How Robots' Unintentional Metacommunication Affects Human-Robot Interactions. A Systemic Approach

Please cite this paper as follows:

Bisconti, P. How Robots' Unintentional Metacommunication Affects Human–Robot Interactions. A Systemic Approach. *Minds & Machines,* 31(4), 487–504 (2021). https://doi.org/10.1007/s11023-021-09584-5

Abstract

In this paper, we theoretically address the relevance of unintentional and inconsistent interactional elements in human-robot interactions. We argue that elements failing, or poorly succeeding, to reproduce a humanlike interaction create significant consequences in human-robot relational patterns and may affect human-human relations. When considering social interactions as systems, the absence of a precise interactional element produces a general reshaping of the interactional pattern, eventually generating new types of interactional settings. As an instance of this dynamic, we study the absence of metacommunicative abilities in social artifacts. Then, we analyze the pragmatic consequences of the aforementioned absence through the lens of Paul Watzlawick's interactionist theory. We suggest that a fixed complementary interactional setting may be produced because of the asymmetric understanding, between robots and humans, of metacommunication. We highlight the psychological implications of this interactional asymmetry within Jessica Benjamin's concept of "mutual recognition". Finally, we point out the possible shift of dysfunctional interactional patterns from human-robot interactions to human-human ones.

Keywords: Human-robot interaction; HRI design; HRI social implications; interactionism; social robots; anthropomorphic robots; robotic metacommunication; behavioral inconsistency; nonverbal interaction; nonverbal cues

1. Introduction

Despite the rapid advancement of interactive technologies, increasingly capable of reproducing humanlike interactions in terms of both verbal and nonverbal communication, the relational setting that humans put in place towards interacting robots is not yet clear. Multiple studies confirm that users adopt communication strategies for relational artifacts similar to human-human interactions (Jung & Copp 2003), even developing an emotional bond (Konok et al. 2018, Shibata et al. 1999). Furthermore, experiments from neuroscience claim that neuronal activations in human-robot interactions (HRIs) do not structurally differ from human-human interactions (HHIs) (Von Der Pütten et al. 2014). Therefore, there is agreement that, to a certain extent, the humans' interactional approach to robots is borrowed from HHI, as claimed by Kramer et al. (2012). These considerations can fall within the well-known concept of anthropomorphizing interactive robots.

However, it is not clear to what extent HHI relational dynamics are transferred in HRI and which are, conversely, the HRI specific relational patterns. Although strong resemblances can be found between HHI and HRI relational settings, it is evident that anthropomorphic artifacts do not yet perfectly simulate human interactions. Therefore, the extent to which HRIs constitute original relational settings, with their own distinctive elements, is not yet defined (Kramer et al. 2011). In short, we lack a theory that explains to what extent users replicate HHI relational settings in HRI. In this paper, we address this issue from an interactionist perspective, which is discussed later.

Within this issue, three questions emerge:

1) Which human-human interactional patterns *cannot* be replicated in HRI?
2) Consequently, what human-robot peculiar interactional dynamics arise in HRI?

"This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this article is on Minds&Machines https://doi.org/10.1007/s11023-021-09584-5

3) Can these HRI-peculiar interactional dynamics be transferred and enacted in HHI?

In this paper, we try to answer these questions by analyzing one specific characteristic of human-human interactions: metacommunication, as understood by psychological interactionism. The concept of metacommunication can be summarized by the following definition: metacommunication is "all exchanged cues and propositions about (a) codification and (b) relationship between the communicators" (Ruesch & Bateson 1951). Therefore, metacommunication is a second-level communication that codifies the relation between the participants; it conveys an implicit content of the communication that describes the relation. The metacommunicative content of an interaction is often produced by the relation between verbal and nonverbal interactional schemas. For example: John greets Dan, saying "Oh, I could not wait to see you" in a flat voice and without smiling. The nonverbal registers (voice intonation, mimicry) are in this case denying the verbal content. In other cases, metacommunication works in a subtler way, which we discuss later[1].

In this paper, we will show why metacommunication abilities cannot be replicated in social robots, what type of interactional setting is therefore produced, and how this HRI-peculiar setting may influence HHI.

In the next chapter, we highlight how the current literature tends to consider relevant for HRI only the robot's anthropomorphic behaviors, ignoring the fact that a robot's non-anthropomorphic behaviors and cues may create new interactional patterns. Consequently, in chapter 3, we adopt an interactionist approach, using Paul Watzlawick's (2011) theoretical framework, to understand how robots' lack of metacommunication understanding shapes HRI and produces new interactional dynamics, significantly different from human-human interactions. Chapter 4 highlights, under Jessica Benjamin's (2013) recognition theory, the psychological effects of interactions with robots. This chapter lays

[1] For a more in-depth definition of metacommunication, please see Watzlawick et al. (2011) and Selvini-Palazzoli & Boscolo (1994)

"This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this article is on Minds&Machines https://doi.org/10.1007/s11023-021-09584-5

the foundation of a problem worthy of further analysis: how will HRI impact human relationality in general?

This paper offers a different standpoint in the discussion on the nonverbal cues: current literature considers the lack of robot understanding of verbal and nonverbal cues as a quantitative decrease in robots' interactionality (Satake et al. 2009). Moreover, it focuses mostly on robots' signal processing of humans' nonverbal cues (Mumm & Mutlu 2011). Instead, we claim that robots' nonverbal cues, analyzed within an interactionist framework, may produce qualitatively new interactional settings. Furthermore, the current literature seems to give little consideration to the fact that nonverbal cues assume significance only in relation to each other and with verbal communication, with most of the literature focusing on the analysis of precise nonverbal cues, detached from their relation with the others (Erden 2013, Mumm & Mutlu 2011, Saunderson & Nejat 2019, Walters et al. 2007). Accordingly, this paper underlines the co-implications between HRI and HHI, highlighting the risk that human-robot interactional patterns may be transferred to human-human interaction, affecting users' psycho-relational setting.

The main contribution of our reflection is to underline the systemic nature of communication and nonverbal cues: if analyzed alone, nonverbal cues are signifiers without a meaning. Instead, a systemic view of interactions, where the consistency between different communicative schemas is the crucial element for effective interaction, can be helpful to enhance HRI effectiveness.

2. The Qualitative Relevance of Non-Humanlike Interactional Behaviors in HRI

In this paper, we use the notion of "interactional elements", which we need to clarify before going further. When talking about interactional elements, we include both nonverbal cues and the verbal side of the interaction. Ideally, an anthropomorphic interaction is one where verbal and nonverbal cues are

"This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this article is on Minds&Machines https://doi.org/10.1007/s11023-021-09584-5

humanlike and coherent between themselves. On the contrary, a non-anthropomorphic interaction involves a lack of coherence between two different interactional elements: laughing without smiling, talking without gazing, greeting without moving, and so forth.

An assumption that seems to be implicit in a significant part of the HRI literature is that the relevant interactional elements are only those successfully reproducing humanlike interaction (Duffy 2003, Kiesler et al. 2008, Luria et al. 2019, Yuan & Dennis 2019); non-anthropomorphic cues and behaviors are only considered a limit and a reduction of social interaction. For example, in current social embodied agents, the lack of facial mimicry and body gesture coordination, consistent with the verbal communication, is only considered as a decrease of robot interactivity, which precludes the artifacts from a socially significant interaction (Tinwell & Sloan 2014). Accordingly, the elements of difference between the robot's interactional abilities and humanlike ones become elements of non-interactivity (Bartneck et al. 2020). Therefore, they supposedly consist only in a decrease of interactionality.

We believe that this approach offers a limited view of the implications of robots' communication. On the contrary, for a comprehensive understanding of HRI, we believe that those elements that deviate from, or fail to reproduce, a humanlike interaction are equally relevant in shaping the interaction. They produce distinctive effects in the interaction setting, which do not just *quantitatively* shape the relation as "less interactive"; they also create new interactional settings, *qualitatively* different from those existing in human relationships. In short, each element of an interaction, including the lack or the roughness of an element, is a significant part of the interaction itself.

This claim is based on two complementary principles: first, the well-known Watzlawick axiom stating that it is impossible *not* to communicate (Watzlawick et al. 2011). This means that, for example, the lack of facial expressions does not amount to a "non-communication". Instead, it produces a precise effect in the interaction: it communicates the refusal to use facial expressions in that relationship. This lack modifies the interaction not only in the sense that it

"This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this article is on Minds&Machines https://doi.org/10.1007/s11023-021-09584-5

diverges from a “richer” human interaction. In fact, the absence of mimicry will be considered as a communicative (and significant) element of the interaction by the human user. Not to communicate means to communicate a refusal, and therefore counts as the communication of something – although Watzlawick’s axiom should be reframed to “humans cannot *not* interpret on the metacommunicative level”, which we explore in the next chapter. Now, we want to tackle in advance a possible criticism of the application of Watzlawick’s rule for communication in HRI. The criticism may be formulated as follows: humans interacting with a robot are perfectly aware of its artificial nature, therefore they know that a robot is not refusing to make facial expressions but is simply not able to, so it does not count as a communication refusal. Certainly, the evident artificial nature of current relational robots decreases the significance of a robot’s “lack” or incoherence in communicating. In the HRI literature, we can find examples of humans “repairing” robots’ lack in communicating (Baker et al. 2018, Plurkowski et al. 2011, Sebo et al. 2019), and this seems to regard mainly trust repair. Nevertheless, we believe that the more the robot is anthropomorphic, the less humans will want to “repair” the interaction. In fact, the more the robot is able to produce a humanlike interaction, the less the non-anthropomorphic interactional elements will be evident and consciously “repaired” by the human. We believe that Serholt’s (2018, p. 14) findings support this claim in an interaction experiment between children and a robot:

“Given the robot as a proficient participator in a learning activity, its lack of social interaction skills, non-existent cooperation skills, while its humanoid appearance provided indications of the contrary, the whole interaction situation became paradoxical. As this study demonstrates, children expected the robot to be able to interpret their intentions, much as human teachers do. When the robot was unable to do so, children either tried to compensate for this themselves, or the interaction broke down.”

It is probable that children are more likely to attribute humanlike interactional abilities to anthropomorphic robots, but Serholt’s claim might be true also for adults when the anthropomorphism is more sophisticated.

“This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this article is on Minds&Machines https://doi.org/10.1007/s11023-021-09584-5

In short, we argue that humans' communication repair strategies decrease when the level of the robot's anthropomorphism increases. This assertion could be linked to theories suggesting that the uncanny effect in HRI is related to conflicting cues, to which the user attributes opposite interactional meaning (in Kätsyri et al. 2015): the "uncanny" effect due to conflicting cues happens only at a certain level of anthropomorphism. Furthermore, the unwanted effects on humans of robots' inconsistent behaviors are widely discussed in HRI literature: the management of proxemics (Mumm & Mutlu 2011); the importance of gazing (Bartnek 2020); the anxiety provoked by incoherent gazing (Ivaldi et al. 2017, Nomura et al. 2019). This supports the assertions made above: that every element of interaction, including communicative "lacks", is relevant and actively modifies the interaction; accordingly, humans' emotional reactions to robots are widely documented to be linked with verbal and nonverbal "leakage" (Mutlu et al. 2009).

Secondly, we agree with Seibt's (2016) argument that there is no room for fictionality in social acts. Social actions are performative and therefore their significance relies only on the fact that they have been performed. In fact, the communication recipient will rely only on the phenomenological observation of the performance in order to attribute significance to the action. The act of pretending to greet someone, from the observer's point of view, cannot be distinguished from the act of greeting itself. Since a social action cannot be fictionalized in absence of a clear formalization (such as in a theatre play), actions in HRI cannot be considered as "fictions of actions" and therefore less significant for the interaction. Moreover, as stated before, multiple considerations support the argument that humans approach HRI with a setting similar to HHI (Kramer et al. 2012). This claim, combined with Watzlawick's and Seibt's considerations on the nature of communication, leads us to conclude that users' awareness of the simulated nature of robot interactionality is not highly relevant from a pragmatic-interactional point of view. This consideration acquires more and more importance as the level of anthropomorphism increases.

"This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this article is on Minds&Machines https://doi.org/10.1007/s11023-021-09584-5

Therefore, we draw a preliminary conclusion from these considerations. As long as the design process is only concerned with the simulation of a humanlike interaction, those interactional elements, far from being anthropomorphic, will be disregarded as insignificant (or, at the most, significant in the form of the decrease of effectiveness). On the contrary, we claim that every action (or lack of action) performed during social interaction is always meaningful for people who are interacting. This significance acquires greater relevance with the increase of a robot's interactional anthropomorphism.

We share Seibt's (2017) suggestion to approach HRI as asymmetric relationships, where "the relevant capacities for sociality are not distributed symmetrically over the interacting systems". The concepts of asymmetric relationship, and the considerations made so far on the relevance of non-anthropomorphic social acts, give an account of HRI beyond a classic philosophical approach that distinguishes only between intersubjective and objectual relations (Hegel 2018). Intersubjectivity presupposes a set of shared attributes (agency, intentionality) among the interaction participants, where the absence of these attributes would qualify it as a non-interaction. Instead, the concept of asymmetrical relationship allows us to consider all the degrees of interactivity, from the total absence of human likeness to the perfect simulation (Seibt 2014), as equally relevant and interactionally significant. Therefore, on the ontological level, human-robot interactions constitute asymmetrical relations. On the phenomenological level – the one perceived by the user – the more the robot is anthropomorphic, the more the relationship is structured as symmetrical, namely anthropomorphic and simulating a human-human interaction. As we show in the following paragraphs, this gap should be bridged to ensure psychologically functional human-robot interactions.

The concept of asymmetric interaction has been further developed by Seibt et al. (2020) – and other scholars linked with the context of Aarhus University's reflection on Robophilosophy – in the concept of sociomorphing, as opposed to the one of anthropomorphizing. In Seibt's paper, it is suggested that users, when interacting with social robots, do not always project human attributes onto the robot. On the contrary, they tend to "sociomorphize" the interaction, following the

expansion matrix of asymmetrical interactions design by Seibt (2014). This insight suggests that humans tend to adapt to the asymmetricity of the interaction when the robot fails to provide an anthropomorphic interactive behavior. We believe that these results are very relevant for this paper, and share a similar approach: "The claim that human social interactions with (animals and) robots are based on sociomorphing, i.e., on the *perception of actual non-human social capacities,* is a claim about the perception of the *manifestation of a capacity*" (Seibt et al. 2021, p. 14). However, Seibt also states that the sociomorphing of robots in most cases happens preconsciously. Therefore, the sociomorphing of social robots can be understood as a substitutive interactional dynamic to "recover" the asymmetricity of the interaction. Seibt compares this pattern to the one enacted in relating with pets; yet, unlike pets that are not anthropomorphic by design, social robots may more or less mimic an anthropomorphic interaction:

"To the extent that the interactive capacities of social robots differ from each other, from animals, and from humans, different types of social robot may afford (relative to interaction context) distinctive new types of sociomorphing and associated new types of experienced sociality." (Seibt et al. 2021, p. 14)

Therefore, we believe that the equation between pets and social robots is partially tenable, since (most) social robots are anthropomorphic by design. We claim, but it should be tested in future experimental works, that the more the robot is anthropomorphized by design, the less an asymmetric setting will consciously take place in the interaction. The less the robot is anthropomorphic by design, the less users will assume the symmetricity of the interaction. Instead of an asymmetric sociomorphing of HRI, we may observe other types of symmetric substitutive dynamics, which we discuss in the next chapters.

We have suggested that when robots fail in reproducing anthropomorphic interaction, this remains a meaningful interactional element and qualitatively modifies the interaction setting.

After establishing these theoretical coordinates, in the next chapter, we investigate the effects of non-anthropomorphic cues in interacting with anthropomorphic-designed robots: what happens when users interact with an

anthropomorphic social robot with a non-humanlike coherence between verbal interactions and nonverbal cues? We show that this absence produces the emergence of an HRI-peculiar interactional dynamic that does not replicate a human-human one. Finally, we hypothesize on the consequences from the pragmatic point of view of the single interaction. Furthermore, we highlight the possible modification of the user's psycho-relational setting.

3. Interactional Consequences of Metacommunication Absence in Social Artifacts

One of the axioms of interactionism, developed by the "Palo Alto school" (Ruesch & Bateson 1951), is the distinction between two fundamental aspects of human communication:

"Every communication has a content and a relationship aspect such that the latter classifies the former and is therefore a metacommunication." (Watzlawick et al. 2011, p. 51)

This theory, developed by Bateson (Ruesch & Bateson 1951), considers language as a set of "messages" and "commands". Metacommunication ("command") is an element present in every communicative act, verbal or otherwise, which informs the recipient on how the emitter perceives the relationship itself. Implicitly conveyed in all interactions, metacommunication allows the interaction participants to cooperatively define the structure and nature of the interaction. Metacommunication is generally defined as "communication about communication itself" and is conveyed through verbal and nonverbal cues. We will now define three fundamental concepts characterizing "metacommunication" inside the interactionist theory: the command level, the feedback circuit, and the systemic nature of communication.

"This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this article is on Minds&Machines https://doi.org/10.1007/s11023-021-09584-5

The concept of “command level” summarizes our previous discussion, when we defined metacommunication as a second-level communication that codifies the relation between the participants. This level conveys an implicit content of the communication, not always directly related to the explicit content, that describes the relation between the participants. For example: someone says “glad to see you” while nervously tapping their fingers, rolling their eyes and with a flat, annoyed voice; all these nonverbal cues are implicitly denying the verbal act of greeting. This level, the one of metacommunication, is highly autonomous from the interaction semantic content (message): sentences that convey the same informative content may have extremely different metacommunicative meanings because of the tone used, the pauses, the facial expressions, gestures and the construction (or “punctuation”) of the sentence (Watzlawick et al. 2011, p. 54).

Moreover, in interactionist theory, language is considered as a feedback circuit, where each metacommunicative element, thus on the command level, acquires meaning on the basis of the response (or “recognition”) it receives from the other interacting individual, therefore retroactively. This means that the intentionality of conveying a specific metacommunicative content is not relevant because the recipient is the one who signifies this content[2]. The concept of feedback circuit is important in order to understand that, in an interaction, meanings are co-constructed by the participants: the metacommunicative content conveyed by X to Y will be reshaped in its meaning on the basis of Y’s response, and vice versa. The importance of recognition is at the heart of many psychoanalytic theories, such as the “recognition theory” (Benjamin 2013, p. 12), and philosophical theories, from Hegel (1807) to Kojeve (1980) to Honneth (1996).

The third key concept is that interactions must be considered in their systemic nature: the effects of each interactional act (e.g. a precise facial expression)

[2] Watzlawick (2011, pp. 187-229) writes about this point in a chapter summarizing “communication failures” due to communicative misunderstanding, involving a discussion on the role of metacommunication in humorous sentences.

change the balance of the entire system-relationship and, retrospectively, may change the meaning of past interactions.

So far, we have discussed the widely accepted assumption that humans largely adopt the same interactional settings of HHI when interacting with robots, including emotional attachment and the artifact inclusion within the subject's personal and intimate life aspects (Turkle et al. 2006a); these considerations fall within the concept of robots' anthropomorphism in HRI. We have also discussed Watzlawick's argument about the impossibility *not* to communicate, and the non-fictionality of social interactions in Seibt. We concluded that every action performed in an interaction is meaningful for the recipient of the communication, in the peculiar way of a feedback circuit. Therefore, we claim also that the expectation of metacommunication understanding is transferred from HHI to HRI, given a high level of robot anthropomorphism. This means that a human, when interacting with a robot, implicitly conveys a definition of the relationship that is taking place, from which she expects a response. However, the robot does not understand the metacommunicative level conveyed in the interaction. The understanding of an implicit, and often non-linguistically formalized, interactional element is something unthinkable even for the most modern systems. They might be able to recognize a single nonverbal cue, but they are not able to manage the complexity of the systemic nature of interactions. Nonetheless, when responding on the level of the explicit content (the message), the robot will also respond on the metacommunicative level (the command). Indeed, the inability of robots to communicate at that level is inexpressible. In this regard, we should reframe Watzlawick's axiom: humans cannot *not* interpret communication on the metacommunicative level. This clarification is necessary in order not to ontologically attribute intentionality to robots: here we are only concerned with the user's point of view of HRI, ascribing significance to the robot's behaviors and a metacommunicative level to its interactions.

In conclusion, any type of response the robot provides at the message level will also be interpreted on the metacommunicative one. We believe that this claim is widely supported by the literature on HRI, even if the role of the

"This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this article is on Minds&Machines https://doi.org/10.1007/s11023-021-09584-5

metacommunicative level has never been explicitly tackled. The literature reports a multiplicity of different reactions (happiness, anxiety, disgust, annoyance) by subjects interacting with relational artifacts, apparently for non-explicit reasons (Nomura et al. 2004, Turkle et al. 2006b). The discussion about seemingly unintentional cues (Mutlu et al. 2009) supports the case that humans attribute intentionality to nonverbal cues, such as eye movement. Other scholars (Dautenhahn et al. 2006; Mumm and Mutlu 2011; Walters et al. 2005) underline the importance of personal space invasion in HRI effectiveness; Syrdal et al. (2006) discuss users' comfort related to different directions of a robot's approaches. The current literature seems to consider each element (personal space, glance, approaching direction, velocity of movements, facial mimicry, etc.) separately, enquiring the effect on the interaction of one precise element. There is the assumption that we can simply put together all the elements providing a satisfactory interaction: eye contact, lateral approach (Dautenhahn et al. 2006), not violating personal space (Syrdal et al. 2007), use of the body, facial expressions, voice intonation (Briggs and Scheutz 2016), pauses in speech (Brinck and Balkenius 2020), and so forth. However, although a simple addition of effective interactional elements may work out for very simple interactive robots, when the interaction becomes complex, the systemic nature of interactionality shows up. This is the difference between nonverbal cues and metacommunication: every nonverbal cue conveys a precise metacommunicative level that acquires meaning only in relation to the other elements of the interactional system (Selvini-Palazzoli et al. 1988). For example, eye contact is widely considered to be one of the fundamental elements for a robot to establish significant social interactions (Scassellati 1998). However, eye contact is also considered a factor causing relational anxiety (Schneier et al. 2011). As human beings, we are able to understand whether a subject likes eye contact by a series of signals (the interlocutor looks away, moves sideways, stops communicating, etc.). We go one step further, supporting that the consistency between different communicative schemas is a crucial element for effective interaction. Therefore, in analyzing the effectiveness of a specific interactional behavior such as making eye contact,

we must be aware that in some cases it enhances the interaction effectiveness: for example, making eye contact when listening to a friend's emotions acquires the meaning of emotional closeness. In other cases, it can cause fear and anxiety: for example, a stranger staring without speaking. Moreover, the absence of an expected behavior, such as smiling when hugging a friend, is also meaningful, and changes the meaning of all the other elements of the interaction. On the level of a single interaction, the inconsistency between two or more communicative elements may produce unsatisfactory results or lead to a reframing of the meaning of the interaction. Since each element (verbal and nonverbal) acquires meaning only in relation to the others, metacommunication understanding is, therefore, a crucial element to provide consistent and satisfactory interactions. The contribution of our reflection is to underline the systemic nature of communication and nonverbal cues themselves: if analyzed alone, nonverbal cues are signifiers without a meaning. Only when we tackle the systemic nature of communication we can understand the role of nonverbal cues and produce effective communication between humans and robots.

In the next section, we discuss the psycho-relational effects of inconsistent metacommunication.

3.1 “Implicit Metacommunication Unidirectionality” and Relational Settings

After discussing the effects of metacommunication absence on the effectiveness of interactions, we examine the type of interactional setting that may be produced because of this lack. It is obvious that the formation of a stable setting implies a multiplicity of interactions between the human and the robot. Therefore, what happens when a robot is engaged in a communicative level it cannot understand, but to which it nevertheless necessarily responds?

First, we define the element of recognition implied in metacommunication. A subject seeks two types of recognition in interactions:

1) The acknowledgment of one's own metacommunicative intentions from the other person (Watzlawick 2011). So, if I have a guest for dinner and it is getting late, I will start talking about how early I have to get up tomorrow, or yawn. If my place neighbor on a plane speaks too much, I will gradually reduce the length of my answers, ending with only "yes" or "no".
2) The second type of recognition that metacommunication conveys concerns the definition of oneself in the context of the interaction. In dyadic interactions, the self-definition also implies the definition of the other person interacting. For example, a man shouts "I am your father!" to a child: this sentence implicitly defines the relation between the two (you are my son, you have to obey me). The semantic meaning can be reduced to a description of the parental relationship; the metacommunicative meaning concerns the power relations between the two. Another example: A is romantically courting B. B says "What matters most in this chapter of my life is freedom, only me with myself". B, defining herself, is also defining the relation between the two. Therefore, in this case, metacommunication is the tool through which the two (or more) subjects define the roles and the relational mechanisms of the interaction.

The level of metacommunication is where the struggle to decide who will define the relationship takes place. The one whose definition of the relation prevails will be in the "one-up" position (master), the other in the "one-down" position, undergoing the relation definition. In functional relations, the interaction participants interchange the one-up position (Selvini-Palazzoli & Boscolo 1994, Watzlawick 2011).

In HHI, when a subject "X" conveys a metacommunicative definition of the relation to "Y", there are three possible answers. When Y answers, she responds to X's definition by confirming it, denying it, or disconfirming it. For example: Y says something that offends X. Then, X looks away and stares into space with a sad expression. Note that the fact that X is offended by the words of Y is not verbalized: X uses nonverbal communication (mimicry, body posture, gazing into space) to express the disappointment with regard to Y's behavior. Here the metacommunicative struggle is about the definition of Y's behavior as

unacceptable and offending. Y may use one of three metacommunicative patterns to respond, which we summarize here[3]:

1) “Sorry, I misbehaved” (Y accepts that her behavior was wrong/unacceptable in the context of this relation).

2) “I don’t care, if you don’t like the way I behave you can stay far from me.” (Y vindicates her behavior).

3) “You are working too hard; you look very stressed” (Y deviates X’s request for recognition, warping the metacommunicative channel).

Human relations rely on the presumption that implicit metacommunicative content is understood correctly by the recipient, who is supposed to respond consistently. Moreover, because of the systemic nature of interactions, every interactional element retroactively shapes the meaning of past interactions.

The robot’s lack of understanding of human metacommunicative definitions produces the inability to manage the user’s relational expectations, at least for currently implemented robots. We will define this situation as “implicit metacommunication unidirectionality”. Unidirectionality is implicit because, on the pragmatic level of interaction, the assumption of metacommunication bidirectionality is anyhow acting, and the robot response has, in any case, an effect on the metacommunication level. This contradiction between the structure of communication and the impossibility of a robotic metacommunication will end up in a relational inconsistency. It may produce one of the following results:

1) In the case of a total and continuous confirmation, a fixed complementary relation may be produced (Watzlawick et al. 2011), with the user in a sort of relational omnipotence. Complementary settings imply that only one subject is supposed to define the relation (one-up), while the other passively accepts (one-down).

[3] This paper is not a theoretical discussion of applications of systemic interactionism. Better examples can be found in Watzlawick (2011) and Selvini-Palazzoli (1994).

“This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this article is on Minds&Machines https://doi.org/10.1007/s11023-021-09584-5

2) In the case of a continuous denial (that is, refusing the metacommunicative level), the relationship will have a symmetric escalation[4] (Selvini-Palazzoli & Boscolo 1994) that will probably result in termination of the interaction. The case of denial should not be understood as an explicit and verbal one, where the robot linguistically states its refusal of a certain type of interaction. To clarify what type of metacommunicative denial we are discussing, consider the following example: a human approaches a robot and extends her arm to shake hands. The robot does not understand the scope of this gesture and remains still. This interactional event counts as a case of denial. The refusal of shaking somebody's hand, in human interactions, may be translated as "we are not friends at all, to such an extent that I do not even respect formalities".
3) In the case of multiple disconfirmations, typically a psychotic setting will be produced, as shown by the work of Palazzoli and other systemic psychologists (Ingamells 1993). In this third case, it is hard to forecast what type of interactional pattern may be produced. It is peculiar of familiar interactions or, otherwise, relations where there is a deep emotional and existential bond, like in love relationships (Selvini-Palazzoli et al. 1988). Disconfirmation is produced when there is significant inconsistency between X metacommunicative claim and Y response, as explained above. To quote Watzlawick (2011, p. 86): "Disconfirmation, as we find it in pathological communication, is no longer concerned with the truth or falsity – if there be such criteria – of P's definition of himself, but rather negates the reality of P as the source of such a definition. In other words, while rejection amounts to the message 'You are wrong', disconfirmation says in effect 'You do not exist'."

We believe that only in the case of frequent and extended HRI may disconfirming situations be produced. In sporadic HRI, this inconsistency on the metacommunicative level may simply create a strong sense of incongruence in

[4] A situation where none of the participants to the interaction accepts definitions given by others.

the robot's behavior, perhaps producing an uncanny sensation as suggested in Kätsyri et al. (2015).

It is our conclusion that metacommunicative content is a crucial part of an interaction, and the robot's impossibility to understand and manage this level has relevant consequences. The implication we have highlighted so far is the possibility that an HRI becomes stressful and not satisfactory for the user.

4. Psycho-Relational Implications of Inconsistent Metacommunication in HRI

The robot's metacommunicative inconsistency does not produce issues only at the level of human-robot interactions. The third dimension that we promised to analyze is still missing, namely the effects that multiple human-robot interactions produce in the user's relational setting. If, as previously assumed, we can consider the influences of HHI and HRI as mutual at least to some extent, we must consider that human-robot interactions may modify the user's psychological-relational structure. For the scope of this paper, we will focus only on the case of a constant confirmation by the robot. Refusal will supposedly simply end the interaction (as stated above), and disconfirmation implies a deep existential bond that is hard to imagine for current HRI.

In this latter part of our discussion, we apply the relational theory, or *recognition theory*, developed by Benjamin to HRI implications for the user's psycho-relational setting. This approach reformulates the previous consideration from the point of view of individual psychology, going beyond the interactional analysis. While the interactionist theory only provided a framework for relational issues, Benjamin's thought also offers a framework to understand the risks produced by HRI for the subject's psychological setting and, therefore, the possible transfer of dysfunctional patterns to HHI.

In the *recognition theory,* mutual intersubjective recognition is at the core of the subject's psychic structure: a balanced structure of the self is possible only in

"This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this article is on Minds&Machines https://doi.org/10.1007/s11023-021-09584-5

the context of mutual recognition (Benjamin 2017). For example, in the dyadic relationship between mother and child, there is a tension between the will to assert one's own subjectivity (the principle of omnipotence) and the subject's need of performing significant actions, only possible if another subject recognizes their significance (Benjamin 2017). As in Watzlawick, meaningfulness is intersubjective and retroactive; actions acquire meaning only if somebody else recognizes it.

The subject psychic balance, between these two contradictory instances of omnipotence and need for recognition, is a "tension" always stretched between the two poles. This process *must* remain dynamic in order to avoid fixation of relation in "doer and done-to", namely the omnipotent subject and the passive one (Henry 2018). The same concept in Watzlawick is defined as "complementary interaction".

This delicate balance is put in crisis if one of the two sides takes over. In the case of constant confirmation, the principle of omnipotence overpowers the recognition of the other's subjectivity. Benjamin (2013, p. 35) reports on excessive confirmation towards a child: "The parents co-opt all the child's intentions by agreement, pushing him back into an illusory oneness where he has no agency of his own."

Therefore, a constant confirmation confines the subject within a bubble of omnipotence of which the result is "emptiness, isolation" (Benjamin 2013). In fact, the omnipotent subject, not recognizing anyone else as a relational partner with subjective status, in the end cannot assess and realize their own subjectivity. In HRI, users might experience a constant confirmation on the metacommunicative level, considering themselves omnipotent in defining the relational space. On the contrary, a healthy relationship presupposes that:

"The recognition process occurs when the subject and the other [...] are conceived as always mutable mirror reflecting the interlocutor, so that this reflection is neither mimetic nor annihilating of the parts at play, but rather allows a continuous and permanently interchange of polarity, not fixed in an oppositional (doer/done-to) form." [translation is mine] (Henry 2018)

Therefore, communication asymmetry must remain dynamic to avoid fixation: a split of the relational roles in a dysfunctional and oppositional form (Benjamin 2013). The functional relationship is not the one avoiding any form of complementarity, but the one where complementary roles (doer/done-to) are dynamic and not fixed. In interactionist terms, in functional relationships, the role of relation-definer (one-up) is interchangeable between the interaction participants. Human-robot interactions seem to be characterized by a fixed complementarity because of the robot's lack of metacommunication abilities that prevents, in a relationship, this dynamic interchangeability and mutual recognition. As a result, the user could be locked up inside a hallucinatory world and, over time, change their relational expectations in HHI (Bisconti & Nardi 2018).

How to set up, both theoretically and practically, a relational model that avoids complementary fixations in HRI must be investigated. HRI should not set up dysfunctional relationships. The asymmetricity of the human-robot relationship, namely the robot's inability to produce coherence between communication and metacommunication, must be assumed and tackled in the design phase in order to remedy the interactional imbalance that we have made clear in this paper.

5. Conclusions

We have suggested giving serious consideration to human-robot interactional elements that do not resemble a humanlike interaction. We have shown how non-anthropomorphic interactional elements qualitatively modify HRI. We should consider interactions as systems, where each element (including the absence of one) influences the entire interactive ecosystem. We have underlined the importance of understanding the effects, at a systemic-relational level, of the absence of a specific element: how does the relationship change? We have analyzed an example of a non-replicable interactional element: the metacommunication understanding of verbal and nonverbal cues. We have

"This preprint has not undergone any post-submission improvements or corrections. The Version of Record of this article is on Minds&Machines https://doi.org/10.1007/s11023-021-09584-5

shown how this absence qualitatively modifies HRI. The user's metacommunicative content receives a response from the robot, albeit involuntary and incoherent, and produces a relational modification. This may cause pathological interactional settings, or it can quickly end the interaction. Specifically, continuous confirmation by the robot can enact a highly complementary relational setting. Within the theoretical framework of *recognition theory*, we have argued that complementary HRI interactions may impact the user's psycho-relational setting: the human-robot relations may produce a fixation of the doer/done-to roles, resulting in a decrease of relational abilities.

In the introduction, we presented three problems that we can now partially answer:

1) Are there interactional characteristics typical of HHI that *cannot* be replicated in HRI?

One example we gave of an HHI interactional characteristic that cannot be replicated in HRI is metacommunication understanding: namely a coherent and consistent use of verbal and nonverbal cues, also in light of the relational bargaining of roles.

2) What substitutive (and original) interactional dynamics are established in HRI?

A peculiar dynamic that may be established in HRI is a fixed complementary interaction.

3) Can these HRI-peculiar interactional dynamics be transferred and enacted in HHI?

We gave only a partial answer to this question: this HRI-peculiar interactional setting may impact the user's psycho-relational setting, producing a worsening of relational abilities and an omnipotent subjective structure. These possible effects on the psychological organization of users depend both on the design of social robots, as this paper has pointed out, and the technological literacy of users. Arguably, a significant marketing of social robots will produce the

adaption of the human social context to new interactional and pragmatic practices, reducing the negative implications thanks to a socio-technical reorganization of the social sphere. Meanwhile, especially in the first steps of social robotics mainly involving users with low social opportunities such as elders and children with autism spectrum disorder, the possibility of an unwanted psychological impact should be assessed and addressed carefully.

References

1) Baker, A. L., Phillips, E. K., Ullman, D., & Keebler, J. R. (2018). Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(4), 1-30
2) Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M., & Šabanović, S. (2020). Human-robot interaction: An introduction. Cambridge University Press
3) Benjamin, J. (2013). *The Bonds of Love: Psychoanalysis, Feminism, and the Problem of Domination*. Pantheon
4) Benjamin, J. (2017). *Beyond doer and done to: Recognition theory, intersubjectivity and the third*. Routledge. https://doi.org/10.4324/9781315437699
5) Bisconti Lucidi, P., & Nardi, D. (2018, December). Companion robots: the hallucinatory danger of human-robot interactions. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 17-22). https://doi.org/10.1145/3278721.3278741
6) Briggs, G., & Scheutz, M. (2016, September). The pragmatic social robot: Toward socially-sensitive utterance generation in human-robot interactions. In 2016 *AAAI Fall Symposium Series*
7) Brinck, I., & Balkenius, C. (2018). Mutual recognition in human-robot interaction: A deflationary account. *Philosophy & Technology*, 1-18. https://doi.org/10.1007/s13347-018-0339-x

8) Dautenhahn, K., Walters, M., Woods, S., Koay, K. L., Nehaniv, C. L., Sisbot, A., Alami, R., & Siméon, T. (2006). How may I serve you? A Robot Companion Approaching a Seated Person in a Helping Context. *Proceeding of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction - HRI '06*, *April 2005*, 172. https://doi.org/10.1145/1121241.1121272

9) Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and autonomous systems*, 42(3-4), 177-190. https://doi.org/10.1016/S0921-8890(02)00374-3

10) Erden, M.S. Emotional Postures for the Humanoid-Robot Nao. *Int J of Soc Robotics* 5, 441–456 (2013). https://doi.org/10.1007/s12369-013-0200-4

11) Hegel, G. W. F., & Inwood, M. (2018). *Hegel: The Phenomenology of Spirit*. Oxford University Press

12) Henry, B. (2018) Voluntary submission as a dark side of adaptive preference. The contribution of relational psychoanalysis to Political Philosophy. *Soft Power 09* , 99.

13) Honneth, A. (1996). The struggle for recognition: The moral grammar of social conflicts. Mit Press

14) Ingamells, D. (1993). Systemic Approaches to Psychosis; Part II—Systemic Psychotherapy. *Australian and New Zealand Journal of Family Therapy*, 14(2), 85-96. https://doi.org/10.1002/j.1467-8438.1993.tb00946.x

15) Ivaldi, S., Lefort, S., Peters, J. et al. (2017). Towards Engagement Models that Consider Individual Factors in HRI: On the Relation of Extroversion and Negative Attitude Towards Robots to Gaze and Speech During a Human–Robot Assembly Task. *Int J of Soc Robotics* 9, 63–86. https://doi.org/10.1007/s12369-016-0357-8

16) Jung, B., Kopp, S. (2003). FlurMax: An Interactive Virtual Agent for Entertaining Visitors in a Hallway. In: Rist, T., Aylett, R.S., Ballin, D., Rickel, J. (eds.) *IVA 2003. LNCS (LNAI)*, vol. 2792, pp. 23–26. Springer, Heidelberg. https://doi.org/10.1007/978-3-540-39396-2_5

17) Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. Frontiers in psychology, 6, 390

18) Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot–like agent. Social Cognition, 26(2), 169-181 https://doi.org/10.1521/soco.2008.26.2.169

19) Kojève, A. (1980). *Introduction to the Reading of Hegel*. Cornell University Press.

20) Konok, V., Korcsok, B., Miklósi, Á., & Gácsi, M. (2018). Should we love robots?–The most liked qualities of companion dogs and how they can be implemented in social robots. Computers in Human Behavior, 80, 132-142. https://doi.org/10.1016/j.chb.2017.11.002

21) Kramer, N. C., Eimler, S., von der Pütten, A., & Payr, S. (2011). Theory of companions: what can theoretical models contribute to applications and understanding of human-robot interaction?. *Applied Artificial Intelligence*, 25(6), 474-502. https://doi.org/10.1080/08839514.2011.587153

22) 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. In *Human-computer interaction: The agency perspectiv*e (pp. 215-240). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-25691-2_9

23) Luria, M., Reig, S., Tan, X. Z., Steinfeld, A., Forlizzi, J., & Zimmerman, J. (2019, June). Re-Embodiment and Co-Embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (pp. 633-644). ACM. https://doi.org/10.1145/3322276.3322340

24) Mumm, J., & Mutlu, B. (2011). Human-robot proxemics: Physical and psychological distancing in human-robot interaction. *HRI 2011 - Proceedings of the 6th ACM/IEEE International Conference on Human-*

Robot Interaction, 331–338. https://doi.org/10.1145/1957656.1957786

25) Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., & Hagita, N. (2009, March). Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. In Proceedings of the 4th ACM/IEEE international conference on Human robot interaction (pp. 69-76)

26) Nomura, T., Kanda, T., Suzuki, T., & Kato, K. (2004). Psychology in human-robot communication: An attempt through investigation of negative attitudes and anxiety toward robots. In RO-MAN 2004. *13th IEEE International Workshop on Robot and Human Interactive Communication*, 35-40. https://doi.org/10.1109/ROMAN.2004.1374726

27) Nomura, T., Kanda, T., Suzuki, T., & Yamada, S. (2019). Do people with social anxiety feel anxious about interacting with a robot?. *Ai & Society*, 1-10.

28) Plurkowski, L., Chu, M., & Vinkhuyzen, E. (2011, August). The Implications of Interactional" Repair" for Human-Robot Interaction Design. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (Vol. 3, pp. 61-65). IEEE

29) Rosenthal-Von Der Pütten, A. M., Schulte, F. P., Eimler, S. C., Sobieraj, S., Hoffmann, L., Maderwald, S., ... & 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

30) Ruesch, J., & Bateson, G., (1951). *Communication: The Social Matrix of Psychiatry*. New York: W. W. Norton & Company, Inc.

31) Satake S, Kanda T, Glas DF, et al (2009) How to approach humans? In: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction - HRI '09. ACM Press, New York, New York, USA, p 109

32) Saunderson, S., Nejat, G. (2019) How Robots Influence Humans: A Survey of Nonverbal Communication in Social Human–Robot

Interaction. *Int J of Soc Robotics* 11, 575–608. https://doi.org/10.1007/s12369-019-00523-0

33) Scassellati, B. (1998). Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In *International Workshop on Computation for Metaphors, Analogy*, *and Agents* (pp. 176-195). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-48834-0_11

34) Schneier, F. R., Rodebaugh, T. L., Blanco, C., Lewin, H., & Liebowitz, M. R. (2011). Fear and avoidance of eye contact in social anxiety disorder. *Comprehensive Psychiatry*, 52(1), 81-87. https://doi.org/10.1016/j.comppsych.2010.04.006

35) Sebo, S. S., Krishnamurthi, P., & Scassellati, B. (2019, March). "I Don't Believe You": Investigating the Effects of Robot Trust Violation and Repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 57-65). IEEE.

36) Seibt, J. (2014). Varieties of the'As If': Five Ways to Simulate an Action. *Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014 (*Vol. 273). Ios Press. (pp. 97-104)

37) Seibt, J. (2016). " Integrative Social Robotics"-A New Method Paradigm to Solve the Description Problem And the Regulation Problem?. In *Robophilosophy/TRANSOR* (pp. 104-115)

38) Seibt, J. (2017). Towards an ontology of simulated social interaction: varieties of the "As If" for robots and humans. *In Sociality and normativity for robots* (pp. 11-39). Springer, Cham. https://doi.org/10.1007/978-3-319-53133-5_2

39) Seibt, J., Vestergaard, C., & Damholdt, M. F. (2021). Sociomorphing, Not Anthropomorphizing: Towards a Typology of Experienced Sociality. *Culturally Sustainable Social Robotics: Proceedings of Robophilosophy 2020*, *335*, 51.

40) Selvini-Palazzoli, M. S., & Boscolo, L. (1994). *Paradox and counterparadox: A new model in the therapy of the family in schizophrenic transaction*. Jason Aronson, Incorporated

41) Selvini Palazzoli, M., Cirillo, S., Selvini, M., & Sorrentino, A. M. (1988). *I giochi psicotici nella famiglia*. Cortina, Milano

42) Serholt, S. (2018). Breakdowns in children's interactions with a robotic tutor: A longitudinal study. *Computers in Human Behavior*, *81*, 250-264 https://doi.org/10.1016/j.chb.2017.12.030

43) Shibata, T., Tashima, T., & Tanie, K. (1999, May). Emergence of emotional behavior through physical interaction between human and robot. *In Proceedings 1999 IEEE International Conference on Robotics and Automation* (Cat. No. 99CH36288C) (Vol. 4, pp. 2868-2873). IEEE

44) Syrdal, D. S., Dautenhahn, K., Woods, S., Walters, M. L., & Kheng Lee Koay. (2006). “Doing the right thing wrong” - Personality and tolerance to uncomfortable robot approaches. *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 183–188. https://doi.org/10.1109/ROMAN.2006.314415

45) Syrdal, D.S., Koay, K.L., Walters, M.L., Dautenhahn, K. (2007). A personalised robot companion?The role of individual differences on spatial preferences in HRI scenarios. In *Proceedings of the 16th IEEE International Workshop on Robot and Human Interactive Communication.* https://doi.org/10.1109/ROMAN.2007.4415252

46) Tinwell, A., & Sloan, R. J. (2014). Children’s perception of uncanny human-like virtual characters. *Computers in Human Behavior*, 36, 286-296. https://doi.org/10.1016/j.chb.2014.03.073

47) Turkle, S., Breazeal, C., Dasté, O., & Scassellati, B. (2006b). Encounters with kismet and cog: Children respond to relational artifacts. *Digital media: Transformations in human communication, 120.*

48) Turkle, S., Taggart, W., Kidd, C. D., & Dasté, O. (2006a). Relational artifacts with children and elders: the complexities of cybercompanionship. *Connection Science*, 18(4), 347-361. https://doi.org/10.1080/09540090600868912

49) Walters, M. L., Dautenhahn, K., Kheng Lee Koay, Kaouri, C., Boekhorst, R., Nehaniv, C., Werry, I., & Lee, D. (2005). Close encounters: spatial distances between people and a robot of mechanistic appearance. *5th*

IEEE-RAS International Conference on Humanoid Robots, 2005., *2005*, 450–455. https://doi.org/10.1109/ICHR.2005.1573608

50) Watzlawick, P., Bavelas, J. B., & Jackson, D. D. (1967*). Pragmatics of human communication: A study of interactional patterns, pathologies and paradoxes*. WW Norton & Company, New York

51) Yuan, L., & Dennis, A. R. (2019). Acting Like Humans? Anthropomorphism and Consumer's Willingness to Pay in Electronic Commerce. *Journal of Management Information Systems*, 36(2), 450-477. https://doi.org/10.1080/07421222.2019.1598691