From Basic Empathy to Basic Trust in Human-Robot Relation: A Phenomenological Proposal

Aboutaleb Safdari[1]

[1] Postdoc Researcher, Institute for Philosophy, Bremen Universität
asafdari@uni-bremen.de

Abstract

There are two types of trust: basic trust (BT) and secondary trust (ST). While ST refers to a rational mental state that is the result of individual-evidential decision making and calculation, BT is a relational state that the subjects experience. In this paper, drawing primarily on resources from the phenomenological-enactive approach to social cognition, I argue that there can be BT in the human-robot relation (HRR). This BT is the result of basic empathy for robots, that has been enriched by a long enough and complicated history of interaction with them. I propose a procedure according to which first basic empathy leads people to experience robots as pseudo-others, resulting in the formation of a thin and simple social relation. Then, through the history of interaction between people and robots, this simple, primary empathic-based social relation evolves into a more complicated and rich form of social relation that fosters the BT.

Introduction

Up until about a decade ago, it was common to see robots performing a variety of tasks in almost every factory. Since then, a fundamental transformation has occurred. The robots have been constantly coming out from these isolated areas, getting closer and closer to our everyday dynamic environments. And, in the not-too-distant future, we will see them involved in almost all of the routine tasks we perform. One example of this is the increasing use of robots in healthcare, where they are being used to assist with surgeries, physical therapy, and other medical procedures. It is also expected that with the advancements in Artificial Intelligence, we will see more and more robots being integrated into our daily lives, performing tasks such as cleaning and cooking. But this is far from the end of the story. They are not only emerging from isolated areas, but they are also finding new virtual bodies known as bots thanks to which they are finding their ways into our so-called onlife. Furthermore, they will be present in the most intimate aspects of our lives, such as sex and care. They will be companions for the elderly, caregivers for children and patients, and perhaps one day our romantic partners. Overall, robots are becoming an increasingly integral part of our society, and their presence in our everyday lives is likely to continue growing in the future.

As a result of this fundamental change, scholars have begun to discuss novel ideas like the "cobot" or "collaborative robot" revolution (Van, 1996) or to confirm the beginning of a new era in which human-computer *interaction* is evolving towards human-computer *integration* (Farooq & Grudin, 2016) and (Brinck & Balkenius, 2020). Machines are partners in this new integration paradigm (IP henceforth), negotiating with people and cooperating with them in meaningful ways, whereas in the interaction paradigm they are mere assistants executing human orders. The IP is expected to continue to change the way we work with robots and the way we understand and experience them.

Trust to robots is one of the most significant changes in the IP. I'll concentrate on this change in this paper and make the case that the nature of trust in the IP has been evolved to what I call basic trust (BT). Furthermore, I argue that the shift toward BT in its turn is result of profound changes in our interactions with robots. To understand this new type of interactions I will employ concepts from phenomenologically inspired enactive social cognition. In section 1 of the paper, the two types of trust are distinguished. In section 2, the paper argues that in the context of the human-robot relationship (HRR) in the so-called IP, a basic form of trust can exist between people and robots. Sections 3, 4, and 5 propose a framework that explains the underlying mechanism responsible for establishing this basic trust in HRR. Specifically, sections 3 and 4 demonstrate how the social relation begins with basic empathy, which leads people to experience robots as pseudo-others. By viewing robots as pseudo-others, people are able to form a simple and primary empathic-based social relation with them. Finally, section 5 argues that basic trust is established through the enrichment of basic empathy with a history of interactions. As people interact more with robots, they develop a more complex and rich social relation, which fosters the development of basic trust. The proposed framework has implications for designing and developing robots that can effectively build and maintain trust with humans, as well as contribute to a better understanding of the complex and evolving nature of the human-robot relationship.

Two kinds of trust

Trust appears to be a complex and multifaceted phenomenon that no single theory can fully comprehend. Generally speaking, it may be said that there are two distinct theories that construct trust in two distinct ways. According to the first approach trust is a *rational mental state* resulted from individual-evidential decision making and calculation. The basic idea is that trust is built on statistical and calculative evidence about the reliability of a person. An individual considers all of these bits of information and then rationally

decides, or calculates, whether the person is trustworthy or not. This is based on the idea that trust is a cognitive-rational decision that is made after evaluating the available evidence. I refer to this sort of trust as a derivative or secondary trust (ST) because I think it is founded on a more basic type of trust, which will be discussed in the paragraphs that follow. ST is roughly comparable to what has been referred to as reliability in the relevant literature (Baier, 1986), (Holton, 1994) and (Hawley, 2014). Or what has been called functionalistic account by (Myskja, 2008). Despite the fact that this approach comes in a variety of forms, (Eikeland & Saevi, 2017) convincingly demonstrate that they are all more or less influenced by game theory and the tradition of rational choice.

In contrast, the alternative phenomenological viewpoint holds that trust is first and foremost experiential and is built through relational experiences, rather than through cognitive or rational processes. In other words, trust is earned through actions and interactions and is built through consistent and positive experiences over time. Some scholars refer to it as operating trust and emphasize that it is the original or constitutive mode of trust (Endreß & Pabst, 2013) and (Endress, 2012). Furthermore, they have argued that it is non-thematic in the sense of being non-negotiable, and non-questionable and something that cannot be subject of any discussion. In other words, this basic form of trust (BT) is not something that can be explained or rationalized and it makes no sense to ask why one chooses to trust (Stern, 2017).

When it comes to this kind of trust, I agree that you can't ask why trust rather than distrust. Because the moment you begin to give reasons, you have moved beyond basic trust and into the cognitive mode of trust. This does not, however, imply that such trust appears out of nowhere. There are undoubtedly some underlying processes that are responsible for its establishment. Indeed, understanding the nature and origins of basic trust is essential to comprehending the dynamics of social relationships, particularly in the context of human-robot interactions in the IP. In what follows, I will make an endeavour to explain this process.

BT in the IP

As stated previously, I believe there can be BT between humans and robots in the IP. Furthermore, in this relationship, BT is not only a possibility but is actually taking place. To support this claim, we must examine the characteristic features of action-interaction between people and robots in the IP. In other words, detecting BT in the human-robot interaction requires looking closely at this action-interaction. To do so, I will start with a position that is somewhat similar to mine, namely a phenomenological-social approach. I will then demonstrate its inadequacies and propose my own positive account.

Mark Coeckelbergh argues in his paper that the nature of trust between humans and technology is similar to that of trust between humans (2012). This implies that the human-robot relation is far more complicated than simple reliance - or ST. To support this idea, he believes we must move beyond instrumentalism regarding technological artifacts, in this case robots, and acknowledge that they are more than just means to an end. To put it another way, if one takes instrumentalism for granted, one must accept that HRR includes mere reliance rather than trust - or mere ST rather than BT. He continues by asserting that we have two approaches for moving beyond the instrumentalist view of technological artifacts.

The first approach, which has its roots in analytical tradition, grants robots artificial agency and treats them as artificial agents. This understanding of robots paves the way for what Coeckelbergh refers to as a *contractarian-individualist* approach to trust in HRR, in which the individual is central, and trust is a result of the individual's attitudes. As such, there are three prerequisites for establishing a trustworthy relation. First, the agents must be able to use language; second, they should have freedom in constructing their relation, as well as some uncertainty about its final outcome and finally, the relation should be a social one in the sense that it is created by the individual agents.

The second approach, developed by philosophers of technology in the phenomenological tradition, primarily Ihde's postphenomenology, suggests that we may perceive robots as more than just machines and instead treat them as animals or even *quasi-others*. This perspective focuses on the ways in which humans interact with technology and how those interactions shape our perceptions and relationships with the technology. It argues that the way we experience and understand robots is not solely based on their functional characteristics. This allows to have a form of "virtual" or "quasi" trust in HRR.

Coeckelbergh, dissatisfied with these approaches, wonders if we can still have basic trust in current robots despite their failure to fulfill these requirements. Here, he proposes the *phenomenological-social* approach as an alternative which emphasizes that trust is given rather than created in social relations. It is a default mode in the sense that in a social relation, the agents do not choose to trust but find themselves thrown into it. Regarding the linguistic prerequisite, this approach possesses the conceptual resources required to encompass both linguistic and non-linguistic trust preconditions. Concerning the freedom prerequisite, trust is largely emergent and the influence of individual freedom is not as necessary as the *contractarian-individualist* approach claims. Finally, when it comes to the third precondition in social relationships, trust is the basic mode *that must be presupposed*. And there's no requirement to view robots as quasi-others when discussing the issue of trust in HRR.

In this regard, I believe the most significant difference between the phenomenological-social approach and the contractarian-individualist approach is the former's strong emphasis on the social aspect of trust, which we might call the thesis of *the primacy of the social*. It is also the difference of this approach with (post-) phenomenological one. This thesis applies not only to human-human relations, but also to human-robot relations. Thus, when it comes to trust in HRR, the social relations are prior to the individual:

> . …if a human-robot relation grows as a social relation, then trust is already there as a 'default' in the social relation… there is no requirement here that the robot

> appears as a quasi-other; the emphasis is not so much on perception but on the relational bond, which is more 'felt' and experienced than seen or acknowledged. Most of the time, no deliberation is needed about who or what to trust (Coeckelbergh, 2012, p. 58).

Coeckelbergh appears to be arguing here that there is BT in HRR, and I completely agree with him. I also agree with him that trust is experienced by the subject more than being acknowledged cognitively. Furthermore, I have complete sympathy with the relational structure he proposes for a trustworthy relation with a robot. However, I believe that the core of his argument, namely *the primacy of the social*, has been left unclear. He emphasizes repeatedly that trust is the default mode in social relations, and thus to have trust in HRR it is sufficient to make it social. As it has been mentioned before, even experiencing the robot as a quasi-other is not required to have this trustworthy relation. One might reasonably wonder what this social relationship is. How can it be established? What mechanisms are responsible for its emergence? Therefore, in this account, the social relation is a mysterious black box, a situation that the subjects have already been given.

Social relation: Opening the black box

How plausible is it to assume a social relation without a counterpart or a quasi-other? It seems that the very nature of social interaction requires the presence of at least two individuals or entities, or more precisely two embodied individuals who exert some degree of mutual influence and communication with one another. Without an "other," there would be no one to interact with and the concept of a "social" relation would not apply. Even in cases where one is alone, one still has a sense of "otherness" within oneself. For example, one might engage in an inner dialogue or self-reflection, which can be seen as an "interaction" with an inner self or an "other" in some sense. It is precisely because of the absence of mutual influence and reciprocal communication that having a social relationship with an inanimate object such as a chair is hard to imagine[1]. It doesn't have the ability to respond, react, or communicate. To put it another way, it's hard to see how an inanimate object like a chair might give rise to the experience of "otherness[2]."

So far, I have discussed the other and its role in social relations without fully clarifying this concept. Therefore, it is reasonable to wonder what I mean by "otherness" in this context. The term "other" refers to any entity or being other than myself that has mental or internal states, which suggests that it has consciousness, thoughts, emotions, or experiences. By treating a different entity as an "other," we acknowledge that it has a subjective perspective or internal experience that is separate from our own. Considering an entity as an other shapes our understanding and perception of it and influences our attitudes and behaviors towards it. From a phenomenological point of view - to which Coeckelbergh also refers - grasping a certain entity as an other takes place on 3 different levels: the *that level* - experiencing an entity as a minded one - the *what level* - determining the other's specific state of mind - and the *why level* - reasoning about the other's past and future mental states - (Zahavi, 2014, pp. 167–168). Imagine you and a friend are having a slice of chocolate cake in a café. In the first place you perceive your friend as an entity *that* has a mind and thus mental or internal states. This allows you to grasp *what* her specific mental states are, e.g. she is enjoying the taste of the cake. Finally, you are able to comprehend *why* she decides to visit this café, for example, because she already knows this café serves delicious cakes and will continue to do so in the future.

Let us continue our investigation by taking a closer look at the most fundamental level; that is, the level at which we grasp an entity as an other experientially rather than cognitively[3]. Here, from a phenomenological standpoint, the action-interaction cycle between the agents is the key to understanding. Following Smith (2010), and based on Husserlian approach, I have argued in my paper that the experience of otherness at this level emerges from a specific action-interaction cycle between the agents known as *harmoniums interaction* (Safdari Sharabiani, 2021). Being harmonious implies that, while the agent's actions and behaviors are constantly changing, these changes or *continuous transitions from phase to phase* are not arbitrary, but rather harmonious. As a result, the participants in the interaction have certain *anticipations* from one another that are constantly fulfilled throughout the interaction. The moment I establish a harmonious action-interaction cycle with an entity, it ceases to be an inanimate object and becomes an *other* to me. In other words, this particular cycle makes me comprehend *that* this entity is a minded creature. This experiential state is referred to as an *empathic* relation with the other.

One might wonder that while we may not typically consider a chair as an other, it is possible to have harmonious interaction with it in the sense that we have certain understandings of how it will response. For example, we *expect* a chair to be stable and support us when we sit on it, or we *expect* that it will topple over to one side if we kick it. However, this interaction does not give us a sense of the chair as a minded entity and it is not considered to be an other in the same way that another human would be. Here I add a second component, that is intrinsically linked to the first: that the horizon of the action-interaction should be *sufficiently vast.* In the case of a chair, our attitude toward it is one of *expectation* rather than *anticipation*: when I sit on it, I do not anticipate it to support me, but I do expect it to. This means that we know, with a high- or maybe full- degree of certainty, how the chair will react when we perform various actions. In contrast, when I interact with the other, I anticipate certain responses. To illustrate the distinction, imagine I'm telling a joke to a friend.

[1] This will be explained in greater detail in the following sections.

[2] However, it is possible to develop an emotional or psychological relationship with inanimate objects, such as through sentimental attachment or through the development of a personal ritual around the object. But these relationships would not be considered social in nature. This point will be clarified in the upcoming parts.

[3] In the next parts, I'll go into more detail about the other two levels.

She could react to my joke in a variety of ways, including laughter, boredom because she has heard it before, surprise, and even sadness because it makes her think of a lost friend. Among these possibilities, I anticipate her to laugh. While, as previously stated, I do not anticipate the chair to support me from a wide range of possibilities, I do expect it to. This is the condition of possibility of establishing a harmonious relation that is non-linear and dynamic- because the horizon of anticipation is broad- in contrast to a definitive relation that is almost linear and pre-defined. Furthermore, It is not merely a quantitatively but qualitatively rich experience. This vast harmonious action-interaction cycle manifests itself in the experiential level to grasping an entity as an other. When it comes to inanimate objects like chairs, the cycle is not vast enough. Thus they cannot be experienced as "others"[4].

So far I have argued that the experience of otherness in its most fundamental level emerges from a sufficiently vast harmonious interaction. More importantly, the aforementioned interaction is inherently social and behaves independently of the individuals. That is, in order to fully comprehend and explain the interaction, you must go beyond the individuals and use a different vocabulary to characterise this new emerging system. Thus, this perspective considers the social system as a whole, rather than just the individual components. For similar reasons, De Jaegher and Di Paolo try to put the concept of interaction, conceived of as something inherently social, at the centre of the phenomenologically inspired enactive theory of social cognition (2007). They propose to call into play dynamical system theory and argue that interaction is not a snapshot event but a *process extended in time with a rich structure*. This rich structure can in its turn be grasped in terms of coordination that is ∀…the non-accidental correlation between the behaviours of two or more systems that are in sustained coupling …∀ (De Jaegher & Di Paolo, 2007, p. 490). More importantly from this perspective, two systems that are correlated with each other construct an emergent system that *has its own life* and autonomy. This emergent system has the capacity to influence the behaviours of both sub-systems[5]. In short, interaction on the one hand is structured through coordination, and on the other hand is continuously structuring the behaviours of its constitutive parts. Furthermore, Thomas Fuchs and Hanne De Jaegher portray a more comprehensive picture of the interaction process as not only having the properties of a dynamical system – coupled, coordinated, synched – but also as enjoying a phenomenal–experiential dimension (2009).

In conclusion, if one has a social relation with an entity, one has already experienced it as the other or a quasi–other. It is true not only in human–human relation, but also in HRR, as I will argue in the following section. This is why, contrary to Coeckelbergh's phenomenological-social approach to trust, I believe the story of BT in HRR begins not with our social relation with robots, but with the moment we experience them as a quasi-other which is the result of our embodied interaction with them.

Social relation with robots

Evidence suggests that individuals have an empathetic attitude toward robots. As an example, electroencephalography (EEG) studies have demonstrated that people have relatively similar empathic reactions to a human and a robotic hand in the same painful situation—here, cutting a finger with scissors (Suzuki et al., 2015). In another study, participants reported having empathetic reactions to a robotic dinosaur named Pleo, both during a "torture" condition and a "friendly behavior" condition (Rosenthal-von der Pütten et al., 2013, 2014). Even in the case of an industrial robotic arm, which lacks a face, a voice, and other identifying human features, people experience the robotic arm as a human arm, when manipulating the movements of it (Hostettler et al., 2022). Because people are certain that these robots are merely inanimate mindless objects this empathic response toward them appears strange to researchers and has led them to believe that it is what people *feel in their hearts* (Gunkel, 2018). However, the surprise will be vanished if we consider the aforementioned phenomenologically inspired enactive approach to social cognition(Safdari Sharabiani, 2021).

Accordingly, I believe that this is the possibility of building a vast harmonious interaction with a robot, causing people to empathically experience them as pseudo-others on that level. This experiential mechanism also clarifies the meaning of the term "pseudo-other" and demonstrates why robots are neither merely things nor perfect others, but pseudo-others. It has been established that the *that* level is an experiential grasping resulted from a certain kind of interaction cycle; at this point, I should add that the two other remaining levels, namely the *what* and the *why* level, are result of more abstract-cognitive social reasoning mechanisms such as simulation, analogy or employing the so-called folk psychology (Wang & Quadflieg, 2015) & (Krämer et al., 2012). People know with absolute

[4] I believe that this argument is sufficient for my purposes, so I will not elaborate on it further. It could be argued, however, that the cycle we establish with a chair is not an action-interaction cycle at all. It is simply a causal-mechanical relation that could be fully explained and understood in a language lacking any sort of mental concepts and vocabulary.

[5] Because the goal of this paper is not to provide a detailed account of this perspective on social cognition, further elaboration on the subject will be avoided. However, there are some intriguing experiments that clearly demonstrate the system's independence. For instance there is an experiment in which the main purpose was to examine unintended coordination between two participants' behaviours (Schmidt & O'Brien, 1997). In this experiment, by using a wrist-pendulum device, each of the two participants was told to find his/her most comfortable tempo to swing his/her pendulum in the first half of the trial (12 seconds). Then in the second half (12 seconds) they were told to look at each other while swinging in their own comfortable tempo. Interestingly, the results of this experiment show that the movements of the participants become coordinated even though they have been told to go ahead at their preferable tempo. Put differently, immediately after the establishment of a perceptual relation – in this case, a visual relation – a kind of unintended coordination occurs.

certainty on the cognitive level that these robots do not have any mental or internal states. Therefore we may conclude that they are not perfect others, as this relation lacks two higher levels of empathy, namely *the what* and *the why* levels. As a result, if we were to question someone who attributes mentality and emotion to a robot- at that level- "what is particularly in this robot's mind", she would either strongly respond with nothing or, in the case of a skeptical philosopher, admit to being unable to adequately answer the question. Thus the possibility of building a vast harmonious interaction with robots is the mechanism allows people to build an experiential-social relation with them as pseudo-others; and this experience is the condition that allows people to ascribe emotions like pain and pleasure to robots.

To be more precise, by “experiential” I mean a non-cognitive relation that cannot be fully explained in terms of cognitive factors. To make this point clear, for the sake of argument let us ignore for the moment the empirical evidence that suggests that prior experience or prior interaction with robots has a positive effect on people's empathic attitude toward them (Sanders et al., 2017) and (Fujii et al., 2021). According to a cognitive approach one may conclude that people's growing familiarity with the robot as a result of their earlier encounters and repeated exposures strengthens their conviction that it is just a robot. You can see that the robot is just a bunch of bolts and nuts with an electronic board below all those fake face expressions, computerized voices, and robotic movements. Thus, from this purely cognitive-technical perspective, a robot will be seen as a collection of hardware components and software algorithms. However, as previously stated, the evidence suggests people can form empathic-emotional connections with robots and be influenced by their interaction. As a result, their relation is much more complex and multifaceted than being captured based on mere cognitive factors.

In addition, from this viewpoint *the social* primarily refers to something inherently non-individual. This means that neither party can take full credit for the relation's evolution or structure. Instead, the relation itself must be viewed as a participatory project in order to be completely comprehended. As a result, we cannot account for it using the attitudes and behaviors of the two separated constitutive parties. In the other words the social aspect of a relationship refers to the interdependence and mutual influence of the parties involved. In the context of human-robot relations, this means that the establishment and maintenance of a social relationship with a robot would involve the collaborative effort of both the human and the robot. If we remind from the previous paragraphs, interaction is a process extended in time with a rich structure; to which we can now add that this rich structure is result of reciprocal interaction between human and robot.

As an illustration, consider two different situations in which Matin (an employee) tries to move a desk from room A to room B: one in which he uses a trolley and one in which he does this in co-operation with a robot. In the former case, he places the desk in a trolley and pulls it from room A to room B. There is no reciprocity, no mutual action-interaction in this case, Matin exerts force to move the trolley and we only have a one-sided mechanical relationship between him and the trolley. In the latter case, however, the robot seeks the best grip and thus positions her body (assuming the robot's voice is female), arms, and legs in a specific manner, requiring Matin to adjust his bodily position. Then, during the relocation procedure, the robot who has been holding the front of the desk notices an obstacle in her path and slows her walking speed, causing Matin to slow down as well. Matin is now tired and wants to take a break; the robot recognizes this and then stops to allow Matin to rest. Finally, as they enter the other room, the robot, who has a complete map of the building in her memory, recognizes that this is not room B and warns Matin that he is in the wrong room. As a result, Matin continues his way to find the right room, which is the next one. This is an example of a structure that is time-extended and has a rich non-linear structure that is not the sole product of either party, a participatory *mutual relationship*. Here, the robot and Matin are interacting in a more complex and dynamic way, with each adjusting their actions based on the actions of the other. It is more like a dance, where both parties are adapting to each other in real-time, rather than a simple mechanical interaction.

More accurately this relation is a *pseudo-mutual relation* rather than a perfect one as the robot is not a perfect equal to the human and may not possess the same level of autonomy, intelligence, and emotions as a human being. The relation is not perfectly mutual from the robot to human side and therefore is limited by the capabilities and limitations of the robot. While the interaction with a robot can be much more complex than with inanimate objects like a desk or chair, it still falls short of a perfect mutual relationship between two human beings. The robot may be able to recognize and respond to certain actions and needs of the person, but it is still limited by its programming and technology. Its horizon of interactions is vast enough to build a pseudo-mutual relation but not vast enough to form a perfect one. In summary, the social aspect of the relation highlights the importance of considering the robot as an active participant in the relationship, rather than simply a tool or an object. By viewing the robot as a pseudo-other, we can better understand the dynamics of the relationship and the ways in which it influences and is influenced by both parties.

Empathic based trust in HRR: the role of history of interaction

So far, a minimalist concept of the social in the sense of a non-individual relationship has been achieved. However, this simplistic concept is insufficient to explain more complex trust relationship between humans and robots. In the other world, one might distinguish two concepts or two levels of the social: the social as a mere non-individual, which is sufficient to explain an empathic relationship with robots, and the social as a rich-perfect social relation, which is required to understand the emergence of trust in the human-robot relationship. What makes this rich-perfect level appear from the first minimalist level? What are the mechanisms responsible for this transition? How does this rich-perfect level produce a trustworthy relationship between humans and robots? Here, I think, the *history of the interaction* plays a key role.

David Krackhardt, albeit in a different context, networked human-human relations, suggests that a history of interaction is necessary for building trustworthy relationships between individuals, and that trust cannot be developed instantly (2003). In the much more relevant context of human-machine interaction, Merritt and Ilgen argue that we should consider the dynamic nature of trust and distinguish between dispositional and history-based trust (2008). Krackhardt refers to trust that one may initially place in another person or machine without any interaction, whereas Merritt and Ilgen refers to trust as being the result of an interaction between a person and another person or machine. From this perspective, there is only dispositional trust at first, which is associated with the person's personality - an extravert has a higher degree of dispositional trust than an introvert - and then, over time and through interaction with the machine, it transfers to a stable history-based trust. In another study, focusing more narrowly on the dynamic nature of trust in human-robot relations, researchers propose that all of the important factors involved change in a non-linear way over the history of interactions (Kaplan et al., 2021). They contend that this is why measuring trust at different points in the interaction yields different, if not contradictory, results.

All of these studies have one thing in common that makes me question them, despite my initial sympathy for this line of thought: they are all overly cognitive. Accordingly, the history of interaction is nothing but exchanging information and gathering evidence. Although Krackhardt contends that time spending creates experience, he defines experience as something informational, allowing a person to know how the trustee will use the confidential information she has shared with him/her (2003, p. 219). Similarly, Merritt and Ilgen make the case that:

> …each interaction provides additional information that can be used to make predictions about machine behavior (2008, p. 197).

Finally, Kaplan (2021) and his colleagues have been repeatedly argued that humans perform calculations during interactions with robots in order to establish trust. In summary, these studies share a common limitation of being overly focused on the gathering and processing of information in interpersonal interactions, treating experiences solely as a means of acquiring information for decision making and prediction and neglecting other important aspects of trust in human-robot interaction namely its experiential aspect.

To accept these overly cognitive readings is to accept that the history of interaction, in all its richness and complexity, can in principle be reduced to a set of informational propositions. If all that matters about the interactions is the information they provide to decide whether or not to trust, or to predict whether or not the trustee is trustworthy, then we can summarise the trustee's past records and behaviours in trustworthy relationships and provide it to the trustor to decide whether or not to trust. Thus, one might wonder why bother designing complicated robots when all we need to do is to provide a brief note - even if fictitious - about the high percentage of successful interactions with a specific robot and then create a trustworthy relation. I think it is evident that while such a brief positive note can certainly be a useful source of information when deciding whether or not to trust a robot, it is not sufficient on its own to create a trustworthy relationship. More importantly, not only is the history of interactions irreducible, but it is also so rich and complex. As a result, some researchers have questioned the ecological validity of a controlled laboratory environment in comparison to a real-world situation when investigating human-robot trust (Flook et al., 2019) and (Baxter et al., 2016). To put it another way, the process of building trust is complicated and involves many different things. It is hard, if not impossible, to fully capture its richness and complexity in an artificial laboratory setting, let alone reduce it to a few informational statements or merely cognitive operations. In HRR, building trust often requires people and robots to interact with each other over time and be willing to be open and vulnerable with each other. This may not be easy to do in a lab setting. This is exactly why there is a growing interest in methods that allow long-term experiments to be conducted in real, wild situations (Bruun et al., 2015) and (Blond, 2019).

A closer examination of the structure of history of interaction in HRR reveals why it is rich, complicated, and not reducible to cognitive operations. From a cognitive standpoint, history is simply interactions between people and a specific robot in a specific context. Participants, for example, interact with a specific robot in a specified task, such as manipulating an object or playing a particular game, and their level of trust is measured by some standard questionaries or interview procedures after repeating the interaction with several participants a number of times. This history is neither complicated nor rich, and it is thus explainable in terms of purely cognitive operations. From a phenomenological perspective, however, the circle of interaction is much wider:

> In the cultural object , I feel the close presence of others beneath a veil of anonymity. Someone uses the pipe for smoking, the spoon for eating, the bell for summoning… (Merleau-Ponty, 2005, p. 405).

Here, Merleau-Ponty proposes that many seemingly irrelevant interactions, such as using a spoon for eating or observing another person using the same spoon in the same way, are indeed part of our history of interaction with others. All these ordinary interactions, which from a cognitive standpoint are completely irrelevant to the establishment of trust—seeing a robot using a spoon or washing the dishes is not evidence to decide whether or not to trust it for a different task—are part of the history[6]. It is because they contribute to the

[6] Due to the dominance of the cognitive approach in the field, these interactions and their impact on trust have received little attention in the literature. In a more or less relevant empirical experiment, the researchers discovered that non-musical interaction tasks, such as the presence of a robot while setting up musical equipment, play a key role in human perception of the robot in the context of human-robot cooperation for doing a musical task or playing an instrument (Savery et al., 2021). This suggests that even seemingly unrelated interactions can shape how humans perceive and interact with robots. There is also evidence that exposure to a properly functioning robot can significantly increase one's confidence in the next different robot they encounter (de Bruijn, 2013). Taken together, these findings underscore the importance of designing

development of a rich, complicated social relationship in which trust plays an essential part. According to this view, when we perceive robots, we don't simply see them as isolated things, but rather as beings situated within their environment and actively interacting with it. By observing how they move and act in their surroundings, we build a rich social relation. Overall, this perspective emphasizes the interconnectedness of people, robots, and their environments as a whole, and takes into account the wide range of embodied interactions that result in a genuine social relation.

So far I have argued that our embodied harmonious interactions with robots lead us to experience them as pseudo-others and to establish an empathic relationship with them; then this primary empathic relationship evolves into a trustworthy one through a rich history of embodied interactions. Thus, the embodied aspect of human-robot interaction plays a crucial role in establishing and evolving the nature of the HRR. Two mechanisms could be proposed here as the underlying mechanisms responsible for this transformation. The first is what could be called *situational pressure.* From this perspective Charles Ess following (Vallor, 2010) argues that:

> …trust as a primary component of our (developing) moral character… in particular, is closely at work with the first virtue of *patience*… (2010, p. 294).

According to this line of thinking, patience (and other closely related virtue of perseverance) can deepen a relation and build trust by demonstrating that its survival is not based entirely on short-term benefits and that:

> Notably, the immediacy and physicality of face-to-face communication often forces us to be patient even when we would rather 'tune out' or 'switch off', to use telling metaphors (Vallor, 2010, p. 165).

In the other world, face-to-face embodied interaction exerts a type of situational pressure that forces humans to exercise patience and perseverance in the context of human-robot interaction, resulting in a richer history of interaction with the robot over time. As humans and robots engage in embodied interactions over time, they may adapt to each other's behaviors and preferences. This mutual adaptation can be seen as a form of co-creation of the relationship between the human and the robot which might lead to the formation of a shared coordination and the effective human-robot collaboration.

The second, more speculative mechanism is related to the first and has drawn on resources from enactivism. It could be hypothesized that the accumulation of interactions gradually changes people's phenomenal experiences from empathic to trustworthy. Although, to the best of the author's knowledge, there is no conclusive argument or empirical evidence in support of this idea, and it requires further investigation and testing, there are some pieces of evidence suggesting a somewhat similar proposal. For example, Paul Bach-y-Rita demonstrated in a series of renowned experiments that by using a tactile-vision substitution system (TVSS) that converts visual stimuli received by a camera to tactile stimuli, and through continuous interaction with the environment, the blind's phenomenal experience changes from tactile to visual:

> …in the process, the students discover visual concepts such as perspective, shadows, shape distortion as a function of viewpoint, and apparent change in size as a function of distance. When more than one object is presented at a time, the subjects learn to discriminate overlapping objects, and to describe the positional relationship of three and four objects in one field (BACH-Y-RITA et al., 1969, p. 963).

Thus, over time, the participants' brains adapted to the TVSS, and they began to experience the tactile information as visual information. This suggests that the phenomenal experiences can be shaped by the accumulation of interactions with the environment, even in cases where the initial experience is quite different from the final experience. While the mechanism behind this transformation in the blind participants' experiences is not fully understood, it provides a potential model for how the accumulation of interactions could shape and transform experiences in other domains, including social interactions. However, it's important to note that the transformation of the blind participants' experiences was not instantaneous, and it required a significant amount of continuous interaction with the TVSS and the environment. It's possible that a similar process of gradual adaptation and transformation would be required for the accumulation of interactions to change people's phenomenal experiences from empathic to trustworthy. Overall, while the example of Bach-y-Rita's experiments is not direct evidence in support of the hypothesis, it does provide some suggestive evidence for the potential of interactions to shape people's experiences from empathic to trustworthy.

In another, more relevant study, researchers have found that a brief interaction with a robot can remove the human bias in the mirror neuron system (Press et al., 2007). Mirror neurons are a type of neuron found in the brain that becomes activated both when an individual performs a specific action and when they observe another individual performing the same action. This type of neuron is part of a larger network of neurons called the "mirror neuron system∀ that is believed to be involved in understanding others' intentions, imitation, empathy, social cognition, and our ability to interact with others. It appears that human actions tend to promote more mirror system activation and automatic imitation than non-biological movements. It is due to the fact that the mirror neuron system has evolved to support social learning and communication, and it is likely that human developmental environments provide more opportunities for individuals to observe and imitate the actions of others, especially those of other humans, than non-biological movements. Research has shown that the mirror neuron system is more responsive to actions that are relevant to social interaction, such as communicative gestures and facial expressions, than to non-biological movements like those of a mechanical device. The

---

interactions that involve richer interactions with robots than simply performing the intended pre-specified functions.

removal of this human bias through a brief interaction with robots could be interpreted as further evidence of the transformative effect of interactions on people's experience of robots.

Concluding Remarks

In this paper, I argue that basic trust *can* exist between humans and robots in the IP. This basic trust stems from our first empathic relationship with robots, in which we experience them as pseudo-others. In turn, the empathic relation is the result of our harmonious interactions with robots, and it is precisely these interactions that are responsible for transforming it into a trustworthy one. Thus, trust does not already exist, nor is it the result of cognitive calculations; rather, it emerges from interaction, or more precisely, the harmonious action-interaction between the human and robot. Accordingly, trust is not the result of human or robot alone but is property of the whole situation. From this perspective, trust in HRR is a matter of degree, beginning with a primary empathic relation and gradually evolving to a mature trustworthy relation in which the subject experiences the robot as a *trustworthy partner* rather than a mere *reliable device*. Furthermore, according to this framework, the body and embodied presence are not only the basis of primary empathic relation, but also the necessary condition for its gradual development.

Now one could possibly argue that not only *can* there be BT in HRR, but that there *should* be BT. Since it strongly binds the participants - in this case humans and robots - I think it could help to establish the IP (Integration Paradigm). So there is a bidirectional relation between BT and IP; while the establishment of IP can help build BT by promoting positive attitudes and behaviours towards robots, BT can also reinforce the IP by promoting an experience of trust and cooperation between humans and robots. This can be beneficial in the long run, as it can help to address concerns and challenges that may arise in the integration of robots into various domains. Overall, I believe that fostering BT in HRR can be essential for building successful and sustainable relationships between humans and robots, and for promoting the adoption and integration of robots into society.

I'd like to emphasise here that I'm not denying the cognitive dimension of trust, either in human-human relations or in HRR. Instead, I have tried to suggest that we should move towards a more comprehensive and unified framework in which both cognitive and experiential sides are considered. This framework draws our attention to one of the necessary parts of the trust-building procedure that has been usually neglected in the relevant empirical literature: the history of interaction in its full and rich sense of the meaning. Recognising the importance of this factor emphasises the importance of consistency and harmony not only in specific tasks, but also in the robot's overall interactions. This approach allows for a more holistic understanding of the factors that contribute to the development of trust in HRR, which can lead to the emergence of a rich and perfect social relationship between humans and robots. By considering both cognitive and experiential dimensions, we can gain a more nuanced understanding of the trust-building process and develop more effective strategies for fostering trust between humans and robots.

Acknowledgment

I would like to express my sincere gratitude to the **Humans on Mars Initiative** for their generous funding, which made this paper possible.

References

BACH-Y-RITA, P., COLLINS, C. C., SAUNDERS, F. A., WHITE, B., & SCADDEN, L. (1969). Vision Substitution by Tactile Image Projection. *Nature*, *221*(5184), 963–964. https://doi.org/10.1038/221963a0

Baier, A. (1986). Trust and Antitrust. *Ethics*, *96*(2), 231–260.

Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., & Belpaeme, T. (2016). From characterising three years of HRI to methodology and reporting recommendations. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 391–398. https://doi.org/10.1109/HRI.2016.7451777

Blond, L. (2019). Studying robots outside the lab: HRI as ethnography. *Paladyn, Journal of Behavioral Robotics*, *10*(1), 117–127. https://doi.org/10.1515/pjbr-2019-0007

Brinck, I., & Balkenius, C. (2020). Mutual Recognition in Human-Robot Interaction: a Deflationary Account. *Philosophy & Technology*, *33*(1), 53–70. https://doi.org/10.1007/s13347-018-0339-x

Bruun, M. H., Hanghøj, S., & Hasse, C. (2015). Studying Social Robots in Practiced Places. *Techné: Research in Philosophy and Technology*, *19*(2), 143–165. https://doi.org/10.5840/techne20159833

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology*, *14*(1), 53–60. https://doi.org/10.1007/s10676-011-9279-1

de Bruijn, M. L. E. (2013). *The Base of Trust in Human-Robot Interaction* [Radboud University Nijmegen]. https://theses.ubn.ru.nl/items/8f1cdd25-a759-4880-8e82-f1f3e58efd15

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences*, *6*(4), 485–507. https://doi.org/10.1007/s11097-007-9076-9

Eikeland, T. B., & Saevi, T. (2017). Beyond Rational Order: Shifting the Meaning of Trust in Organizational Research. *Human Studies*, *40*(4),

603–636. https://doi.org/10.1007/s10746-017-9428-6

Endress, M. (2012). Trust and the dialectic of the familiar and the unfamiliar within the life-world. In H. Nasu & F. C. Waksler (Eds.), *Interaction and Everyday Life. Phenomenological and Ethnomethodological Essays in Honor of George Psathas* (pp. 115–133). Lexington Books.

Endreß, M., & Pabst, A. (2013). Violence and Shattered Trust: Sociological Considerations. *Human Studies*, *36*(1), 89–106. https://doi.org/10.1007/s10746-013-9271-3

Ess, C. M. (2010). Trust and New Communication Technologies: Vicious Circles, Virtuous Circles, Possible Futures. *Knowledge, Technology & Policy*, *23*(3–4), 287–305. https://doi.org/10.1007/s12130-010-9114-8

Farooq, U., & Grudin, J. (2016). Human-computer integration. *Interactions*, *23*(6), 26–32. https://doi.org/10.1145/3001896

Flook, R., Shrinah, A., Wijnen, L., Eder, K., Melhuish, C., & Lemaignan, S. (2019). On the impact of different types of errors on trust in human-robot interaction. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, *20*(3), 455–486. https://doi.org/10.1075/is.18067.flo

Fuchs, T., & De Jaegher, H. (2009). Enactive intersubjectivity: Participatory sense-making and mutual incorporation. *Phenomenology and the Cognitive Sciences*, *8*(4), 465–486. https://doi.org/10.1007/s11097-009-9136-4

Fujii, A., Okada, K., & Inaba, M. (2021). A Basic Study for Acceptance of Robots as Meal Partners: Number of Robots During Mealtime, Frequency of Solitary Eating, and Past Experience with Robots. *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 73–80. https://doi.org/10.1109/RO-MAN50785.2021.9515451

Gunkel, D. J. (2018). The other question: can and should robots have rights? *Ethics and Information Technology*, *20*(2), 87–99. https://doi.org/10.1007/s10676-017-9442-4

Hawley, K. (2014). Trust, Distrust and Commitment. *Noûs*, *48*(1), 1–20. https://doi.org/10.1111/nous.12000

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy*, *72*(1), 63–76. https://doi.org/10.1080/00048409412345881

Hostettler, D., Mayer, S., & Hildebrand, C. (2022). Human-Like Movements of Industrial Robots Positively Impact Observer Perception. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-022-00954-2

Kaplan, A. D., Kessler, T. T., Sanders, T. L., Cruit, J., Brill, J. C., & Hancock, P. A. (2021). A time to trust: Trust as a function of time in human-robot interaction. In *Trust in Human-Robot Interaction* (pp. 143–157). Elsevier. https://doi.org/10.1016/B978-0-12-819472-0.00006-X

Krackhardt, D. (2003). The Strength of Strong Ties : The Importance of Philos in Organizations. In *Networks in the Knowledge Economy*. Oxford University Press. https://doi.org/10.1093/oso/9780195159509.003.0008

Krämer, N. C., von der Pütten, A., & Eimler, S. (2012). *Human-Agent and Human-Robot Interaction Theory: Similarities to and Differences from Human-Human Interaction* (pp. 215–240). https://doi.org/10.1007/978-3-642-25691-2_9

Merritt, S. M., & Ilgen, D. R. (2008). Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(2), 194–210. https://doi.org/10.1518/001872008X288574

Myskja, B. K. (2008). The categorical imperative and the ethics of trust. *Ethics and Information Technology*, *10*(4), 213–220. https://doi.org/10.1007/s10676-008-9173-7

Press, C., Gillmeister, H., & Heyes, C. (2007). Sensorimotor experience enhances automatic imitation of robotic action. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1625), 2509–2514. https://doi.org/10.1098/rspb.2007.0774

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An Experimental Study on Emotional Reactions Towards a Robot. *International Journal of Social Robotics*, *5*(1), 17–34. https://doi.org/10.1007/s12369-012-0173-8

Rosenthal-von der Pütten, A. M., Schulte, F. P., Eimler, S. C., Sobieraj, S., Hoffmann, L., Maderwald, S., Brand, M., & Krämer, N. C. (2014). Investigations on empathy towards humans and robots using fMRI. *Computers in Human Behavior*, *33*, 201–212. https://doi.org/10.1016/j.chb.2014.01.004

Safdari Sharabiani, A. (2021). Genuine empathy with inanimate objects. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-020-09715-w

Sanders, T. L., MacArthur, K., Volante, W., Hancock, G., MacGillivray, T., Shugars, W., & Hancock, P. A. (2017). Trust and Prior Experience in Human-Robot Interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1809–1813. https://doi.org/10.1177/1541931213601934

Savery, R., Zahray, L., & Weinberg, G. (2021). Before,

Between, and After: Enriching Robot Communication Surrounding Collaborative Creative Activities. *Frontiers in Robotics and AI*, *8*. https://doi.org/10.3389/frobt.2021.662355

Schmidt, R. C., & O'Brien, B. (1997). Evaluating the Dynamics of Unintended Interpersonal Coordination. *Ecological Psychology*, *9*(3), 189–206. https://doi.org/10.1207/s15326969eco0903_2

Smith, J. (2010). Seeing Other People. *Philosophy and Phenomenological Research*, *81*(3), 731–748. https://doi.org/10.1111/j.1933-1592.2010.00392.x

Stern, R. (2017). 'Trust is Basic.' In P. Faulkner & T. Simpson (Eds.), *The Philosophy of Trust* (pp. 272–294). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198732549.003.0016

Suzuki, Y., Galli, L., Ikeda, A., Itakura, S., & Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports*, *5*(1), 15924. https://doi.org/10.1038/srep15924

Vallor, S. (2010). Social networking technology and the virtues. *Ethics and Information Technology*, *12*(2), 157–170. https://doi.org/10.1007/s10676-009-9202-1

Van, J. (1996). *Mechanical Advantage: Two Northwestern University engineers are developing cobots -- machines that, unlike robots, cooperate with workers without displacing them*. https://peshkin.mech.northwestern.edu/cobot/chitrib/jonvan.html

Wang, Y., & Quadflieg, S. (2015). In our own image? Emotional and neural processing differences when observing human–human vs human–robot interactions. *Social Cognitive and Affective Neuroscience*, *10*(11), 1515–1524. https://doi.org/10.1093/scan/nsv043

Zahavi, D. (2014). Self and Other: Exploring Subjectivity, Empathy, and Shame. In *Oxford University Press*. https://doi.org/10.1017/CBO9781107415324.004