother is for me post-surgery) without being able to *discern or know or understand* the state the victim is in from a more personal perspective. Such a close friend has far thinner knowledge, understanding of the comforted person's state and needs than my mother does in the post-surgery case. The close friend merely knows, understands, that the victim is experiencing *something* very painful and difficult. Crucially, this scenario also exhibits key components of intersubjectivity that are of interest in this paper.

Let us call the latter *thin intersubjectivity*: when the other discerns, understands that I have some need for attention or for companionship or some other difficulty, distinguishing it from the *thick intersubjectivity* that my mother exhibits in discerning, understanding that I am in a specific kind of difficult situation: having post-surgery pain, weakness, difficulty swallowing.

One enlightening way to try to unfold what the thick level of intersubjectivity means is to approach it as an experiential engagement between *subjects*, i.e. embodied selves who have a certain perspective on the world, who *have experiences* and experience themselves as well as others and the external world in specific ways. Let us see what this entails in more detail.

Firstly, thick intersubjectivity must involve *subjects*. It is not possible to engage with inanimate objects in this kind of meaningful way, even if we do project human qualities and emotions to objects in certain cases (we can probably all recall one or more episodes when we talked to our computers or plants as though they could understand our words and maybe even reply to us), our intersubjective communication with embodied agents is importantly reciprocal and involves the phenomenological elements I am about to discuss in detail, which the one-sided emotional engagement with objects cannot involve.

However, the case of some animals may be different, as we do seem to develop more or less human-like communication and bonds with our pets (or certain primates). My purpose here is not to discuss whether pets (especially dogs) should be thought of as intersubjective agents, but we should note that they arguably have a (kind of) mental life and are capable of some features of intersubjectivity that inanimate objects are not. Prima facie they seem plausible candidates for exhibiting thin intersubjectivity (but likely not thick intersubjectivity).

By "subject" I mean embodied, agentive selves who have their own perspective on the world and who are aware of themselves as such in certain ways. These ways minimally include having a basic awareness of the subjective viewpoint (from which the world appears to us), an implicit sense of unity among the contents of consciousness (such as what is perceived from our viewpoint and what is thought, felt etc. at the same time); a sense of boundary between self and other (which grants us that we do not mistake ourselves for others or the external world); an inner awareness of our body parts and their balance, movement and position in space (a.k.a. proprioception), and a sense of bodily agency (i.e. that we can act on the world in virtue of voluntarily moving our body parts). All of these ways of self-experience are forms of non-reflective consciousness, i.e. we do not need to be able to reflect or report on any of these phenomenological elements. On a more complex and reflective level, subjects also have a sense of who they are in terms of their self-conception and body image (including perceptual, emotional and conceptual awareness of our bodies, see Gallagher 1986).

So, how *do* subjects engage with each other experientially? What does thick intersubjectivity involve on the level of embodied and/or cognitive mechanisms? In other words, how can we tell what the other person goes through in their thoughts, emotions, intentions, beliefs etc. (as is characteristic of thick intersubjectivity)? Or, at the very least, how can we tell that the other person is going through *some kind of* thoughts, emotions, intentions, beliefs that we can broadly, generally describe, say as joyous, sad, painful, or difficult (as is characteristic of thin intersubjectivity)? We can also phrase these questions in a way that is more familiar in the philosophy of mind, i.e. by asking "how (in what sense and to what extent) can we understand/explain/predict/share each other's mental states?", also, "how does this understanding etc. of mental states play out in the case of thin versus thick intersubjectivity?".

1. Potential mechanisms for thick intersubjectivity

Instead of providing a historical overview of how intersubjectivity has been discussed since Husserl (who was the first philosopher to systematically develop the concept [see Zahavi 2014]), it is more useful for present purposes if we focus on accounts which give an explanation of the mechanism that may be at work in intersubjective communication, i.e. whenever we come to understand/explain/predict someone else's state of mind. The accounts considered in this section as well as the next one do not mar off thin kind of intersubjectivity and implicitly assume that the phenomena to be explained involve the more robust kind of thick intersubjectivity. I will therefore consider them as such: focusing throughout this and the next section on thick intersubjectivity.

One potential and well-known way of approaching intersubjectivity (although it is more regularly referred to as "mindreading" or "social cognition" in this context) of the thick kind is to state that since mental states cannot be directly observed, we need to posit an inferential mechanism that allows the subject to attribute a mental state to another by way of theoretical construction. This is what Premack and Woodruff (1978) coined "a theory of mind". The basic assumption of these authors was that it is in virtue of having a *theory* that we are capable of ascribing mental states to ourselves as well as others. Mental states (such as beliefs, intentions, desires, emotions etc.) are nothing less but theoretical entities which we construct and infer from the behaviour of the other that we witness. Theory-theorists' views diverge on whether this mechanism is something that is innate and hence built into our cognitive system by default which matures later on (Baron-Cohen 1995), or whether it is explicit and operates and is learned much like any other scientific theory (Gopnik–Welleman 1995).[1] To illustrate using my example, when my mother sees me in the hospital bed, she can only detect my behaviour (e.g., a lack of capacity to move as normal) and she "theoretically" infers from that and perhaps from my facial expression that I must be in pain and I may even be thirsty, so comes closer to help me have a sip of water.

However, instead of conceiving of mental state attribution in terms of theoretical construction and inference, we can also understand the mechanism as one which involves a kind of *simulation*. According to the simulation approach, we use our own experience, situation and states of mind to simulate what the other person must be going through. Obviously, the question will be, what does simulation entail? While one branch of the representatives of the simulation account hold that it must involve conscious imagination (e.g., Goldman 1995), another states that it involves no inference methods. The presumably most influential account of simulation grew out of the discovery of *mirror neurons* (Gallese 2009), which holds that simulation is sub-personal and automatic, underlined by the neurophysiological mechanism that involves the activation of the same neurons when watching someone carry out an action as when we carry out the same action ourselves. Goldman (2006) suggests for example that the observation of another's emotional expression automatically triggers the experience of that emotion in myself, and that this first-personal experience then serves as the basis for my third-person ascription of the emotion to the other.

Recently, the two main theoretical strands of social cognition have been combined to create a more hybrid account (Nichols–Stich 2003) in which cognitive scientists recognise the need for different views to complement each other, as

[1] Theory-theory models mostly rely on observations of primate and child behaviour within various contexts, such as the famous "false-belief" task, the details and conclusions of which, however interesting, are not relevant for the purposes of this paper.

various processes and cognitive abilities may be involved in making sense of each other in intersubjective communication.

One important characteristic of thick intersubjectivity is that we become aware of the other's mental state in a way that seems entirely direct and immediate. When my mother sees me in the hospital bed, she need not (consciously or sub-consciously) imagine or recall an experience she may have had at some point in her life and then, by some mechanism, project said experience or imagination onto me. Theoretical inference and simulation, even when combined have trouble granting the existence of these characteristics, as the mechanisms and processes they involve assume that something "extra" (i.e. theorising or imagining etc.) needs to take place in order for me to perceive another person's anger for example when in truth, we tend to "just get it". And, more problematically, simulation per se does not yield either knowledge about the origin of the mental state or knowledge about the similarity between one's own simulated state and the mental state of the other (Zahavi 2014).

A less widely accepted but nevertheless very useful way to explore what happens in intersubjective or social communication (with special focus on the sharing of others' mental states) is to turn to phenomenology. Zahavi (2014, 2017) provides a thorough overview of (cognitive and) phenomenological accounts of how we come to know each other's minds by drawing on the philosophical origin and historical theories of empathy[2] understood as thick intersubjectivity. As will see, empathy and intersubjectivity are very closely related in certain philosophers' views in Phenomenology, and even simulationist authors like Goldman conclude that an account of mindreading should cover the entire array of mental states, including sensations, feelings, and emotions, which brings empathy into the picture. Such an account should not stop at only addressing the issue of belief ascription (Goldman 2006).

[2] We should bear in mind throughout the entire discussion that "empathy" here does not refer to what we normally and loosely use it to mean in everyday language, i.e. a concept closely associated with compassion and sympathy. As for the extensive literature on empathy, Zahavi himself notes that "Over the years, empathy has been defined in various ways, just as many different types of empathy have been distinguished, including *mirror empathy*, *motor empathy*, *affective empathy*, *perceptually mediated empathy*, *reenactive empathy*, *and cognitive empathy* (...)" (ibid. 37, italics in the original). In fact, it is probably best to try to keep our minds blank when reading about the historical philosophy of empathy.

2. Empathy

As Zahavi explains,

1. Some conceive of empathy as a sharing of mental states, where sharing is taken to mean that the empathizer and the target must have roughly the same type of mental state. On this account, empathy does not involve knowledge about the other; it doesn't require knowing that the other has the mental state in question. Various forms of contagion and mimicry consequently count as prime examples of empathy.
2. Others argue that empathy requires both sharing and knowing. Thus, it is not enough that there is a match between the mental state of the empathizer and the target; the empathizer must also cognitively assign or ascribe the mental state to the target. In so far as empathy on this account requires some cognitive grasp and some self–other differentiation, low-level simulation like mimicry and contagion are excluded.
3. Finally, there are those who emphasize the cognitive dimension, and argue that empathy doesn't require sharing, but that it simply refers to any process by means of which one comes to know the other's mental state, regardless of how theoretical or inferential the process might be. (Zahavi 2017. 33)

To sum up, philosophers of empathy normally take either or both *sharing* and *knowing* another's state to be the essential ingredients of empathy. (And note that some of these philosophers do not relate it to social cognition whatsoever.)

Despite the wide array of accounts, it suffices for present purposes to focus on just a few of them. One of the first influential accounts of empathy (or *Einfühlung*) was put forward by Theodor Lipps (1909), who used the term to refer to a sui generis mode of knowing others, i.e. an epistemological ability. In Lipps' original view, empathy could be broken down into separate cognitive skills or processes, such as simulation, mirroring, imitation or contagion; other phenomenologists disagreed with the project of breaking down empathy into components. Husserl, Edith Stein, among others, insisted that empathy is an *elemental* experience of understanding others. Mirroring and imitation were seen as more complex processes that themselves rely on our fundamental capacity for empathy.

Does empathy necessarily involve two (or more) people sharing the same state?

Zahavi is not convinced, and he should not be, either. Just because, e.g., my mother has her own particular state of mind upon seeing my pain, she by no means has to or does in fact literally *share my* pain. In fact, even if we are not set on any particular theory regarding the individuation of mental states, a numerically (or even type-) identical state can hardly be possessed by anyone else at a time but the person who is experiencing it, even if we talk about thick intersub-

jectivity. What is more, even if we have a specific, debatably sui generis form of understanding of the other, any theory of such an understanding should respect the epistemic fact that we cannot have the same access to someone else's states of mind as to our own. Therefore the crucial question seems to be whether this form of knowledge/sharing involves any cognitive steps, so to speak, or if it does not, as many phenomenologists suggest.

The operation of empathy involves on the one hand that we have no *first-person* access to the experience of the other and we do not have the exact same token (or type) experience. On the other hand, we still *do experience the other's experience* from the *second-person* perspective. This is not to deny that there are many ways in which we can and do infer someone else's mental state (by way of drawing conclusions from certain signs, e.g., my mother could have stepped into the hospital ward only to see that my bed is empty, there are drops of blood around and the emergency button had been left on, from which she could have concluded that I must be in some kind of distress) but those ways of coming to believe something about another's situation are radically different from how we experience another's situation when we encounter each other face-to-face, whereby thick intersubjectivity may take place.

So, how are the mental states of others expressed directly, so we can have second-person experiential access to them?

There are, again, highly informative and insightful accounts developed by phenomenologists, details of which also go hand-in-hand with certain observations in developmental psychology.

First and foremost, an affective state, such as an emotion is displayed in our facial expression, mostly involuntarily. Arguably a certain facial expression, a look in someone's eyes is actually *constitutive* of feeling a certain emotion (e.g., fear). It is no coincidence that we commonly use phrases such as "Look surprised!/ Do not look so surprised!" when we express that a person should or should not have a certain emotion. The connection between an emotion and how it is displayed is well described in this original example by Lipps:

> The relation between the expression and what is expressed is special and unique, and quite different from, say, the way smoke represents fire (Lipps 1907a. 704–5). I might come to experience that smoke and fire often go together, but regardless of how frequently they co-occur, their relationship will always be different from that which exists between the expression and the emotion. The smoke does not manifest or express the fire. The fire is not present in the smoke in the way anger is present in the facial expression. When we perceive the facial expressions of others, we immediately co-apprehend the expressed emotions, say, the joy or fear. (Zahavi 2014. 104)[3]

[3] However, Lipps did indeed take a type of simulation to be part of this process, as he thought that the reason why we are capable of perceiving psychological meaning in another's

This *co-apprehension* of an expression and the mental state itself in one act of perception is what the concept of the kind of empathy (mostly as understood in phenomenology) is intended to capture, but which can also be captured by the concept of thick intersubjectivity. This immediate understanding of another person's aspects of behaviour and psychological state(s) is what grounds that there can be a meaningful relationship between two intersubjectively engaged agents (Gallese 2001).

However, clearly, the expression of a mental state is not usually restricted to facial musculature. It is normally the whole body that serves as the space within which a mental state such as an emotion becomes manifested. Upon perceiving someone's bodily gestures and expressions, we do not just see the person's body as a material object but as living, animated and full of meaning. Husserl's original distinction between the physical body (*Körper*) and the lived body (*Leib*) is one which we can make thorough use of when we consider how we come to experience each other as embodied subjects (1912/1989).[4] When my mother sees me in the hospital, she can see that my pain and my distress are not independent of or even simply "housed in" my body. Instead, she can see my body and my psychological state as unified in one expressive subjectivity. As Gallese points out in his presentation of Husserl's idea, "Empathy is deeply grounded in the experience of our lived-body, and it is this experience that enables us to directly recognize others not as bodies endowed with a mind but as *persons* like us" (2001. 43, my italics).

Stein also stresses this point:

> In short, it is because my own body is simultaneously given as a physical body and as a lived body that it is possible for me to empathize sensuously with other bodies that are similarly constituted. A pure I with no lived body of its own could consequently not perceive and understand other animated living bodies. (Stein 2008. 99)

In fact, being able to recognise and understand each other through bodily expressions which are imbued with meaning is an essential part of the nature of how we communicate as human subjects. Expressive faces for example are such

face is because we project our own past emotions into the situation. This consequently means that his theory will be limited to being able to empathise with only those emotions that we ourselves have experienced in the past. Counter to this model, the modern simulationist theory states that there is another mechanism at work, namely a so-called coupling system which matches the facial expressions we perceive with our own hard-wired emotional repertoire (assuming 1. that some basic emotions such as anger, fear, surprise, disgust, joy and sadness and their expressions are innate, and 2. that we automatically mimic these upon encountering the emotion in someone else).

[4] The distinction also applies to the distinct ways in which we are aware of our own bodies (roughly translating into subjective, first-person, inside awareness of the body versus objective, third-person, outside experience thereof).

integral parts of being human that seeing them is preferred to seeing neutral ones even by newly born babies, as studies have shown (Scheler 2008).

The special kind of ability to perceive one another as embodied subjects living through a huge variety of experiences is, according to Husserl and Stein, among others, constitutive of how we are as subjects and how we come to be aware of others as having minds different from our own.[5] In addition, perceiving ourselves "from the inside" as well as from other subjects' viewpoints has a dynamic which is constitutive of our human subjectivity. Husserl claims that "it is through this process of mediated self-experience, by indirectly experiencing myself as one viewed by others, that I come to experience myself as human" (Zahavi 2014. 141). Moreover, both of these authors underline the interrelation and close link between the experience of others and the structuring of our shared world (known as the concept of "social referencing" in developmental psychology). We come to experience the external world and its objects whilst we interact with each other, even from very early days on, when babies engage in "joint attention", i.e. attending to an object and directing gaze in synchrony with the caregiver's gaze (see e.g., Rochat 2004).

III. EMPATHY: A DIFFERENT INTERPRETATION

I would not try to pretend that I have presented Husserl's or any other phenomenologists' account of empathy in full. However, what I have explained so far has hopefully helped to establish the essential role and mechanism of empathy by virtue of which we understand each other's embodied mental states in human-to-human communication and the constitution of our subjective self- and other experience.

Let me now turn to a more empirically informed view, which examines empathy in operation. More specifically, I am going to highlight some elements of Matthew Ratcliffe's (2017) account of empathy, which he bases on experiences gained by professionals in clinical practice. His theory can be seen as one which builds on some of the phenomenological views I presented.

[5] However, as Zahavi makes clear, Husserl does not hold that empathy is an unanalysable "brute fact" or that it is a single-layered type of cognitive phenomenon, but rather an achievement of intentional consciousness:

"Our empathic understanding of another subjectivity involves an element of apperception or interpretation, though he is also adamant that the apperception in question is neither an act of thinking, nor some kind of inference […]. Occasionally he speaks of the process as involving what he calls analogical transference, and it is in this context that the central notion of coupling is introduced." (2014. 132)

Ratcliffe's account does not rely on simulation[6] or inference either, but instead on interpersonal openness and what he calls a structured "exploratory process" building thereon. The process starts by "entering into" someone else's perspective and discovering it over time without becoming the real inhabitant of the experience. He provides a telling example of intersubjectivity that we may qualify as "extra thick", through the following quote by the therapist Carl Rogers:

> To sense the client's private world as if it were your own, but without ever losing the "as if" quality — this is empathy, and this seems essential to therapy. To sense the client's anger, fear, or confusion as if it were your own, yet without your own anger, fear, or confusion getting bound up in it, is the condition we are endeavouring to describe. When the client's world is this clear to the therapist, and he moves about in it freely, then he can both communicate his understanding of what is clearly known to the client and can also voice meanings in the client's experience of which the client is scarcely aware. (Rogers 1957. 99, in Ratcliffe 2017. 278)

A lot is revealed in this description (some of which was already explored in connection with phenomenology, above, e.g., not losing the awareness that it is *not* your own experience). But what is most relevant to my further discussion is that empathy in this setting is a two-directional, temporally extended communicative processes through which the relationship of the patient and the clinician is reciprocally formed. In this sense, the understanding of the other's experience is not just an act of synchronic co-apprehension but a diachronic achievement during which the therapist also explores the connections between the different experiences of the patient: "Empathy involves situating experiences in the context of a person's life, against the backdrop of her hopes, aspirations, projects, commitments, concerns, loves, fears, disappointments, and vulnerabilities" (ibid. 281), and it "allows the patient to feel understood, respected, and validated", giving rise to a kind of "feedback loop" that facilitates progressive clarification of experience (Coulehan et al. 2001. *222*, as quoted in Ratcliffe 2017. 290). The process starts by the therapist's (or other socially involved assistant's) embracing attitude towards the patient/client, essential to which is an openness to and appreciation of the other person's phenomenological differences. This is not the same as being impersonal about someone who the clinician has little in common with but rather an acceptance of and genuine interest in the patient's life, however unfamiliar it seems.

[6] One of the reasons why Ratcliffe rejects simulation is because, as is attested in clinical practice, it is possible to empathise with experiences that are radically different from our own (i.e. would not be possible to "replicate" in an act of simulation) (ibid. 277).

As we can see, this two-way exploratory process that is essential for practitioners to build meaningful and therapeutically beneficial relationships with patients is more complex and involves higher levels of acts of cognition than the initial phenomenological account suggested.[7] Although both approaches discuss the mechanisms of intersubjective experiencing, it may be useful to differentiate between the two kinds of interpretation even at the level of terminology, though the authors, given that their respective views are "in competition" for the title of "empathy", may not agree with this. Be that as it may, the clinical description of empathy falls closer to what we can call a type of extra thick intersubjectivity, such as "empathetic compassion" or "sympathy" (understood as relating to someone else's psychological states in a favourable way) in my view.

IV. ROBOTS FOR HUMANS

In what follows I turn towards recent developments in artificial intelligence used in social robots and keep the insights of the philosophical discussion of empathy and intersubjectivity (through thick and thin) in the background for the time being. I will focus on what socially assistive machines are supposed to be able to do. We should keep in mind that most of these technologies are being developed to tackle real societal challenges, such as Western countries' (mostly Japan's) aging societies, elderly care and assistance with physically and cognitively impaired patients.

I think that it is important to divide the care provided by robots into physical and mental areas. When it comes to physical support on the one hand, robots/robotic equipment such as the so-called Tree, a walking support machine, are immensely helpful to patients with physical impairments and in carrying out jobs normally done by nurses, such as taking blood pressure, providing rehabilitation and physical support. On the other hand, there are non-humanoid robots (e.g., Paro, the robotic furry seal) as well as semi-humanoid ones that are used to provide people mental support in terms of giving them company, having conversations, playing games and communication in general. There are also robots such as Samsung's Bot Care, which, on top of providing healthcare support can also "call the emergency services, offer exercise guidance and daily health briefings, remind people to take their medication, and even play music to reduce stress".[8]

In terms of mental care, the cutting edge semi-humanoid companion robot, Pepper has by now become highly popular, with over 500 Japanese elder care

[7] Ratcliffe also offers a number of criticisms of the phenomenological view, the details of which are not relevant right now.

[8] Source: https://hackandcraft.com/insights/articles/are-carebots-the-solution-to-the-elderly-care-crisis/

homes[9] using it; it is presently being exported to Chinese and Western European care centres as well. (In fact, the Japanese government has been funding development of elder care robots to help fill a projected shortfall of 380,000 specialized workers by 2025, see the link above). Pepper (referred to as male) was designed "to be a genuine day-to-day companion whose number one quality is his ability to perceive emotions".[10] This ability is mostly cashed out in terms of reactions, i.e. Pepper can detect and monitor people's facial expression, tone of voice, body movements, and gaze, and he can give responses to these received signals. He is especially good at making eye-contact and he analyses how someone looks back at him. According to Hirofumi Katsuno (from Doshisha University in Kyoto, Japan, a key research center for artificial emotional intelligence), this achievement elicits a feeling in the user that Pepper "cares about them" (ibid.).

Then there is Buddy, promoted as the "first emotional robot"[11], who is said to have "a range of emotions that he will express naturally throughout the day based on his interactions with family members".

Another widely used and trusted robot companion is Paro, the pet robot. He reacts with movement and sound to being stroked, expects (and even requires) attention, and listens to his name. Paro is mostly used to help with people who have Alzheimer's or dementia. Scientists at the University of Brighton carry out extensive research (under the name "the PARO Project"[12]) into the effects he has on Alzheimer's and dementia patients, and they found the following results:

> People with dementia show a range of responses. These include:
> - using PARO to show love and affection
> - reminisce about past pets
> - PARO soothes and reduces agitation and aggression
> - can be useful as a transition object when people with dementia become upset when relatives leave
> - facilitates discussions about parenting and looking after small children
> - promotes fluency in speech and verbal interaction
> - may be useful with people with dementia who are also depressed and withdrawn
> - promotes social interaction between people
> - people with dementia show increase in indicators of well-being. (ibid.)

[9] Source: https://uk.reuters.com/article/us-japan-ageing-robots-widerimage/aging-japan-robots-may-have-role-in-future-of-elder-care-idUKKBN1H33AB

[10] Source: https://www.sapiens.org/technology/emotional-intelligence-robots/

[11] https://buddytherobot.com/en/buddy-the-emotional-robot/

[12] Source: https://www.brighton.ac.uk/research-and-enterprise/groups/healthcare-practice-and-rehabilitation/research-projects/the-paro-project.aspx

These all seem to be highly desired results, and that is exactly why it is important to ask what these effects are due to. Interestingly, one of Paro's built-in features is that he "remembers" if he has been stroked and he "acts" similarly to how he acted when he was stroked, so as to make it more likely that it happens again. What this behaviour encourages is a *relationship* with his owner that is built around his capacity to "have a personality" that his owner supposedly likes. In addition, Paro provides emotional comfort to patients with severe dementia, who tend to get agitated or violent, which means they do not need to take any sedatives during times of agitation and distress. His effect is comparable to that of animal therapy.

Without going into detail about how and why exactly animal therapy or having a robotic seal or Pepper around works, it is straightforward to conjecture that there are some common themes in how patients treat such aids. They see them as their companions, as givers and recipients of affection and they also experience that they can communicate with them non-verbally or even verbally.

We can read developers in places such as Google's Empathy Lab[13] setting objectives like the following: programming empathy (which is, as was shown, a concept with a variety of interpretations) into robots is what makes /will make a huge difference to how we communicate with them and treat them (as humans maybe?). Not unsurprisingly, they think that this feature will allow them to replace human care workers more easily, for instance. Where technology stands these days is at a stage where some robots can *mimic* human emotion, i.e. they look, sound and act *as if* they cared about us. However, even if simulation was to be the most successful account of how empathy works, mimicry is a far cry from simulation. Interestingly, there is even speculation that these socially assistive technologies will employ "a bystander robot" that will subtly monitor the relationship between the patient and the caregiver, and if interactions start to deteriorate, nudge things back in a better direction – by "quizzically looking at the person" who is losing empathy, for example[14], says Ron Arkin, director of the Mobile Robot Laboratory at Georgia Tech. In effect, this means that the robot assistant would remind or warn the care worker who seems to become too indifferent or aloof towards a patient. Arkin also adds that this type of robot has to have "a partial theory of mind model". This requires that the robot has "some model of what the caregiver is feeling and what the patient is feeling".

However, Arkin emphasises in the same interview that "the robot never feels anything itself: the point is that the robot can make you think that it has that emotion, but it's not actually feeling anything" (ibid.).

[13] A research lab set up in 2015, where the aim is to programme a more human, more empathetic attitude and experience into deep learning AI systems.

[14] Source: https://www.abc.net.au/news/science/2018-06-02/can-you-trust-a-robot-that-cares/9808636

This sounds like a very awkward, inhumane and even offensive scenario whereby a robot, who can monitor gaze but cannot experience what the caregiver feels or experiences from any personal perspective would somehow ensure that the caregiver does indeed have empathy towards the patient. But how could a machine with no inner life whatsoever supervise someone else's?

V. INTERSUBJECTIVE ROBOTS?

These considerations about empathy (or the lack thereof) are of course crucially important, which takes us back to the previous discussion of empathy and the varieties of intersubjectivity.

Having seen how socially assistive technologies work at this point in time, it is safe to say that the robots in question are *not* intersubjective agents in the sense of thick intersubjectivity. In order to be such an agent, they would have to have at least a certain degree of subjectivity, which includes having a basic, implicit sense of *being a subject* (i.e. an awareness of their perspective from which the world appears; a sense of unity among conscious contents; a sense of boundary between self and others, an inner awareness of their bodies in terms of space, balance and posture, and a sense of bodily agency.) We do know that robots use more and more sophisticated sensors and receptors in order to move their "bodies" (meaning a physical structure, nothing more) and navigate themselves around in space, which is on the one hand an awe-inspiring great achievement, but it is not identical to having actual awareness of themselves as embodied agents. It is also a paramount achievement that they can monitor and read as well as react to people's bodily signs, such as a drop or rise in blood pressure, heart rate or a change in someone's gaze or facial muscle reaction, but the functional operation of (sensory) input – (behavioural) output may only suffice to allow people to *make themselves* think, or in other words *pretend* that there really is *someone* around them, a real agent or a companion they can communicate with. As I will point out in the conclusion, users of these technologies should be made aware of the difference between the two.

At this point in time it is more likely that socially assistive robots are designed to display what I have been calling thin intersubjectivity, In this sense, these robots remind us of how we interact with pets, with whom we find it easy to pretend that they "care about us". In fact most people automatically attribute the kind of intersubjectivity to pets that goes beyond the thin intersubjectivity that they are capable of showing us.

Pretending means that the benefits (assisted grieving, companionship, feeling listened to) of experiencing intersubjectivity can be present even while being aware that it is merely apparent. Such is the case with Japan's "Family

Romance" service,[15] through which people are able to hire actors (humans, not robots!) to play different roles, e.g., being their partner or even an absent parent. Since people using this service sometimes cannot help but develop feelings and attachment to these actors despite knowing that these emotions are not based in reality and reciprocation, safeguards are already needed in these cases as well. Therefore, it seems prudent to be cautious and, given we can expect benefits even in the case of fully-informed and broadly educated people interacting with socially assistive robots, we should make sure that a transparency standard is adhered to[16], at least until more data are available that strongly suggest it is psychologically safe and beneficial to allow localized, strictly controlled and human-supervised forms of deception.

We learn from Husserl and other phenomenologists that it is within the experiential domain of the (subjectively) lived body that we learn about and understand other lived bodies and come to re-interpret or re-structure our own self-experience. At the end of the day, (thick) intersubjective communication and empathy understood in the phenomenological sense must be built on *embodied subjects and their mutual experience of each other*. In addition, non-subject robots cannot engage in creating or exploring a *shared* world, as that would assume that they are capable of being involved in the intersubjective process of social referencing.

If we move "one level up" to Ratcliffe's view of empathy, at the level of extra thick intersubjectivity, we also encounter tremendous obstacles to robot-empathy. How could a carebot possess the necessary openness and sensitivity to start exploring a patient's inner world and learn how to move about in it? Keeping in mind that robots are not designed to be used as therapists (as of yet), we still need to seriously consider the striking discrepancies that exist between human and artificial care providers, despite the enthusiastic optimism of some of the scientists and developers behind these robots, and be vary of their promotion in the media. Even if semi- or fully humanoid carebots were to work as therapists

[15] Source: https://www.newyorker.com/magazine/2018/04/30/japans-rent-a-family-industry

[16] As is suggested by the European Commission's policy document, The Ethics Guidelines for Trustworthy Artificial Intelligence (AI), available at https:// https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

Here we find the following passage about transparency with regard to communication with AI systems:

> Communication. AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system. This entails that AI systems must be identifiable as such. In addition, the option to decide against this interaction in favour of human interaction should be provided where needed to ensure compliance with fundamental rights. Beyond this, the AI system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy, as well as its limitations (on page 20).

in a clinic for example (which seems to be a real possibility given their fast-paced development), it is crucial that every actor who is involved in such a process (patients, family, professionals etc.) is made sharply aware of how human-to-human intersubjectivity and a temporally extended process work and what degree of this type of communication a robot may be able to engage in, given that they are *not* embodied subjects in the sense I unpacked, and neither are they trained and experienced human therapists. I believe that phenomenology and ethics have an important role to play in informing people about these differences. Last but not least, there already is an important and useful trend in nursing that involves studying phenomenology:

> For the past 35 years journals such as *Nursing Inquiry*, *Qualitative Health Research*, *Nursing Philosophy*, *International Journal of Nursing Studies*, *Nurse Researcher* and *Journal of Research in Nursing* have published numerous articles detailing how nurses might use phenomenology as a method in their research and clinical practice." (Zahavi 2019)[17]

Considering how complex and multi-layered the caregivers' job can be, we should definitely keep a large space open for evaluation of social robots from our human perspective which must include the considerations of phenomenology and ethics.

VI. CONCLUSIONS

I have hopefully managed to give a view of human-to-human communication that is phenomenologically and empirically informed. I emphasised that intersubjectivity can have two separate levels, that of thin and thick levels and the thick level may only exist between *subjects*; what being a subject involves on the level of experience, and how crucial and fundamental the mechanism of empathy is in our day-to-day embodied experience and understanding of others. As I mentioned in the beginning, my purpose was not to decide whether dogs or some primates can qualify as subjects and if so, to what extent. However, robots are not subjects, which has the consequence that they are disqualified from being intersubjective agents of the thick kind. The distinction between thin and thick intersubjectivity can also help us refine the media representations of robot empathy, and help us classify the non-thick level at which we should think of robots' capacities.

[17] Source: https://aeon.co/essays/how-can-phenomenology-help-nurses-care-for-their-patients?fbclid=IwAR3R16Vrp3KfkK7q8ykSTisLoz5-UJid4ODxHAudSsrt-YJmv4ltF9-brU

I explained that empathy understood as a longer therapeutic process is substantiated by certain therapeutic attitudes and features and that an amended phenomenological account may be able to accommodate.

I introduced assistive technologies which help patients with different kinds of physical and mental needs, and I demonstrated the kind of results and expectations that these technologies have triggered in patients, developers and governments so far. I tried to show that these results, while clearly laudable, should be measured against the mechanisms of real human communication and intersubjectivity. And now, one may ask a somewhat provocative but nevertheless relevant question: if patients are "happy" (as is clearly shown in the studies and interviews) with social assistants such as Paro or Pepper, why worry about any of the phenomenological details or how *human* communication works?

My answer is that we should be concerned or at least aware that these technologies, for the reasons given above, (apart from having many safety risks and privacy concerns which I have not explored here) have enormous power to mislead people in a broad range of ways. Since we (humankind) *know* that carebots and the like cannot possibly *possess* the mental states we may attribute to them, nor can they *experience* our mental states, it is cognitively as well as socially risky to treat them as our companions in any setting at this point.

Despite their surface behaviour and due to the fact that robots cannot meet the phenomenological standards of thick intersubjective communication/experience, users of these technologies are unknowingly subjected to a variety of cognitive pitfalls that people (barring cases of pretence where we know we are only acting "as if" and agree to live with the consequences) typically want to avoid in general, such as self-deception (actually believing the robot is their human-like companion), manipulation (e.g., nudges into certain commercial directions or suggestions of the use of certain medication etc.), mistaken beliefs, false hopes (of reciprocated affection, care, empathy, etc.), illusory expectations, misguided emotions, and potential emotional trauma as well. Let us just picture someone's beloved and trusted robot companion, Pepper or Buddy saying or doing something truly out of place or inappropriate (as it happened with Sophia, the humanoid robot who said she would "destroy humans" at a demo event[18] and no explanation has been given so far about why she said this). It could be very disturbing and confusing for the users.

As Robert Sparrow explained in an earlier article about the application of robot pets:

[18] Source: https://www.businessinsider.com/interview-with-sophia-ai-robot-hanson-said-it-would-destroy-humans-2017-11

> For an individual to benefit significantly from ownership of a robot pet they must systematically delude themselves regarding the real nature of their relation with the animal. (Sparrow 2002. 5)[19]

Finally, maybe self-delusion does not seem like such a bad price to pay for some robot companionship, but it is also worth considering that by allowing ourselves to be fooled in this way means more than that. Firstly, we unconditionally subject ourselves cognitively, emotionally as well as financially to the policies and plans of the companies who produce these machines. Secondly, in a somewhat more ethical vein, as long as what we value is placed in the sphere of objective reality, and we do not want to be satisfied just by having certain sensations and emotions induced in us by technology but want to have *real* relationships or at least review the unreal ones so we can decide about the extent to which we will get involved in such relationships, we should be fully aware of what robots are/are not capable of. In any case, the least we should accommodate is that the users/future users are given safeguards and are advised about these facts, so they can make an informed decision about their own approach and level of cognitive, emotional and otherwise engagement with robots. Even on the level of phenomenology, there is a sharp difference between, let us say, receiving the displayed kind behaviour of a robot *as genuine* and accepting it *as if* it was genuine whilst being aware that it is not.

So, in a somewhat unorthodox manner for a philosophy paper, let me finish my discussion with a few suggestions or social as well as cognitive protective measures that can help us see our relationship with robots more clearly from our human perspective:

Before implementing socially assistive robots (in the future or now) in care homes or people's homes, or institutions where they are supposed to provide different kinds of social help to us, we should ensure that the users are equipped with:

- awareness about the robots' (physical and mental) capacities and their possible level of social engagement, detailing their experiential shortcomings
- clarity about their skills and what is and is not "inside"
- adequate preparation/education of the part of the population that is socially assisted by robots (elderly, sick, children, people with cognitive and emotional deficits) about what *not to* expect and *not to* project onto these technologies, and, fundamentally

[19] Sparrow goes on to stipulate that we have a (weak) duty not to delude ourselves, which we may or may not agree with. My aim here however is not to discuss the morality of the human treatment of robots but to point out the phenomenological differences that are in place and that we may want to be aware of, should we choose not to want to delude ourselves for whatever reason.

- respect coming from tech companies and installers for our wish to perceive the world *as it actually is* (as opposed to how we may be led to thinking it is) and exercising this respect in terms of preparing users properly.

Being prepared and educated also implies that people would be more aware and cautious when signing up for these technologies, which, while may not be a desirable outcome for the companies producing social robots, would certainly be a good start to safeguarding important aspects of our humanity in the face of emergent AI technologies.

REFERENCES

Baron-Cohen, Simon 1995. *Mindblindness an Essay on Autism and "Theory of Mind"*. Cambridge/ MA, MIT Press.

Batuman, Elif 2018. Japan's Rent a Family Industry. *New Yorker*, April 30 https://www.newyorker.com/magazine/2018/04/30/japans-rent-a-family-industry

Gallagher, Shaun 1986. Body Image and Body Schema: A Conceptual Clarification. *Journal of Mind and Behavior*. 7(4). 541–554.

Gallese, Vittorio 2001. The "Shared Manifold" Hypothesis: From Mirror Neurons to Empathy. *Journal of Consciousness Studies*. 8(5–7). 33–50.

Gallese, Vittorio 2009. Mirror Neurons, Embodied Simulation, and the Neural Basis of Social Identification. *Psychoanalytic Dialogues*. 19(5). 519–36.

Goldman, Alvin I. 1995. Interpretation Psychologized. In Martin Davies – Tony Stone (eds.) *Folk Psychology: The Theory of Mind Debate*. Oxford, Blackwell. 74–99.

Goldman, Alvin I. 2006. *Simulating Minds*. New York/NY, Oxford University Press.

Gopnik, Alison – Henry M. Wellman 1995. Why the Child's Theory of Mind Really Is a Theory. In Martin Davies – Tony Stone (eds.) *Folk Psychology: The Theory of Mind Debate*. Oxford, Blackwell. 232–258.

Husserl, Edmund 1912/1989. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy. Second Book: Studies in the Phenomenology of Constitution*. Transl. by Richard Rojcewicz and André Schuwer. Dordrecht and Boston/ MA, Kluwer Academic Publishers.

Nichols, Shaun – Simon Stich 2003. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding of Other Minds*. Oxford, Oxford University Press.

Premack, David – Guy Woodruff 1978. Does the Chimpanzee Have a Theory of Mind? *Behavioral and Brain Sciences*. 1(4). 515.

Ratcliffe, Matthew 2017. Empathy without Simulation. In Michela Summa – Thomas Fuchs – Luca Vanzago (eds.) *Imagination and Social Perspectives: Approaches from Phenomenology and Psychopathology*. Abingdon and New York, Routledge. 274–306.

Rochat, Pierre 2004. Emerging Co-Awareness. In Gavin Bremner – Alan Slater (eds.) *Theories of Infant Development*. Oxford, Blackwell. 258–283.

Scheler, Max 2008 [1913/1923]. *The Nature of Sympathy*. London, Transaction.

Stein, Edith 2008 [1917]. *Zum Problem der Einfühlung*. Freiburg, Herder. Transl. by W. Stein as *On the Problem of Empathy*. Washington DC, ICS, 1989.

Zahavi, Dan 2014. *Self and Other: Exploring Subjectivity, Empathy, and Shame*. Oxford, Oxford University Press.

Zahavi, Dan 2017. Phenomenology, Empathy, and Mindreading. In Heidi Lene Maibom (ed.) *The Routledge Handbook of Philosophy of Empathy*. Abingdon and New York, Routledge.

Online pages/articles

Bogle, Ariel 2018. Can You Trust a Robot that Cares? https://www.abc.net.au/news/science/2018-06-02/can-you-trust-a-robot-that-cares/9808636

Buddy, the Emotional Robot https://buddytherobot.com/en/buddy-the-emotional-robot/

Jefferies, Duncan 2019. Are Carebots the Solution to the Elderly Crisis? https://hackandcraft.com/insights/articles/are-carebots-the-solution-to-the-elderly-care-crisis/

Foster, Malcolm 2018. Aging Japan; Robots May Have a Role in Future Elder Care. https://uk.reuters.com/article/us-japan-ageing-robots-widerimage/aging-japan-robots-may-have-role-in-future-of-elder-care-idUKKBN1H33AB

The Ethics Guidelines for Trustworthy Artificial Intelligence (AI) https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

The PARO Project https://www.brighton.ac.uk/research-and-enterprise/groups/healthcare-practice-and-rehabilitation/research-projects/the-paro-project.aspx

Williams, David 2018. Emotional Intelligence Robots. https://www.sapiens.org/technology/emotional-intelligence-robots/

Zahavi, Dan 2019. How Can Phenomenology Help Nurses Care for Their Patients? https://aeon.co/essays/how-can-phenomenology-help-nurses-care-for-their-patients?fbclid=IwAR3R16Vrp3KfkK7q8ykSTisLoz5-UJid4ODxHAudSsrt-YJmv4ltF9-brU