The Credibility Problem in Human Robot Interaction

Víctor Fernández Castro, Elisabeth Pacherie

THE CREDIBILITY PROBLEM IN HRI

Víctor Fernández Castro[+*] & Elisabeth Pacherie[*]

[+]LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France

[*]Institut Jean Nicod, CNRS UMR 8129, Département d'Etudes Cognitives, École Normale Supérieure & PSL Research University, Paris, France.

1.Introduction

An increasing body of work in the philosophy of mind and action has emphasized the importance of commitments for joint actions (Cohen & Levesque 1991; Gilbert 1997; Michael & Pacherie 2015; Roth, 2004). For instance, Michael and Pacherie (2015) argue that commitments facilitate joint actions by stabilizing expectations, reducing the uncertainty of the interaction, providing reasons to cooperate or improving action coordination. However, commitments can only serve these functions if they are credible in the first place. In other words, commitments can only play a function in joint action as far as the participants, more often than not, comply with their commitments. Arguably, such motivation for complying with commitments is connected with *the need to belong* (Fernandez Castro & Pacherie, manuscript), the human need to affiliate with others and form long-lasting bonds with them. Such a need is what primarily motivates us to interact and engage with those around us and act so as to preserve and reinforce the bonds we have forged with them. Other motivational forces may be at work (e.g. care for reputation, social emotions, but, arguably, the need to belong is the most basic proximate motivation for conforming to commitments and serves as a scaffold for other social motivations.

The need to belong, however, is absent during human-robot interaction (HRI). Empirical evidence in psychology suggests that humans do not recognize robots as social peers, or at least, humans do not exhibit the same tendencies and prosocial motivations to engage with robots that they exhibit when engaging with other humans (Sahaï et al 2017, 2019). Thus, we have reasons to believe that the need to belong (NTB for short) is not in place during HRI. The lack of NTB motivation during HRI may reintroduce a human-robot credibility problem, where the human motivation to comply with their commitments, and thus, their credibility is absent. Let us call this problem the human-robot credibility problem.

The aim of this paper is twofold. First, we introduce the human-robot credibility problem and show how it can undermine the interaction between human and robots. In particular, we argue that the problem is especially challenging when considering how commitments are maintained during joint action. Second, we review some recent literature in psychology and philosophy of mind in order to draw different strategies that can be used in social robotics for overcoming the problem and compensate for the absence of the need to belong in HRI.

2.Commitments in Human-Human Interaction

In order to be characterized as an instance of joint action, the action must be the result of a joint intention where joint intention can be characterized as a persistent goal that the individual members of the group aim to achieve in a condition of mutual knowledge (Cohen and Levesque 1991; Gilbert 1997). Such a condition of mutual knowledge establishes that the participants must know that the participants individually intend the goal to be achieved and that they will behave as necessary to perform the goal until it is

achieved. In this sense, in order to engage in a joint action, the participants must be committed to achieving the persistent goal (G) as a group or informing their partners that the persistent goal is not achievable anymore (Participatory commitments). Furthermore, the participants must be *contralaterally committed* to performing the necessary actions and sub-goals to achieve the overall goal (Roth 2004). For instance, acting as expected given the appropriate circumstances (e.g. not performing an action that is at odds with the goal) or helping the other participants when they have problems to perform a particular sub-goal.

Establishing participatory commitments requires the participants involved to generate reliable expectations that they intend to do G in order to make mutually manifest their readiness to achieve G. In this sense, a straightforward way to establish a commitment is to make a promise (Austin, 1975). However, participatory commitments require not only that participants be committed but also that they know that their partner is committed too. In order to do that, participants often establish the participatory commitments through what Clark (2006: 131-133) calls a projective pair (e.g. proposal/acceptance), where one of the participants proposes a particular goal (Let's do G!; Should we do that?) and the other can accept or reject it (Ok). Furthermore, there are other mechanisms that can be used to generate reliable expectations that one is committed to G. First, if a participant exhibits an intention to perform G, the other participant must produce gestures or non-verbal signals to indicate that she is also committed to G (Siposova 2019). Second, social interactions are often mediated by social norms, rules, and scripts that establish a participatory commitment. For instance, if I go to a restaurant, the waiters, cooks and I assume the persistent joint goal of serving me dinner. In a nutshell, humans use different

verbal and non-verbal devices to establish participatory commitments to engage in a joint action.

However, during the joint action, there is always a risk that one of the participants revises her intention and motivations to achieve G and abandon the joint action. Such motivational uncertainty would require one to constantly and actively monitor others' intentions and behavior in order to control the risk that the other interactant may partially or fully *disengage* from a task that s/he is performing together with a partner. In this sense, a central problem with the maintenance of participatory commitments is the risk of instrumental and common ground uncertainty related to contralateral commitments. In several circumstances, even when a participatory commitment is established, the participants can have different instrumental beliefs about what their contralateral commitments regarding the action are. This uncertainty does not only undermine the coordination between the two agents, for instance, hindering dyadic and triadic adjustments; but also, the perception that one of the participants is violating a contralateral commitment can be perceived by the co-actor as a signal of a lack of implication in the task or even as a signal that the co-actor refuses to comply with the participatory commitment.

In this sense, maintaining participatory commitments partially relies on the monitoring and maintenance of contralateral commitments. Participatory commitments require the completion of different layers of sub-participatory and contralateral commitments, which create a hierarchy of commitments (Clark, 2006: 137-138). Such a hierarchy creates an interlocking set of commitments that accumulates different layers of obligations that make the participants more motivated to remain engaged in the action, ,thus reducing the

risk of leaving. Now the question is how do we deal with such contralateral commitments?

Often, the contralateral commitments regarding sub-goals and tasks are clear. For instance, when there is an asymmetric relation between the participants (e.g. boss/employee) that automatically assigns different roles and specifications to the actors. Furthermore, individuals exploit different norms of practical rationality that dictates what is the most appropriate course of action to commit to given the situation and the general participatory commitments. However, even in these situations, especially when the two co-actors do not have a history of mutual interactions, each member must display different strategies to decide which to contralateral commitments they must comply; but most importantly, each member must insure that the other party involved knows what he will do in every particular situation in order to coordinate with him.

Participatory and contralateral commitments can be broken for different reasons. First, the author of the commitment can fail to perform effectively the relevant action or fail to do so at the appropriate time. Second, the recipient of the commitment can over-interpret or under-interpret the action. For instance, one may perceive a particular action as an instance of the execution of a contralateral commitment when it is not or vice versa. Finally, the author may just find different reasons or motivations to renounce to comply with the commitment and abandon the task.

To avoid such failures, human participants must deploy monitoring and repair strategies that secure the compliance with contralateral commitments. In order to facilitate monitoring, humans often employ different ways to signal or anticipate the appropriate

course of action. First, apart from obvious verbal communication, one may provide different gestures and non-verbal signals (ostensive gaze direction) to give reliable clues to the other co-actor that one is going to initiate the relevant course of action (Siposova, 2019). Second, one may use different *coordination smoothers*, modifications of one's own behavior to make it easier for others to predict one's upcoming actions. For an agent can exaggerate her movements or reduce the variability of her actions (Vesper et al. 2007).

Facilitating monitoring is not enough to keep our contralateral commitment alive. As we mentioned, a participant may consider that the other participant's course of action is not relevant or performed at the appropriate time or maybe she could just perceive that the other participant is not acting it as she should. In these occasions, the participants in the joint actions must trigger different repair strategies to insure that the co-actor comply with her contralateral commitments. As in the case of facilitating monitoring, there are different strategies that one may use to repair contralateral commitments. Such strategies may differ depending on how the violation of the commitment is committed. First, when the commitment is perceived as a relatively small deviation from what is expected, the recipient of the commitment may just attempt to display implicit repair strategies, for instance, to *compensate* for the deviation himself (e.g. displaying a helping behavior) or to automatically *express negative emotions* about the outcomes of other participants' actions in order to motivate a change of the course of action (Michael 2011). However, when the perceived deviation is larger, the repair strategies could become more explicit. For instance, *protesting*, *reprimanding* or *asking for explanations* (Roth 2005).

Similarly, the reactions of the author of the violations to the repair strategies could differ depending on the type of strategies. While the reaction to implicit repair could be to

(automatically or intentionally) compensate her behavior, more explicit repair strategies may provoke explicit reaction such as apologizing and intentional compensation. However, in some cases, we can expect the author of the commitment to openly *explain* or *discuss* what he is doing, which can lead to a re-negotiation to the commitment in place that could end up with an acceptance, rejection or counter-proposal by the recipient (Clark 2006).

At this point, it is worth mentioning that monitoring and repair presuppose, contrary to other psychological devices (see Székely & Michael, 2018), a certain normative force. For instance, when two participants establish a commitment they are automatically entitled to reprimand or sanction the other when the other abandon the task (see Gilbert, 1997), but also, entitled to monitor what the other is doing in the context of the joint task. However, monitoring and repair have important cognitive and behavioral costs. For instance, reprimanding for not complying with his contralateral commitments could have negative consequences for the socio-affiliative relation between the participants. So why do we engage in such strategies in spite of their potential negative consequences for the interaction itself?

The answer must be found in the pro-social tendency to engage in social interactions embodied in the need to belong. Human beings exhibit a need to affiliate with others and form long-lasting bonds with them. Such a need is what primarily motivates us to interact and engage with those around us and act so as to preserve and reinforce the bonds we have forged with them. But also, it is the more basic proximate motivation for conforming to commitments, and thus, for deploying the necessary strategies for maintaining them

alive. Our disposition to perceive social interactions as intrinsically rewarding may compensate for the possible costs of monitoring and repair.

3. Commitments in Human-Robot Interaction

In the previous sections, we have presented different strategies that humans exhibit for establishing and maintaining participatory and contralateral commitments in joint action. In this section, we would like to envisage a possible problem that the maintenance of commitments may generate in human-robot interactions. This problem, we argue, poses a general challenge to roboticists whose goal is to build robots capable of interacting cooperatively with humans; that is, the challenge of dealing with different sources of opacity and with the resistance of the human participant to display reparative behaviors.

In the previous section, we argued that a successful joint action requires co-agents to maintain different contralateral commitments that facilitate the maintenance of the participatory commitment to a persistent overall joint goal. This involves not only making the intentions to perform the goal mutually manifest but also monitoring and repairing such contralateral commitments. These strategies presuppose a pro-social motivation that offsets the possible costs associated with the strategies. For instance, humans must find social interactions rewarding and pleasant in order to be ready to assume the costs of being entitled to sanctions or being ready to provide social cues that facilitate the monitoring of their own actions.

However, a number of findings in psychology and neurosciences suggest that humans interact differently when their partner is a robot rather than a human. (Sahaï et al. 2017, 2019; Wiese et al. 2017). To give an example, while different studies in neurosciences indicate that humans can recruit different motor simulation mechanisms to understand

others' behavior even during passive observation of others (Elsner et al., 2012 Manera et al., 2011), studies with PET suggest that humans predictive neurological devices are not responsive to non-human generated actions (Perani et al. 2001; Tai et al. 2004).

Although the reasons for these differences could be diverse, we believe that the fact that humans do not recognize robots as social peers would automatically inhibit prosocial dispositions associated to the need to belong, and thus, the central motivations to deploy different socio-cognitive capacities we often exercise in social encounters with human partners.

The lack of the need to belong and the prosocial tendencies associated with it during HRI poses two sources of difficulties and challenges for social roboticists. First, from the robot perspective, the lack of need to belong motivation or pro-social behavior on the part of the human can be an important source of opacity. Human's lack of affiliative motivation may produce resistance to be cooperative but also to provide the type of social cues that facilitate monitoring during the interaction. Furthermore, the lack of prosocial motivation can make the human participant more intransigent with the robot's failures or inappropriate behaviors which would increase the risk of abandoning the joint task. Second, from the human perspective, given the robot's *underperformance* and *overperformance* of the task, there is another source of opacity that it is difficult to compensate with reparatory strategies if the human lacks the prosocial motivation to trigger them.

These sources of difficulties are amplified by other specific features of HRI. To see how, notice that an important number of HRI presuppose an asymmetric relation between the two participants where the robot is a helper or servant and the human is the boss or figure

of authority. This asymmetry could reinforce some of the factors presented above, for instance, the resistance on the part of the human to provide social cues to help the robot monitors her commitments. Even in cases where the robot is the figure of authority (e.g. therapeutic environments), human resistance to perceive and recognize the robot as an agent with whom to create a long-standing relation may provoke fatal failures in the interaction.

In a nutshell, HRI interactions exhibit certain features that are absent in HHI. The fact that humans may not recognize the robot as a peer may inhibit their affiliative tendencies which amplify the asymmetry of the interaction and may reduce the tendency of the human to comply with their commitments. This imposes different problems for social roboticists. First, the robot will receive fewer social cues to facilitate its monitoring. Second, the opacity of its behavior and the lack of affiliative tendencies in the human may result in the human making fewer reparative behaviors to facilitate comprehension by the robot.

4. How to Solve the Human-Robot Credibility Problem: Some preliminary strategies

Addressing the challenges presented above requires focusing on two different tasks: improving monitoring for both the human and the robot co-actors and providing the robot with accurate devices for displaying reparatory behaviors when the human or itself fails to comply with a commitment.

Improving transparency and monitoring during HRI requires making robots able to reliably produce and understand social signals. The objective of creating robots which

produce social signals can be instantiated in different ways. For instance, Glas et al (2016) and Nishio et al. (2007) attempt to create robots that look and act like humans, while other researchers (Zecca et al. 2004) have concentrated on imitating the biomechanics of human movements rather than making their appearance human like. However, mechanisms for social signaling do not requires imitating human sociality or providing sophisticated expressions. For instance, (Triebel et al. 2016) equipped Spencer, a socially aware service robot for passenger guidance, with the capacity for anticipating his next movement by looking at the direction he is going to take before he does. Providing robots with more sophisticated mechanisms of detection is the other key aspect for improving transparency. In this respect, there have been important advances in the development of more robust actions or emotions recognition in robots in the last years (Hoffman & Breazeal, 2007; Karg et al. 2013; Koppula & Saxena, 2013; Palinko et al. 2014). In a nutshell, the current state of the art in robotics allows us to be optimistic regarding the implementation of this type of strategies for transparency and elicitation of social behavior.

However, social robotics, but also research in human-human interactions, often overlook the importance of repair during joint action. As we understand it, the notion of repair is intimately related to the *normative aspect of commitments*. The way we respond to others violation of their commitments often implies a normative attitude which implies holding the other on demand of what he must do. To understand this aspect, consider the distinction between normative and descriptive expectations (Greenspan 1978; Paprzycka 1999; Wallace 1998). Descriptive expectations are tied to the notion of prediction and their violation or frustration does not necessarily triggers reactive attitudes. For example, you can expect your friend to have a beer because this is what she always does but if she does not, this may surprise you but not bother you. However, normative expectations are

connected to the idea of holding someone on demand and their frustration triggers reactive attitudes like blame, request for justification or sanctions. In other word, the normal response to a violation of a normative expectation is to impose a negative reaction to the other agent for not acting as expected. Such negative reactions are more emotionally loaded and directed to regulate the others behavior. For instance, you can feel entitled to sanction your friend when he frustrates your expectation that he will cede his seat to an older person on the subway. Such a normative attitude is also intrinsic to joint action and commitments (Gilbert 2009; Roth 2004). In fact, some recent studies suggest that people who judge that two persons are walking together in certain conditions were more likely to judge that one of the participant has the right to rebuke the other when he peels off (Gomez-Lavin and Rachar 2018). In other words, sanctions and repairs are pivotal responses to violation of commitments, and thus, social robotics must consider them in its general strategy to avoid the problems associated to the develop of HRI.

In section 2, we discussed several strategies aimed at insuring that the co-actor complies with her commitments. Those strategies include more implicit strategies like compensation or emotional responses; and more explicit strategies like reprimanding or protesting. Now, the question is whether we could redeploy the repair strategies that human display during interaction in HRI. Given the difficulties reviewed above regarding the lack of need to belong during HRI, focusing on repair strategies must be an important ally for improving the interaction between humans and robots. Expectably, the specific sources of prosocial motivation and opacity that emerge during HRI could be compensated by designing robots able to display repair strategies to avoid the break-up of commitments.

However, the question of how we can equip robots with repair devices leads to two basic problems. First, as in the case of monitoring, when the robot is the author of the commitment but fails to perform the expected action, it would have to be made sensitive to the emotional reactions of its human partner to engage in reparative behavior or to detect when the human has engaged into compensatory strategies. This would require building robots with more robust mechanisms for understanding human emotions and behavior. Second, when the robot is the recipient of the commitment, we must be careful of the reparative strategies he recruits. After all, given that humans are often the authoritative figure during the interaction, certain repair actions could be perceived by the human co-actor as damaging or threatening.