Robots and Resentment: Commitments, recognition and social motivation in HRI

Elisabeth Pacherie, Víctor Fernández Castro

Draft version. Fernández Castro, V. & Pacherie, E. (forthcoming) "Robots and Resentment: Commitments, recognition and social motivation in HRI" In C. Misselhorn, T. Poljansek and T. Störzinger (eds.) *Emotional Machines. Perspectives from Affective Computing and Emotional Human-Machine Interaction.* Springer.

Robots and Resentment: Commitments, recognition and social motivation in HRI

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

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

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

1. Introduction

A fundamental challenge for robotics is to develop agents capable of interacting with humans in collaborative tasks. In recent years, considerable resources and effort have been directed at designing and manufacturing robots for use in numerous social contexts, such as companionship to the elderly, education, therapy or service. To make further progress, social robotics needs to design robots capable of meaningfully engaging with humans, to ensure the robots can collaborate with humans in shared activities and joint actions with high levels of coordination. This need explains the fast expansion of the field of human-robot interaction (HRI) and the various avenues of research explored within this field in an effort to enable robots to engage more successfully in social interactions. As part of this expansion, HRI research has taken inspiration from some important findings in psychology, philosophy of mind and neuroscience regarding

human-human interaction (HHI) to provide robotic agents with the necessary cognitive capabilities for achieving joint actions.

In the wake of these studies, our chapter proposes that some of the current problems confronting robot-human interaction in joint tasks can be solved by equipping robots with capabilities for establishing mutual recognition of social agency. After first arguing in favor of the fundamental role of the notion of commitment in mutual recognition, we show how the attribution and maintenance of commitment requires fundamental affective states such as social emotions or the prosocial motivation of the need to belong. Finally, we survey three proposals on how social robotics could implement an architecture of commitment by addressing the centrality of these emotions and exposing their weaknesses and strengths.

The chapter is structured as follows. In section 2, we argue that prediction and motivation are two pivotal aspects of joint action which are especially challenging for social robotics. In section 3, we discuss a way of meeting these challenges according to which we must equip robotic agents with the capacity of establishing mutual recognition with humans. We argue that, although promising, the minimal version of the recognitional view proposed by Brinck and Balkenius (2018) has important limitations. In section 4, We propose understanding mutual recognition in terms of the capacity for attributing, undertaking and signaling commitments and argue that this version of the recognitional view is better posed to improve prediction and motivation in HRI. In section 5, we argue that social emotions or the need to belong are indispensable for the establishment of commitments. In section 6, we consider three different proposals for enabling such affective states or for replacing them with functionally relevant substitutes and discuss some of their problems and limitations.

2. Joint Action, motivation, and prediction in HRI

Broadly considered, a joint action is a social interaction where two or more individuals coordinate their behavior in order to bring about a common goal or to have a particular effect in their environment (Sebanz et al., 2006). Joint action has been the subject of debate between philosophers and psychologists for some time now. On the one hand, philosophers have traditionally claimed that joint action requires the participants to share certain intentions and they have extensively discussed the nature of such shared intentions. For instance, Margaret Gilbert (1992; 2009) and Michael Bratman (2014) have extensively argued about whether shared

intentions depend on the establishment of joint commitments or about the continuity between individual and shared intentions. On the other hand, psychologists have focused on elucidating the psychological mechanisms facilitating the coordination of behavior among these participants, including devices like motor predictions (Prinz 1997) entrainment (Harrison & Richardson, 2009), perception of joint affordances (Ramenzoni et al., 2008) or mimicry (Chartrand & Bargh, 1999).

Both philosophers and psychologists share the objective of elucidating why and how humans spend an important amount of their time engaged in collaborative action and of characterizing the sophisticated abilities that support collaborative action. Collaborative action confronts two main challenges. First, successful cooperative interactions are premised on action *prediction*. Agents need to coordinate their actions at various levels and must be able to make accurate predictions regarding their partner's actions and their consequences. Shared intentions serve to plan joint action and decide on general and subsidiary goals or sequences of actions to be performed (Bratman, 2014; Pacherie, 2013) and a variety of psychological devices help insure their implementation by facilitating coordination and mutual adjustments among co-agents (Knoblich et al., 2011). Second, successful cooperative interactions are also premised on *motivation*. Humans exhibit a strong proclivity to engage in social interactions. Their motivation can have a variety of sources, both endogenous (e.g. need to belong or general pro-social tendencies) or exogenous (e.g. social pressure). We find some interactions with others intrinsically rewarding (Depue & Morrone-Strupinsky, 2005) and even when we do not, other sources of motivation (e.g., moral values or social pressure) can lead us to help others or collaborate with them. In fact, some of the predictive mechanisms mentioned above seem to have close ties to pro-social motivation. For instance, people who exhibit more cooperative tendencies and possess more empathic dispositions are also those who exhibit more nonconscious mimicry of postures or expressions. Chartrand and Bargh (1999) have demonstrated the motivational aspect of the *chameleon-effect*, which refers to the nonconscious mimicry of postures, expression and other behaviors when interacting with a partner.

In a nutshell, motivation and prediction are two pivotal aspects of joint action among humans and, as such, the two elements must be considered central in the design of social robots. Moreover, both create important challenges for social robotics, not just because of the difficulty of implementing mechanisms capable of instantiating such functions in robots but also because, as a number of studies have demonstrated, there are different negative elements of HRI that may undermine motivation and prediction in various ways.

First, several findings in psychology and neuroscience suggest that humans interact differently when their partner is a robot rather than a human (Sahaï et al., 2017; Wiese et al., 2017). To give an example, while different studies in neurosciences indicate that humans can recruit motor simulation mechanisms to understand others' behavior even during passive observation of others (Elsner & Hommel, 2001), studies with Brain Positron Emission Tomography suggest that the premotor mirror system activated when observing human grasping actions is not responsive to non-human generated actions (Tai et al., 2004).

Second, the gap between the expected and the actual capabilities of the robot seems to impact the predictive capacities of humans (Kwon, Jung, and Knepper 2016). While the physical appearance of the robot and some of their social capacities –e.g. mimicry—may generate high expectations regarding the autonomy and sophistication of the robot, its real capacities might actually be heavily context-dependent and its autonomy quite limited. Such a difference between expectations and reality may provoke frustration in humans but more importantly, it may negatively influence their attunement to the robot and decrease their capacity to generate reliable predictions of the robot's behavior. Thus, Kwon, Jung, and Knepper (2016) created a series of surveys where subjects were presented with vignettes that described a human collaborating with an industrial robot, a social robot or another human in either an industrial or a domestic context. In their first study, they found that people often generalized capabilities for the social robot, attributing the same confidence to both robots in industrial settings. In contrast, they only attributed low confidence to the industrial robot in domestic settings. In a second study, they created two video clips of a human-robot team and a human-human team each completing a simple block-building task, with one teammate responsible for one color of blocks. They programmed the robot to be incapable of stacking blocks without the help of their human partner and introduced the same limitation in the human-human team. Participants were asked to rate the capabilities of the more "limited" teammate at various points of the human-human and human-robot videos. The experimenters found that the subjects were "more willing to modify their expectations based on a robot's perceived capabilities compared to a human". In other words, the gap between expectations and perceived capabilities is more pronounced when the observed agent is a robot, which could increase the risk of failures during HRI.

Similarly, several studies have uncovered a number of factors that can undermine human motivations for engaging with robots. First, an often-voiced problem in social robotics is the well-

known *Uncanny Valley Effect*, the phenomenon whereby humans experience a feeling of discomfort or revulsion when perceiving a machine or artifact that acts or looks like a human (see Wang et al., 2015 for a review). The first reference to this effect appears in the work of Mori (1970) and it has been observed not only in humans but also in primates (Steckenfinger & Ghazanfar, 2009). There are different hypotheses concerning the possible causes underpinning the discomfort, for instance, that the feelings are associated with evolutionary adaptations for some aesthetic preferences or the avoidance of pathogens. However, it is important to emphasize that whatever these causes are, they can interfere with the motivation of human agents to engage in social interactions with robots.

Second, the empirical studies reporting negative attitudes toward robots are not restricted to those regarding the Uncanny Valley Effect. In a series of experiments using implicit association tests, that measure the reaction times of participants depending on the associations between a target (a robot or a human) and a positive or negative attribute, (Sanders et al., 2016) found that the participants in the experiments exhibited implicit negative attitudes toward robots even when their explicit assessments of the robotic agent were positive. Those experiments seem to demonstrate that people exhibit adverseness toward robots, or at least, less positive stances than they project toward humans. Moreover, other aspects of robots like their human-like appearance or personality can be perceived as deceptive (Vandemeulebroucke et al., 2018) or impact the human levels of the trust, which could produce resistance to start interacting with the robot or lead one to abandon too quickly the collaborative task, ending up in the disuse of the robot (M. Lewis et al., 2018)

In conclusion, despite the enormous advances made towards endowing robotic agents with socio-cognitive capacities, there are reasons to believe that attempts at establishing a better mutual understanding between humans and robots do not always meet with success. In particular, the mentioned studies have discovered some important elements that could negatively impact the objective of designing robots able to collaborate with humans as a team. In the next section, we review two general approaches to social robot design that may help solve these problems and we present some of their limitations.

3. Social Robotics and Recognition

How can we eliminate or mitigate the threats to prediction and motivation in joint action for HRI?

A promising strategy aims at identifying and characterizing the fundamental social capacities deployed in human-human interactions in order to design robots with similar capacities for understanding the behavior or mental states of their human partners. Such an approach is exemplified by the work of several laboratories that attempt to design robots with social capacities like joint attention (e.g. Huang & Thomaz, 2010), recognition of emotion or body gestures (Benamara et al. 2019) or theory of mind (Pandey & Alami, 2010; Sisbot et al., 2010). To take a few examples, Huang and Thomaz (2011) carried out a study with a Simon robot able to recognize human attention (by recognizing eye gaze, head orientation or body pose), to initiate joint attention (by using pointing gestures, eye gaze and utterances) and to ensure joint attention (monitoring the focus of attention and soliciting the partner's attention with several communicative strategies). In their study, they found that people tend to have a better mental model of the robot and to perceive the robot as more sociable and competent when it manifests a capacity for joint attention. Another example of this type of strategy is Pandey and Alami's (2010) theory of mind-like implementation which uses information about the human's position, posture or visibility of the relevant space and objects to enable the robot to reason about the human reachability and her visual perspective. Moreover, these types of designs enable robots to exploit different social cues to respond to the human's action, plan different courses of behavior or implement collaborative interactions. Some labs have designed planners and decision-making devices that take into account the information provided by these mechanisms. For instance, Sisbot et al. (2010) designed a system that integrates information from perspective-taking, the human point of view on the relevant space and objects, and human-aware manipulation planning to generate robot motions that consider human safety and their posture along with task constraints.

Endowing the robot with such highly sophisticated capacities enables it to understand humans and reason about their behavior and mind. However, as Brinck and Balkenius (2018) argue, the problem with this strategy is that such capacities are not often oriented towards making the human user feel *recognized* by the robot or towards establishing a *mutual recognition* between the human and the robotic agent (for another view of recognition in HRI see Laitinen, 2016). Humans can understand others' behavior, for instance, in terms of mental states. However, they are also capable of expecting different behaviors from others depending on their physical aspect or the shared environment. Moreover, humans are constantly exhibiting different proactive and reactive strategies to make the other aware of such expectations or to acknowledge the expectations the other has towards them. Such responsiveness and proactive and reactive strategies serve to establish a mutual recognition, that is, the participants are not merely passive subjects of

prediction and control by the other (Davidson, 1991, p. 163; Ramberg, 2000, p. 356), but active agents who attribute to each other certain rights and obligations concerning the joint action.

The notion of recognition is complex, but in the minimal sense, recognizing and being recognized as social agents give rise to the type of behaviors and responses that are at the root of our sociality and of the way we adjust to each other's actions. In this sense, Brinck and Balkenius argue, the design of the robot must consider the human as a social being with needs, desires, and mental states, and thus, the robot must be designed to be aware of, and responsive to, these features. As a result, the robot should be able to take the human into consideration and react to his presence, body, and actions, but also to signal its own presence and acceptance of the human, so the recognition becomes mutual.

Now, the question is which is the best way to implement this recognitional approach in social robotics? Brinck and Balkenius have put forward a minimal recognitional approach according to which the design of the robot must focus on embodied aspects of mutual recognition. According to several authors (Brandom, 2007; Satne, 2014), recognition involves high-level cognitive capacities including being able to give, and ask for, reasons or the ability to respond to blame and reproach. However, Brinck and Balkenius argue that recognition can be manifested in more minimal social capacities including attending to the other, searching and making eye-contact, engaging in turn-taking or mimicking postures. According to this minimal approach, recognition involves three cognitive capacities: First, *immediate identification*, the processes that assign certain properties to the other individual and generate certain expectations about how others will engage in a mutual activity based on perceptual information available here and now. This identification, they argue, requires the perception of movement and action, gaze, vocalization, and emotions. Second, *anticipatory identification* requires anticipating the actions of the other based on previous interactions and the available information in the context, for instance, the perception of the others' actions in the interaction. Finally, mutual recognition requires, what Brinck and Balkenius call *confirmation*, which involves reacting to the presence of the other or signaling pro-actively to show that identification has taken place and one is ready for the interaction. As a result, when these elements are combined, the individual shows that their behavior can be influenced by the other's, thus exhibiting a willingness to engage in the interaction. Such mutual recognition is crucial not only to establishing the readiness to interact but also the dynamic of signals, actions and reactions that facilitate the interaction. Although in principle, the three basic components of recognition can be instantiated by different capacities,

Brinck and Balkanius emphasize the importance of embodied recognition, that is, identification and confirmation based on physical constitutions of the body and sensory-motor processes. Some of the processes involved in this type of recognition are attentional engagement, mimicry of postures or gestures, emotional engagement, responding to other's gaze and attention, exaggeration of the movements or explicit modification of their kinematics, turn-taking or active eye-contact.

This general framework gives us a solution to the problems presented in section 2. Equipping robots with capacities for establishing mutual recognition may solve the motivational problem to the extent that people may feel recognized by the robot. Robotic embodied recognitional capacities could make the human feel perceived as a social peer, and thus, bring about the same type of motivations that an interaction with a human social peer may trigger. Moreover, this strategy could, in principle, dissipate some of the prediction problems. Designing robots with capacities for confirmation can help control the type of expectations that the human generates. For instance, if the robot is designed with confirmation strategies regarding some expectations (e.g. the robot confirms his capacity for facilitating physical therapy) but not others (e.g. he is unable to establish conversation), it could reduce the aforementioned expectation gap.

Despite its virtues, the embodied recognitional approach suffers from important limitations, however. First, it lacks the level of abstraction that would be needed for it to be a general approach to the design of social robotics. In robotics, we find a large number of different robotic agents, which possess different perceptual and behavioral capabilities and physical features. To give a few examples, we can find robots able to move their heads and arms (like Pepper or i-Cub) while others cannot (e.g. Rackham). Some agents can introduce different signals through channels like a screen (Pepper and Rackham), while other agents, like a humanoid, are restricted to human-like expressive capabilities or language. Similarly, while most robots are equipped with visual capacities in the form of cameras, many others use tactile sensors, thermal cameras or heart rate detectors. Such a variability creates a problem for strategies that lack a sufficient degree of abstraction, precisely because some embodied strategies may be available to some robotic agents and not others. We need to model social interactions in a way that abstracts away from some aspects of implementation, so we can adjust different social strategies to different robots depending on their perceptual and behavioral capabilities.

Second, the embodied recognition strategy fails to take advantage of some important procedures that human exploits during joint action and that are available to social robotics. This is especially obvious in the case of identification. As Brinck and Balkanius state, humans generate expectations about others' actions based on different embodied aspects and physical features. However, the sources of information we use to generate expectations about others during joint action are not restricted to physical features or even contextual factors. Humans often anticipate or predict others' actions by the mediation of patterns of rationality (Dennett, 2009; Fernandez Castro, 2020; Zawidzki, 2013), scripts and social norms (Maibom, 2007; McGeer, 2015) or the structure and features of the joint action itself (Török et al., 2019). To give an example, in a recent study, Török, et al. (2019) demonstrated that when people perform joint actions, they behave in ways that minimize the costs of their own and their partner's movement and they make rational decisions when acting together. Arguably, the capacity for diminishing the costs of the partner's choice must be partially based on the assumption that the partner will behave rationally. In other words, we can assume that people expect each other to behave as it is rationally demanded by the joint action. A second example has to do with the type of expectations that we generate depending on the nature of the joint action itself. As several authors have suggested (Knoblich et al., 2011; Pacherie, 2011; Vesper et al., 2017), when we engage in a joint action, we form representations of the joint plan. Such representations not only involve an individual's own actions in relation to the joint action but also in relation to their partner's actions. In this sense, individuals anticipate and predict partner's courses of action in relation to the representation of the joint plan. In other words, the human capacity for identifying each other relies on a variety of informational sources that are not restricted to embodied physical information. These types of strategies can be extremely helpful for social robotics, so there is no reason why we should not exploit similar strategies during HRI.

Thus, while we do not want to deny that the embodied recognition strategy may contribute to solve or attenuate the prediction and motivation problems in social robotics, we think that a more general strategy is necessary. Such a strategy may be elaborated upon a model that abstracts away from specific implementations of the robotics agents, so it can be adjusted to the embodied and non-embodied idiosyncrasies of every robot while exploiting the socio-cognitive capacities of humans and establishing a mutual recognition between humans and robots. In the next section, we present our proposal according to which the recognition between humans and robots must be understood in terms of commitments. So, the general strategy to design robots able to overcome the aforementioned problems with prediction and motivation must be oriented towards developing

robots with the capacity for establishing, tracking, and responding to, individual and joint commitments.

4. Commitments and Recognition

Part of the rationale behind the notion of *recognition* is the idea that the stance that humans have toward each other's behaviors is not passive, as if one were distantly observing a mere object whose movements one needs to predict. In fact, in joint actions, humans adopt an active stance, in which they pro-actively provide social cues to facilitate prediction and anticipation by their partner, regulate their behavior to make it more transparent and actively influence others' actions to facilitate the realization of a joint goal. Evidence that humans adopt such a proactive stance is provided by empirical findings suggesting that, during human-human interactions, people provide others with information about their own actions. For instance, some studies on sensorimotor communication demonstrate that people exaggerate their movements to allow their partners to better recognize the action goal (Vesper & Richardson, 2014). Moreover, humans are sensitive to implicit cues (gaze signals) that manifest an agreement to carry out with them a task their partner intends to perform (Siposova et al., 2018).

In our view, such a proactive stance and the repertoire of actions and capacities necessary to adopt it requires agents to attribute and undertake different participatory and individual commitments. As a first approximation, we submit the idea that mutual recognition in joint action is established when the partners recognize each other as authors of different commitments involved in the interaction and hold each other responsible for such commitments. People proactively give social cues, regulate and adjust their behavior in response to others as a way to establish, negotiate and track a set of individual and joint commitments related to the goal and plan associated with a joint task.

Now, a commitment is in place when the recipient generates an expectation regarding the author as a result of having an assurance that the author will act according to the expectation in a condition of mutual knowledge. To give an example, Sara is committed to helping Andrew repair his bike when Andrew expects Sara to do it as a result of Sara having made a promise to do so and Andrew having acknowledged the promise. Traditionally, philosophers have connected the establishment of commitments to explicit verbal actions, e.g., one agent, the author of the commitment, commits to another, its recipient, to a course of action X by intentionally

communicating that one intends to X through a promise or other speech act (Austin, 1962; M. Gilbert, 2009). However, commitments are not necessarily established through explicit verbal agreements. For instance, one might indicate through gestures or facial expressions that one will perform the appropriate action (Siposova et al., 2018). Moreover, as several authors claim, some factors like situational affordances, social norms and scripts (Fernández Castro & Heras-Escribano, 2020; Lo Presti, 2013), or the identification of another agent's goal (Michael et al., 2016), for example, can lead an agent to undertake commitments and attribute commitments to others.

Understanding our social stance in terms of commitments allows us to understand the notion of identification in a way that helps us cover a greater range of strategies than the embodied version of the recognitional view. *Identifying* another agent as a social peer implies attributing to her a set of commitments from which we can generate different expectations that anticipate and predict her actions. We can attribute different commitments depending on the physical appearance of the author, but also, depending on social norms, general patterns of rationality, scripts or the structure and features of the joint action itself. For example, we can attribute to our partner the commitment to behave in the most rational way and minimize the costs of the overall action. In other words, we can assume that people identify each other as committed to behave as rationally demanded by the joint action, which generates expectations that facilitate anticipation and prediction. Another example has to do with the type of commitments that we attribute to each other depending on the nature of the joint action itself. As Roth (2004) has emphasized, when we engage in joint action, we do not only undertake a joint commitment to pursue the joint goal but we also undertake a set of contralateral commitments regarding individual actions and sub-goals necessary to the success of the collective task. For instance, if we agree to go for a walk together, we can attribute to our partner an individual commitment to walk at the same pace.

Now, identification (and confirmation) are not the only types of capacities involved in the social stance we adopt during joint action. When we perceive our partner as such, we do not only generate expectations and wait for confirmation; we often use *exhibitory signals* to indicate to our partner what we expect her to do. In other words, one pro-actively gives cues to one's partner regarding what behavior one believes should be performed. For instance, as Michael et al. (2016) suggest, people often use investment of effort in a task as an implicit cue for making the perceiver aware that we expect him to behave collaboratively. In other cases, the cues are more explicit, for instance, when we negotiate what to do during the task through what Clark (2006:131–33)

calls a projective pair (e.g. proposal/acceptance), where one of the participants proposes a particular goal to another (Let's do G!; Should we do that?), who then accepts or rejects the proposal. Besides these exhibitory actions, social agents manifest different *regulative actions* directed toward the performance of others. For instance, we often use positive or negative emotional expressions, like smiling or wrinkling one's nose as a signal of approval or disapproval toward the action of the other (Michael, 2011). Moreover, humans exhibit a robust repertoire of regulative actions directed toward others when they have frustrated our expected social interactions, including blaming, reprimanding or asking for reasons (McGeer, 2015; Roth 2004). Such regulative actions are manifested during joint action (Gilbert 2009; Roth 2004). Some recent studies suggest that people who judge that two persons are walking together in certain conditions are more likely to consider that one of the participants has the right to rebuke the other when he peels off (Gomez-Lavin & Rachar, 2019).

An interesting aspect of these exhibitory and regulative actions is that they are hard to accommodate in a framework that does not presuppose that social agents can hold each other responsible for certain actions. In other words, exhibitory and regulative actions presuppose that agents feel enabled or justified to hold their partners responsible for the expectations they have generated. Such a *normative attitude* is accommodated in our framework to the extent that social partners recognize others' actions as living up to such commitments or frustrating them. In other words, exhibitory and regulative actions are motivated by the normatively-generated expectations of commitments. The normative attitude underpinning exhibitory and regulative actions is explained by the fact that the expectations associated with commitments are normative (Greenspan, 1978; Paprzycka, 1998; Wallace, 1994). That is, when we expect an agent A to do X because she is committed to G, we do not just predict and anticipate X (descriptive expectations) but we are entitled to demand X from A on the basis that she has the obligation to G (normative expectation).

The existence of such a normative attitude can also explain why, in joint actions, social partners may feel motivated to perform actions whose goal is not instrumental but communicative (Vesper et al. 2017: 4). As Clark (2006) emphasizes, joint actions can be divided into two types of actions: the *basic actions*, aimed at achieving the goal per se and *coordinating actions* aimed at facilitating the prediction, adjustments, and coordination between the partners. In other words, social peers pro-actively ensure that the other partner generates the appropriate expectations. Notice that coordinating actions also involve a normative component: they involve the agent's obligation and

accountability. When someone produces a social cue that signals what he is going to do —e.g. making eye contact to indicate one's readiness to engage in a collaborative task— She is implicitly embracing the responsibility to act accordingly. These actions exhibit a high component of social exposure. Making public your intention to perform a particular joint action entitles others to sanction or blame you if you decide to abandon the action. Thus, it is difficult to see how agents could undertake such a responsibility without having a particular understanding of their actions and themselves in terms of commitments.

Now, we can see how commitments play a pivotal role in both prediction and motivation in joint action. Regarding prediction, as we mentioned, when one identifies a partner as such and she confirms this identification, the two agents are establishing a set of commitments. As Michael and Pacherie (2015) have argued, commitments stabilize expectations regarding actions, beliefs, and motivations that reduce different types of uncertainties regarding how to proceed during the joint action, shared background knowledge or whether or not the participants have a common goal. In this sense, both attributing and undertaking individual and joint commitments facilitate prediction and coordination by prescribing courses of actions and individuals' behavior more transparently.

Regarding motivation, commitments can serve as an important catalyst for joint actions in different ways. First, commitments impose an obligation on the author of the commitment to fulfill the appropriate expectation. Such an obligation can be enforced in different ways. For instance, the obligation entitles the recipient to sanction their author or to protest if the expectation is not fulfilled which could provide reasons to the authors to engage in a joint action when he has previously committed to it. Secondly, the author of the commitment may feel motivated to act not because she is inclined to avoid the recipient's possible sanctions but because she may feel identified with the expectations in place. Humans often find others' expectations about their own behavior appealing when they are reasonable (D. K. Lewis, 1969; Sugden, 2000). Thirdly, expressing commitments can also motivate the action of the recipient to the extent that they are costly signals which provides evidence of the author's motivation and serves as a reason for the recipient to engage in the joint action (Michael and Pacherie, 2015; Quillien, 2020). Finally, the author's signals of commitments can prompt a so-called *sense of commitment* (Michael et al., 2016) on the recipient that may act as an endogenous motivation to engage in the joint action. Sense of commitment is the psychological motivation to collaborate, to engage in a joint goal or cooperate with someone because this person expects you to do so and he has somehow manifested such expectation.

These aspects are sufficiently important and pervasive in human-human interaction to motivate a general strategy in social robotics that puts the notion of commitments at the center of the design of robotic agents. This would consist of taking seriously the idea of making robots able to identify human partners as social partners and attribute them a set of commitments depending on relevant features of the situation, but also, able to monitor and respond to the fulfillment or frustration of the relevant expectations. Moreover, robotic agents should be able to undertake commitments, and thus, pro-actively signal them to establish mutual recognition with their partner.

Some exploratory ideas of how robot design can be oriented to the establishment of commitment lay on the use of social signals in robot design. Several studies have demonstrated that equipping robots with the capacity to produce behavioral cues, facial expressions or gaze cues can boost transparency, mutual understanding and trust in HRI (Normoyle et al., 2013; Sciutti et al., 2018; Stanton & Stevens, 2014). For instance, different studies indicate that stereotypical motions, along with straight lines and additional gestures (see Lichtenthäler & Kirsch, 2013 for a review) are pivotal factors for legible robot behavior. In this line, Breazeal et al. (2005) have demonstrated that equipping robots with subtle eye gaze signals—for instance, enabling Leonardo to reestablish eye contact with the human when it finishes its turn, and then, communicating that it is ready to proceed to the next step in the task— improves the subject's understanding and her capacity to quickly anticipate and address potential errors in the task. Moreover, different laboratories have designed different expressive capacities in robots that boost the motivation to interact in humans or that maintain her engaged in the collaborative task. For instance, work in this direction involves facial expressions in anthropomorphic faces with many degrees of freedom (Ahn et al., 2012; Kedzierski et al., 2013), posture (Breazeal et al., 2007) or body motion (Kishi et al., 2013). The efficiency of such approaches is confirmed by studies that demonstrate that human users find robots more persuasive when they use gaze (Ham et al., 2015) or more cooperative when they use cooperative gestures like beckon, give or shake hands (Riek et al., 2010). In sum, we have reasons to believe that robots equipped with the capacity for advancing certain expectations or exhibiting a certain degree of commitment to a particular task can improve human trust and proclivity to engage with a robot.

5. The Affective side of Commitments

In the previous section, we have presented a general approach to the design of social robots capable of engaging in joint action with humans. Such an approach emphasizes the necessity of implementing robots whose capabilities enable them to attribute and undertake commitments. These capabilities must include abilities like pro-active signaling of robot's expectations regarding the human, monitoring and reacting to the frustration and fulfillment of such expectations or manifesting its readiness to behave as expected. The main virtues of this strategy are its level of abstraction and generality, which would allow implementing the mentioned abilities differently depending on the specificity and constraints of specific robotic agents. Now, in which sense are affective states important for joint action in HRI and which role do they play in this general strategy? What can a capacity for emotional expression bring to this commitment approach? Are emotions, motivations or other affective states necessary for establishing and maintaining commitments between humans and robots?

To appreciate the role of affective states in the dynamics of establishing and monitoring commitments, we must consider again the regulative actions associated with holding someone responsible for their commitments. As we mentioned, when we expect an agent A to do X because she is committed to G, we do not just predict and anticipate X (descriptive expectations) but we are entitled to demand X from A on the basis that she has the obligation to G (normative expectation). A first affective aspect associated with commitment seems to be the often emotionally loaded character of these regulatory reactions. Regulatory actions are a series of behaviors oriented toward the other agent or their behavior in order to acknowledge that they have violated (or fulfilled) a commitment. Regulatory actions include subtle cases like manifesting surprise or merely warning the other but also more dramatic actions like expressing resentment and anger or manifesting disapproval or blaming. However, even the milder reactions, finding fault with someone or communicating a judgment of violation, seem to register an emotional state or charge.

But what are emotions and why do we experience negative emotions when someone breaks a commitment? Emotion theorists generally understand emotions as having intentional, evaluative, physiological, expressive, behavioral and phenomenological, components, although they tend to

disagree on which of these components are central or essential to emotion (Michael, 2011; J. J. Prinz, 2004; Scherer, 2005). Social emotions are a particular subset of emotions, both self-directed and other-directed, that depend upon the thoughts, feelings or actions of other people, "as experienced, recalled, anticipated or imagined at first hand" (Hareli and Parkinson 2008: 131). Examples include shame, embarrassment, jealousy, admiration, indignation, gratitude, and so on. Now, commitments impose obligations on the author of the commitment and enable or entitle the recipient to demand these obligations be met. A possible explanation of the emergence of emotional states like feelings of offense, disappointment, disapproval or resentment is precisely their function as a cultural mechanism to prompt punishment and prevent free-riding. While the explanation of small-scale collaboration can appeal to indirect-reciprocity or genetic affinity, the evolution of large-scale cooperation is a puzzling phenomenon from an evolutionary point of view to the extent that it occurs among strangers or non-relatives in large groups. Several authors have proposed that cultural evolution solved the problem by impacting human social psychology (Henrich & Henrich, 2007). In this view, the solution to the problem will lie on a series of evolved norm oriented psychological mechanisms that enable humans to learn and acquire social norms and cultural patterns but also to respond to norm violation and engage in punishment. Such norm-psychology will facilitate large scale cooperation by creating and reinforcing more stable groups (Guzmán et al., 2007).

In this sense, one may argue that social emotions are part of our norm-psychology and that their biological function is precisely to sustain large-scale cultural dynamics by prompting monitoring and punishing responses to norm violations or non-cooperative behaviors in our cultural niche (Fehr & Gächter, 2002). Such a perspective may help us to understand how emotional states can contribute to the maintenance of commitments or why commitments are regarded as credible. Social emotions enable humans to monitor others' commitments and norm compliance and trigger the appropriate punishing or sanctioning response when appropriate. Moreover, one may suggest, the existence of these emotional responses may have promoted, in the appropriate evolutionary circumstances, an inclination to signal and fulfill the commitments, not only in order to avoid punishment and sanctions but as an outcome of developing some sort of avoidance of others' negative emotions — e.g. aversion to others' distress or guilt (Decety & Cowell, 2018; Vaish, 2018) — or as part of a reputation management mechanism. As a result, social emotions

seem to play a fundamental role in the establishment and maintenance of commitments, and thus, in the well-functioning of joint actions. But would this be a definitive positive answer to the question of whether or not commitments require affective states? Does social robotics need to develop emotional robots to establish and maintain joint commitments and protesting as a reaction to the violations of commitments? To what extent should such emotional states be associated with sanctioning or punishment?

Although in evolutionary theories of cooperation, punishment seems to play a fundamental role, it is not so obvious that punishment is necessarily linked to an increase of cooperative behaviors among a population. Several authors have argued that punishment can be ineffective to sustain compliance and even can produce an erosion in cooperation (Dreber et al., 2008; Ostrom et al., 1992). Moreover, the reactive behavioral pattern associated with emotional states like disappointment or disapproval seems to be protesting, attempting to jolt the wrongdoer into seeing things more from the wronged party's perspective or, in the worst case, blaming, rather than punishing or sanctioning. Thus, it is not so obvious that punishing or sanctioning would be the ultimate functional role of social emotions in the development of commitments.

In our view, the role of social emotions in the emergence and maintenance of commitments makes more sense when we have a look at the entire dynamic exchange where these emotional states are often inserted when a violation of commitment is produced. Such an exchange, we will argue, shows that these social emotions play a fundamental role in managing commitment when they work in conjunction with another pivotal affective state: the need to belong. The functional role of social emotions in the context of commitment management only makes sense when the author of the commitments is not only ready to comply with the commitment but also when he is ready to review, re-evaluate and regulate their behavior under the light of the emotional charge produced by a violation. Such inclination to review, re-evaluate and regulate her behavior cannot be explained without postulating a central motivation to care about the others that we will characterize in terms of the need to belong (Baumeister & Leary, 1995; Over, 2016).

To see how, let us consider an observation that several philosophers have made in the context of the debate on moral responsibility (Fricker, 2016; Macnamara, 2013; McGeer, 2012). A fundamental debate in philosophy of mind regarding moral responsibility revolves around the

function of reactive attitudes like blaming in the practice of moral responsibility. However, according to these authors, while the debate has often focused on the fact that reactive attitudes are backward-looking responses to actions and attitudes, which manifest that we profoundly care about other moral actions and responsibilities, the debate has often overlooked the forward-looking dimension of such reactive attitudes and that seems to be fundamental for understanding how such reactive attitudes scaffold the moral agency of others (McGeer 2012). In other words, although reactive attitudes towards others often manifest negative emotions, they also communicate a positive message, namely that we see them as moral agents capable of understanding and living up to the norms of a moral community. In understanding this message, what is essential is that "the recipients of such attitudes understand – or can be brought to understand—that their behaviour has been subjected to normative review, a review that now calls on them to make a normatively "fitting" response" (McGeer 2012: 303). Such a responsiveness involves the wrongdoers behaving reactively in ways commensurate with treating them as responsible agents, as manifested by the co-reactive attitudes of apologizing, giving reasons and reviewing her behavior in a way that reflects a moral sensitivity. So, moral agency is reflected in the disposition to respond reactively to others' reactive attitudes.

Considering commitments from a similar angle, we can see reactions to violation or fulfillment of commitments (e.g. asking for reasons, reprimanding or manifesting surprise or disapproval) precisely as just one step in a dynamic practice where reacting and co-reacting (e.g. giving an excuse, adjusting one's action) to the violation and fulfillment only make sense in the context of the forward-looking function of reactive attitudes: urging the wrongdoer to review her behavior and mental attitude in the light of the expectations generated by the commitments that the wronged party attributed to her. Such a forward-looking function is manifested in the co-reactive responses to reactive attitudes, which are often oriented to apologizing or justifying the violation but more importantly to increase the agential capacities of the agents to the extent that they scaffold their capacity to evaluate their behavior in the light of the commitments involved. As Fricker (2016) claims, even when the wrongdoer does not admit the violation or does not acknowledge her previous commitment, the reactive attitude can produce sufficiently psychological friction on the wrongdoer to orient her mind toward an evaluative stance and lead her to review her motives or reasons to behave as she did.

These dynamics have an important function in terms of interpersonal alignment of intentions and joint beliefs. Imagine two friends, Pablo and Sara, who decide to paint a house together and when Sara takes her brush to start, Pablo takes his equipment and goes to another room. The following exchange may ensue:

- Pablo: “I thought we were going to do this together."
- Sara: “This way we'll go faster."
- Pablo: “I don't want to get bored doing this.”
- Sara: “Okay, you're right."

This kind of dynamic, where two agents react and co-react to the frustration of an expectation associated with a joint commitment, shows us how regulative actions may serve to align intentions and beliefs during a joint action, which has important consequences for prediction and motivation. However, the lesson we would like to draw from this analysis is different. Notice that the dynamic trajectory of reactions and co-reactions is based on the premise that both agents care about each other, the joint commitments and their mutual expectations. Reactive and co-reactive attitudes only make sense if the agents involved are the kind of agents that care about living up to the expectations and demands of commitments and care about exercising their agential capacities expressed through evaluating and regulating their actions in accordance with commitments and their normative expectations. Such capacities, then, can only make sense if the agents involved care and value their social relations and bond with other agents, which seems to imply an important affective factor.

Elsewhere, we have argued that a major human motivation for explaining why commitments are credible is the need to belong (Fernández Castro & Pacherie, 2020). The need to belong is the need that individuals have for frequent, positively valenced interactions with other people within a framework of long-lasting concern for each other’s welfare (Baumeister and Leary 1995; Over 2016). The need to belong is a need, in the sense that long-term social bonds are crucially important to well-being and, conversely, their lack leads to ill-effects. The need to belong explains why humans find acting with others rewarding, why they tend to give attentional priority to social cues, or why many joint actions are motivated not just by the desire to achieve the intended outcome of their shared intention but also by the desire to obtain this social reward. However, in

contrast with other postulated general prosocial motivations (Godman, 2013; Godman et al., 2014), the need to belong is neither indiscriminate nor unbounded but rather manifests selectivity. So, humans prefer repeated interactions with the same persons to interactions with a constantly changing sequence of partners, they devote more energy to preserving and consolidating existing bonds than to interacting with strangers and once the minimum quantity and quality of social bonds are surpassed, their motivation to create new bonds diminishes.

For our purpose, the importance of the need to belong lies in its capacity for giving an account of why we care for, or value, both our commitments to others and the commitments others have to ourselves. Without such a capacity, we could not explain why one experiences social emotions when someone breaks a commitment or why one feels the psychological pressure to evaluate one's performance and commitments in light of the reactions of others. Fernández Castro & Pacherie (2020) argue that although there is a large diversity of motivations that may be involved in why we commit ourselves and why we remain faithful to those commitments –e.g. reputation or avoidance of negative social emotions—, the need to belong is, from a developmental point of view, a more basic motivation. Now, the presented dynamic also allows us to see how the need to belong might be involved in the emergence of these other motivations. First, an agent's social emotions emerge as reactive attitudes to others' attitudes or behavior that may trigger her capacity for reviewing, re-evaluating and regulating her behavior. However, social emotions can only serve such a function if the wronged party cares about the wrongdoer's commitments, even in cases where the wronged party does not necessarily benefit from the result of the joint action, and if the wrongdoer in turn cares about the wronged party sufficiently to motivate a co-reaction. Second, although we may abide by our commitments or provide justifications to explain why we violated a commitment simply in order to promote or preserve a positive reputation, such management of reputation can only emerge as a result of the dynamics of previous reactions and co-reactions premised on the idea that we care about others. The notion of prestige and reputation is tied to the image that others have of us and how such an image may impact our social relations with them. Without a motivation to engage in such social relations, and probably without a dynamic of manifested positive or negative attitudes toward the resulting outcomes of such relations, the notion of reputation no longer makes sense.

In a nutshell, the establishment and maintenance of commitments involve a dynamic trajectory of signals, reactions, and co-reactions. This trajectory seems to involve at least two emotional components. First, a series of socio-emotional states triggered by frustration (and fulfillment) of commitments that produce a series of regulative behaviors signaling recognition of the violation, disapproval of it and warning and aimed at making the wrongdoer review her behavior. Second, a pro-social motivation that prompts both parties to establish commitments but more importantly makes them ready to review, re-evaluate and regulate their behavior in the light of their violation, and thus, scaffold their agential capacities as a team.

6. Emotional Robots and Commitments

We have proposed that the establishment of commitments during joint action necessitates the mutual influence of two types of affective states: social emotions and prosocial motivations. The interaction between these types of affective states explains different features of the interaction between partners like how they hold each other responsible for commitments, how they react and co-react when the expectations associated with the commitments are violated or fulfilled or why one would assume the social costs of undertaking them. If our argumentation is compelling, one must wonder whether implementing these types of affective states is necessary for the design of social robots able to establish commitments with a human partner in collaborative contexts: May developers attempt to implement affective states in robots? Or would it be sufficient to mimic or imitate the behavioral profile of such states? Could we find a design solution that would devise functional substitutes for such affective states without implementing them properly? In this section, we propose three possible answers to these questions and discuss their potential scopes and problems. First, we discuss a minimalist option that would involve faking emotional states. Second, we discuss a maximalist option that would endow robots with affective states or quasi-states that are equal or at least similar to human affective states. Finally, we discuss an alternative option that attempts to design different solutions aimed at establishing commitments without using or faking affective states.

To begin with, the minimalist option would attempt to endow robots with a capacity to manifest certain behavioral reactions recognizable by the human as stereotypical affective reactions that facilitate the establishment and maintenance of commitments without necessarily implementing other dimensions of affective states like arousal or specific action tendencies. Although not straightforwardly connected to commitments, the use of emotional signals to communicate

different information or to maintain the human engaged in a collaborative task is a common strategy in social robotics. Several labs (Breazeal, 2003; Craig et al., 2010; Kishi et al., 2013; Oberman et al., 2007; C. Wendt et al., 2008;) are equipping robots with different emotional expressions in order to prompt empathy or pro-social attitudes in the human agent, so robots "could potentially tap into the powerful social motivation system inherent in human life, which could lead to more enjoyable and longer lasting human-robot interactions" (Oberman et al. 2007: 2195). Indeed, we can find some developments that could somehow support the realization of the minimal option. On the one hand, some implementations have used emotional expressions as indications of robot's failures (Hamacher et al., 2016; Reyes et al., 2016; Spexard et al., 2008); these emotional expressions can facilitate human interpretation and trigger helping behavior in a way similar to the types of reactions triggered by the social emotions involved in commitments mediated interactions. For example, Hamacher et al. use a BERT2, a humanoid robotic assistant, in a making-omelet task in order to test users' preferences. BERT2 was able to express sadness and apologize when dropping an egg. The studies demonstrated that subjects preferred to interact with the robot able to display such expressions than with the more efficient robot without such social capacities. On the other hand, we can find some studies that give support to the idea that some indications of motivation and commitments on behalf of the robot can boost the human feeling of obligation to remain committed to the action (Michael & Salice, 2017; Powell & Michael, 2019; Vignolo et al., 2019). In Vignolo et al.'s experiment, an iCub robot interacted with children in a teaching skills exchange. In the experiment, the subjects were exposed to two different conditions: in the high effort condition, the iCub slowed down his movements when repeating a demonstration for the human learner, whereas in the low effort condition he sped the movements up when repeating the demonstration. Then, the human had to reciprocate teaching the iCub a new skill. They found that subjects exposed to the high-effort condition were more likely to reciprocate and make more effort to teach the robot. These experiments seem to provide some partial support to the idea that exhibiting prosocial motivations, which indirectly ensure the level of commitment to a task, may facilitate the maintenance of commitments in HRI. In a nutshell, we have reasons to believe that faking emotional expressions is viable, and thus, in principle, a good way to attempt to implement a mechanism for establishing and maintaining commitments in HRI.

The minimalist option, however, presents two significant problems. The first is the problem of responsible agency. As we argued above, the role of affective states in the establishment and maintenance of commitments is twofold. First, social emotions play a role in the production of regulative behaviors that acknowledge the partner's responsibility tied to commitments while

triggering the appropriate co-reaction. Second, such co-reaction is motivated by a general prosocial motivation which is manifested in the tendency to provide excusing explanations or reparations in the form of apology but more importantly in the inclination to review one's own behavior in the light of the partner's acknowledgment and reactive attitudes. Now, keeping this latter function in mind, we can see one of the problems of the minimalist option. The role of social motivation in the dynamic of reactions and co-reactions that maintain commitments is not just to trigger expressive or behavioral responses; its function is also to induce changes in the dispositional profile of the subject and thus shape responsible agency. When one receives blame or approval for violating or fulfilling a commitment, one's prosocial motivation may mobilize the appropriate change in the capacity for being properly sensitive to one's commitments, in the care taken to regulate one's actions according to these commitments and in the amount of attention paid to the relevant aspects of the situation in subsequent actions. In the case of fulfillment, the change propitiated by the need to belong can be instantiated in a feeling of reward associated with the action that can translate into a reinforcement of the appropriate dispositions and cognitive processes associated with it. In the case of violation, the change can be produced by a feeling of discomfort because the violation is causing a negative balance in our relationship with the other. Be that as it may, the function of the need to belong as a motivation for establishing and maintaining commitments is not only connected to expressions that can be mimicked but to changes in the dispositions and cognitive processes of the agent. Thus, the minimalist option does not seem to cover all the relevant functions of the affective states necessary for producing commitments.

Moreover, there is a second problem with the minimalist option. Faking emotional responses seems to produce important ethical concerns. As Brinck and Balkenius (2018)) have argued, sociable robots that fake emotions exploit the emotional vulnerability of human users, which has potential harming consequences for their integrity:

> The fact that [robots] mimic human emotion and interact via bodily and facial expression of emotion encourages users to grow emotional attachments to them, whereas the robots themselves do not have feelings of the human kind, but display cue-based behavior. Users who invest themselves in the robot and become emotionally dependent on them risk being hurt, suffer depression, and develop mental and physical illnesses (Brinck and Balkenius, 2018: 3 - 4)

In other words, social robotics construct HRI in a way that exploits the emotional profile of the users and could have serious harming implications for them. Moreover, exploiting human emotions for efficiency purposes, without considering human preferences or needs during the situation, goes against the very basis of sociality itself where the partners often negotiate in order to align their intentions and beliefs to motivate each other to remain engaged in the task.

These problems may lead us to opt for a maximalist option regarding robotic affective states. According to this option, one may attempt to design social robots with real internal states that do not just fake human emotional responses but have real powers to modify the behavior of the robots and guide its adaptation to the social or non-social environment. The key aspect of affective states is to provide the agent with the capacity for selecting behaviors depending on different parameters (Canamero, 2003). For instance, several developers have tried to implement architectures able to adapt the robot to the social or learning environment thanks to modules or devices that assign different quasi motivational internal states depending on the information they receive and that trigger different responses according to such states (Hiolle et al., 2012, 2014; Tanevska et al., 2018, 2019). In an experiment, Tanevska et al (2019) equipped an iCub robot with an adaptive architecture with a state machine that represented the robot's level of social comfort and with a social adaptative machine able to track the state of the robot and produce different reactions depending on the level of saturation. For instance, when its level of social comfort was optimal, the iCub would play with its toys and interact with the user while it would try to attract her attention when the level was non-optimal and it would disengage when getting oversaturated.

Following this idea, one may attempt to develop robots able not only to properly expressively react to social cues or violation of expectations in a way similar to what a human would do but also to regulate their dispositions and cognitive processes correspondingly. For instance, a robot could be more cautious (double-checking human cue-based behaviors) and less engaging with those users who had violated a commitment as a way to instantiate a type of quasi-disappointment state or display more prosocial behavior and engagement strategies when it has violated a commitment to repair the relationship with the user. Such a maximalist strategy would facilitate the avoidance of the problem of responsible agency. Enabling robots with the capacity for reviewing or assessing their own behavioral and cognitive capabilities in response to the reactive attitudes of the human with consequences for the rest of the joint action or future social interactions with the same or distinct users is precisely the type of learning capacities one may expect from the normative

aspect of commitments. As such, the affective states will play the necessary functions associated with commitments in joint action.

On the other hand, the maximalist option also comes with more technical problems and computational costs than the minimalist option. Implementing the capacities to detect emotions, gestures, and actions has turned out to be an especially challenging enterprise in realistic environments (Yang et al., 2018). In order to deal with such a problem, developers often consider different proxies like the mere presence of the human or her face, the distance to the robot or tactile stimuli as inputs that trigger particular emotional responses (see e.g. Breazeal 2004). In the case of joint actions and commitment instantiation, one may opt for a similar solution to detect the relevant aspects of the situation and trigger social emotions as the appropriate reaction to violation or fulfillment of commitments. To take an example, Clodic et al. (2006) implemented a robot guide at a museum. In this experiment, they defined the task of the robot in terms of commitments and assumed that the human was fulfilling the commitment of following the robot when detecting his presence behind. In the minimalist view, one may use such a type of proxies to react to the appropriate emotional responses, for instance, looking back and smiling as a sign of approval or ask for explanations in an angry tone of voice if the user stops following the robot. However, it is difficult to see how the maximalist option may exploit such proxies when implementing co-reactivity. To successfully modify its behavior according to human social emotions, the robot must be able to distinguish very subtle human reactions like indignation, approval, disapproval, guilt, disappointment and so on. As such, the technical limitations associated with emotional recognition in social robotics is much more pressing in the case of the maximalist option.

Moreover, as we stated before, the maximalist option requires not only the capacity to detect the appropriate emotional responses but also the capacity to evaluate and learn to change one's behavior in accordance with these responses along with the capability to execute the appropriate repair strategy in every case. The conjunction of these capacities does not only multiply the problems regarding technical issues but also the computational costs generated by the necessity of processing a larger quantity of information, by the necessity of having more perceptual and behavioral modules or devices, and by the necessity of integrating all this information. Thus, the maximalist option does not only have to deal with some ethical concerns on its own, but also, with more technical problems and computational costs than the minimalist option.

Finally, an alternative to the minimalist and maximalist solution would be to replace emotions and affective states with functional substitutes, that is other types of reasoning or communicative devices which do not necessarily simulate human-like emotional responses or affective states, but can served the commitment-supporting functions served by emotions in humans; so, robots could be enabled to signal commitments, communicate their violations and fulfillment, to negotiate reparations or evaluate and select their subsequent actions by using alternative mechanisms. On the one hand, given that the pivotal function of emotional states like social emotions in establishing commitments depends largely on their expressive power, alternative expressive strategies like explicit verbal communication (Mavridis, 2015) for reacting to frustration or fulfillment of commitments or more neutral signals like lights or symbols in a screen (Baraka et al., 2016) might be used for the same purpose. On the other hand, the role of affective states in the agent's self-evaluation and in the selection of behavior could be implemented through reasoning capacities. To the extent that robots may be enabled to understand humans' reactive attitudes and commitments signals, they could process the given information to evaluate their own behavior and cognitive processes, so in principle, this alternative solution could also serve to implement the dynamic set of reactions and co-reactions associated with the maintenance of commitments.

Now, the alternative option could, in principle, avoid the first ethical concern to the extent that they can use emotionally neutral expressive strategies to establish and manage commitments, so humans would be less emotionally engaged with the robotic agent and less vulnerable to emotional exploitation. However, like in the case of the maximalist option, the second concern can only be avoided if we put the human preferences, values and integrity at the center of the reasoning capacities that modulate the robotic behavior. Now the question is could we substitute quasi-motivational states for reasoning capacities without missing an element? As we stated above, affective states inform us about how the world is in relation to our own well-being. For instance, the state of fear informs us that a particular object or feature of our environment is dangerous in the sense that it can damage our physical integrity. Moreover, these affective states are also intrinsically connected to actions and can trigger effective behavioral responses. Certainly, a reasoning architecture could infer the relevant commitments a robot should undertake given certain human responses or in what ways it should modify its behavior depending on a state it infers from the human's action. However, in the wild, autonomous robots may have to decide between different courses of action, some of which may involve decisions that have consequences for the human partner or for itself. Selecting one course of action over others is, at the end of the day, something that may involve preferences or motivations that relate to what the

robot “cares about” or not. As such, it is difficult to see how we could have an autonomous robot without an architecture that regulates or modulates its behavior in order to maintain certain homeostatic levels that we may identify with preferences. In this particular case, a preference for maintaining a well-balanced relation with the human partner, and thus, a preference for behaviors that facilitate the fulfillment of joint commitments and goals.

7. Concluding Remarks

Solving the problems of motivations and predictions that social robotics encounters in joint action for HRI requires, we believe, enabling robots with the capacity for establishing and maintaining commitments. While improving the prediction of robots’ behavior and boosting human motivation to interact with them necessitate establishing a mutual recognition between the partners, we have argued that such mutual recognition cannot be simply implemented through embodied strategies. The different physical and functional features of robotic agents along with the diversity of strategies one may use to identify others as social agents demand that we attribute to them different commitments depending on different physical features, social norms, or contextual parameters. In this sense, our proposal has the advantage of providing a framework to improve prediction and motivation while remaining at the right level of abstraction and being compatible with a larger set of communicative strategies to implement recognition.

Further, we have defended that the establishment and maintenance of commitments in human-human interaction depends on at least two fundamental affective states: social emotions and the pro-social motivation associated with the need to belong. In this sense, we have asked ourselves to what extent a robotic architecture could either incorporate such affective states and provide functional substitutes for them. Finally, we have proposed three options and evaluated their possible ethical and technical problems.

References

Huang, C. M., & Thomaz, A. L. (2011). Effects of responding to, initiating and ensuring joint attention in human-robot interaction. In *2011 Ro-Man* (pp. 65-71). IEEE.

https://doi.org/10.1080/713752551

Quillien, T. (2020). Evolution of conditional and unconditional commitment. *Journal of*

Theoretical Biology, *492*, 110204.

Ahn, H. S., Lee, D.-W., Choi, D., Lee, D.-Y., Hur, M., & Lee, H. (2012). Difference of Efficiency in Human-Robot Interaction According to Condition of Experimental Environment. In S. S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, & M.-A. Williams (Eds.), *Social Robotics* (Vol. 7621, pp. 219–227). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34103-8_22

Anjum, M. L., Ahmad, O., Rosa, S., Yin, J., & Bona, B. (2014). Skeleton Tracking Based Complex Human Activity Recognition Using Kinect Camera. In M. Beetz, B. Johnston, & M.-A. Williams (Eds.), *Social Robotics* (Vol. 8755, pp. 23–33). Springer International Publishing. https://doi.org/10.1007/978-3-319-11973-1_3

Austin, J. (1962). *How to do things with words.* Clarendon Press.

Baraka, K., Paiva, A., & Veloso, M. (2016). Expressive Lights for Revealing Mobile Service Robot State. In L. P. Reis, A. P. Moreira, P. U. Lima, L. Montano, & V. Muñoz-Martinez (Eds.), *Robot 2015: Second Iberian Robotics Conference* (pp. 107–119). Springer International Publishing. https://doi.org/10.1007/978-3-319-27146-0_9

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497–529. https://doi.org/10.1037/0033-2909.117.3.497

Benamara, N. K., Val-Calvo, M., Álvarez-Sánchez, J. R., Díaz-Morcillo, A., Ferrández Vicente, J. M., Fernández-Jover, E., & Stambouli, T. B. (2019). Real-Time Emotional Recognition for Sociable Robotics Based on Deep Neural Networks Ensemble. In J. M. Ferrández Vicente, J. R. Álvarez-Sánchez, F. de la Paz López, J. Toledo Moreo, & H. Adeli (Eds.), *Understanding the Brain Function and Emotions* (Vol. 11486, pp. 171–180). Springer International Publishing. https://doi.org/10.1007/978-3-030-19591-5_18

Blakemore, S. J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience*, *2*, 561–567.

Brandom, R. B. (2007). The structure of desire and recognition: Self-consciousness and self-constitution. *Philosophy & Social Criticism*, *33*(1), 127–150. https://doi.org/10.1177/0191453707071389

Bratman, M. E. (2014). *Shared Agency*. Oxford University Press.

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, *59*(1–2), 119–155. https://doi.org/10.1016/S1071-5819(03)00018-1

Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G., & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference On*, 708–713.

Breazeal, C., Wang, A., & Picard, R. (2007). Experiments with a robotic computer: Body, affect and cognition interactions. *Proceeding of the ACM/IEEE International Conference on Human-Robot Interaction - HRI '07*, 153. https://doi.org/10.1145/1228716.1228737

Brinck, I., & Balkenius, C. (2018). Mutual Recognition in Human-Robot Interaction: A Deflationary Account. *Philosophy & Technology*, *33*(1), 53–70. https://doi.org/10.1007/s13347-018-0339-x

Canamero, D. (2003). Designing Emotions for Activity Selection. In R. Trappl, P. Petta, & S. Payr (Eds.), *Emotions in humans and artifacts* (Vols 115–148). The MIT Press.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893–910. https://doi.org/10.1037/0022-3514.76.6.893

Chudek, M., & Henrich, J. (2011). Culture-gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences*, *15*(5), 218–226. https://doi.org/10.1016/j.tics.2011.03.003

Clark, H. H. (2006). Social actions, social commitments. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of human sociality: Culture, cognition, and interaction* (pp. 126–150). Berg.

Clodic, A., Fleury, S., Alami, R., Chatila, R., Bailly, G., Brethes, L., Cottret, M., Danes, P., Dollat, X., Elisei, F., Ferrane, I., Herrb, M., Infantes, G., Lemaire, C., Lerasle, F., Manhes, J., Marcoul, P., Menezes, P., & Montreuil, V. (2006). Rackham: An Interactive Robot-Guide. *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 502–509. https://doi.org/10.1109/ROMAN.2006.314378

Craig, R., Vaidyanathan, R., James, C., & Melhuish, C. (2010). Assessment of human response to robot facial expressions through visual evoked potentials. *2010 10th IEEE-RAS International Conference on Humanoid Robots*, 647–652. https://doi.org/10.1109/ICHR.2010.5686272

Davidson, D. (1991). Three Varieties of Knowledge. *Royal Institute of Philosophy Supplements*, *30*, 153–166. https://doi.org/10.1017/S1358246100007748

Decety, J., & Cowell, J. M. (2018). Interpersonal harm aversion as a necessary foundation for morality: A developmental neuroscience perspective. *Development and Psychopathology*, *30*(1), 153–164. https://doi.org/10.1017/S0954579417000530

Dennett, D. (2009). Intentional Systems Theory. In A. Beckermann, B. P. McLaughlin, & S. Walter (Eds.), *The Oxford Handbook of Philosophy of Mind*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199262618.003.0020

Depue, R. A., & Morrone-Strupinsky, J. V. (2005). A neurobehavioral model of affiliative bonding: Implications for conceptualizing a human trait of affiliation. *Behavioral and Brain Sciences*, *28*(03). https://doi.org/10.1017/S0140525X05000063

Devin, S., & Alami, R. (2016). *An implemented theory of mind to improve human-robot shared plans execution*. 319–326.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, *452*(7185), 348–351. https://doi.org/10.1038/nature06723

Elsner, B., & Hommel, B. (2001). Effect anticipation and action control. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 229–240.

https://doi.org/10.1037//0096-1523.27.1.229

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140. https://doi.org/10.1038/415137a

Fernandez Castro, V. (2020). Regulation, Normativity and Folk Psychology. *Topoi*, *39*(1), 57–67. https://doi.org/10.1007/s11245-017-9511-7

Fernández Castro, V., & Heras-Escribano, M. (2020). Social Cognition: A Normative Approach. *Acta Analytica*, *35*(1), 75–100. https://doi.org/10.1007/s12136-019-00388-y

Fernández Castro, V., & Pacherie, E. (2020). Joint actions, commitments and the need to belong. *Synthese*. https://doi.org/10.1007/s11229-020-02535-0

Fricker, M. (2016). What's the Point of Blame? A Paradigm Based Explanation: What's the Point of Blame? *Noûs*, *50*(1), 165–183. https://doi.org/10.1111/nous.12067

Gilbert, M. (1992). *On social facts*. Princeton University Press.

Gilbert, M. (2009). Shared intention and personal intentions. *Philosophical Studies*, *144*(1), 167–187. https://doi.org/10.1007/s11098-009-9372-z

Godman, M. (2013). Why we do things together: The social motivation for joint action. *Philosophical Psychology*, *26*(4), 588–603. https://doi.org/10.1080/09515089.2012.670905

Godman, M., Nagatsu, M., & Salmela, M. (2014). The Social Motivation Hypothesis for Prosocial Behavior. *Philosophy of the Social Sciences*, *44*(5), 563–587. https://doi.org/10.1177/0048393114530841

Gomez-Lavin, J., & Rachar, M. (2019). Normativity in joint action. *Mind & Language*, *34*(1), 97–120. https://doi.org/10.1111/mila.12195

Greenspan, P. S. (1978). Behavior Control and Freedom of Action. *The Philosophical Review*, *87*(2), 225–240. JSTOR. https://doi.org/10.2307/2184753

Guzmán, R. A., Rodrıguez-Sickert, C., & Rowthorn, R. (2007). When in Rome, do as the Romans do: The coevolution of altruistic punishment, conformist learning, and

cooperation. *Evolution and Human Behavior*, *28*, 112–117.

Ham, J., Cuijpers, R. H., & Cabibihan, J.-J. (2015). Combining Robotic Persuasive Strategies: The Persuasive Power of a Storytelling Robot that Uses Gazing and Gestures. *International Journal of Social Robotics*, *7*(4), 479–487. https://doi.org/10.1007/s12369-015-0280-4

Hamacher, A., Bianchi-Berthouze, N., Pipe, A. G., & Eder, K. (2016). Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction. *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 493–500. https://doi.org/10.1109/ROMAN.2016.7745163

Hareli, S., & Parkinson, B. (2008). What's Social About Social Emotions? *Journal for the Theory of Social Behaviour*, *38*(2), 131–156. https://doi.org/10.1111/j.1468-5914.2008.00363.x

Harrison, S. J., & Richardson, M. J. (2009). Horsing Around: Spontaneous Four-Legged Coordination. *Journal of Motor Behavior*, *41*(6), 519–524. https://doi.org/10.3200/35-08-014

Henrich, N., & Henrich, J. P. (2007). *Why humans cooperate: A cultural and evolutionary explanation*. Oxford University Press.

Hiolle, A., Cañamero, L., Davila-Ross, M., & Bard, K. A. (2012). Eliciting caregiving behavior in dyadic human-robot attachment-like interactions. *ACM Transactions on Interactive Intelligent Systems*, *2*(1), 3:1–3:24. https://doi.org/10.1145/2133366.2133369

Hiolle, A., Lewis, M., & Cañamero, L. (2014). Arousal regulation and affective adaptation to human responsiveness by a robot that explores and learns a novel environment. *Frontiers in Neurorobotics*, *8*. https://doi.org/10.3389/fnbot.2014.00017

Huang, C.-M., & Thomaz, A. L. (2010). Joint Attention in Human-Robot Interaction. *AAAI Fall Symposium*.

Kedzierski, J., Muszynski, R., Zoll, C., Oleksy, A., & Frontkiewicz, M. (2013). EMYS—Emotive

Head of a Social Robot. *International Journal of Social Robotics*, *5*(2), 237–249. https://doi.org/10.1007/s12369-013-0183-1

Keller, P. E., Novembre, G., & Hove, M. J. (2014). Rhythm in joint action: Psychological and neurophysiological mechanisms for real-time interpersonal coordination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658), 20130394–20130394. https://doi.org/10.1098/rstb.2013.0394

Kishi, T., Kojima, T., Endo, N., Destephe, M., Otani, T., Jamone, L., Kryczka, P., Trovato, G., Hashimoto, K., Cosentino, S., & Takanishi, A. (2013). Impression survey of the emotion expression humanoid robot with mental model based dynamic emotions. *2013 IEEE International Conference on Robotics and Automation*, 1663–1668. https://doi.org/10.1109/ICRA.2013.6630793

Knoblich, G., Butterfill, S., & Sebanz, N. (2011). Psychological Research on Joint Action. In *Psychology of Learning and Motivation* (Vol. 54, pp. 59–101). Elsevier. https://doi.org/10.1016/B978-0-12-385527-5.00003-6

Kwon, M., Jung, M., & Knepper, R. (2016). *Human Expectations of Social Robots*. 463–464.

Laitinen, A. (2016). Robots and Human Sociality: Normative Expectations, the Need for Recognition, and the Social Bases of Self-Esteem. In J. Seibt, M. Nørskov, & S. S. Andersen (Eds.), *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016 / TRANSOR 2016* (Vol. 290, pp. 313–322). IOS Press. https://www.medra.org/servlet/aliasResolver?alias=iospressISBN&isbn=978-1-61499-707-8&spage=313&doi=10.3233/978-1-61499-708-5-313

Lewis, D. K. (1969). *Convention: A philosophical study*. Harvard University Press.

Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In *Foundations of trusted autonomy* (pp. 135–159). Springer.

Lichtenthäler, C., & Kirsch, A. (2013). Towards Legible Robot Navigation—How to Increase the Intend Expressiveness of Robot Navigation Behavior. *International Conference on Social*

Robotics - Workshop Embodied Communication of Goals and Intentions. https://hal.archives-ouvertes.fr/hal-01684307

Lo Presti, P. (2013). Situating Norms and Jointness of Social Interaction. *Cosmos and History: The Journal of Natural and Social Philosophy*, *9*(1), 225–248.

Macnamara, C. (2013). “Screw you!” & “thank you”. *Philosophical Studies*, *165*(3), 893–914. https://doi.org/10.1007/s11098-012-9995-3

Maibom, H. L. (2007). Social Systems. *Philosophical Psychology*, *20*(5), 557–578. https://doi.org/10.1080/09515080701545981

Mavridis, N. (2015). A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, *63*, 22–35. https://doi.org/10.1016/j.robot.2014.09.031

McGeer, V. (2012). Co-reactive attitudes and the making of moral community. In C. MacKenzie & R. Langdon (Eds.), *Emotions, Imagination and Moral Reasoning (eds., C. MacKenzie & R. Langdon)*. Psychology Press.

McGeer, V. (2015). Mind-making practices: The social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations*, *18*(2), 259–281. https://doi.org/10.1080/13869795.2015.1032331

Michael, J. (2011). Shared Emotions and Joint Action. *Review of Philosophy and Psychology*, *2*(2), 355–373. https://doi.org/10.1007/s13164-011-0055-2

Michael, J., & Pacherie, E. (2015). On Commitments and Other Uncertainty Reduction Tools in Joint Action. *Journal of Social Ontology*, *1*(1), 89–120. https://doi.org/10.1515/jso-2014-0021

Michael, J., & Salice, A. (2017). The Sense of Commitment in Human–Robot Interaction. *International Journal of Social Robotics*, *9*(5), 755–763. https://doi.org/10.1007/s12369-016-0376-5

Michael, J., Sebanz, N., & Knoblich, G. (2016). The Sense of Commitment: A Minimal

Approach. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01968

Michael, J., & Székely, M. (2018). The Developmental Origins of Commitment. *Journal of Social Philosophy*, *49*(1), 106–123. https://doi.org/10.1111/josp.12220

Mori, M. (1970). The uncanny valley. *Energy*, *7*(4), 33–35.

Normoyle, A., Badler, J. B., Fan, T., Badler, N. I., Cassol, V. J., & Musse, S. R. (2013). Evaluating perceived trust from procedurally animated gaze. *Proceedings of Motion on Games*, 141–148. https://doi.org/10.1145/2522628.2522630

Oberman, L. M., McCleery, J. P., Ramachandran, V. S., & Pineda, J. A. (2007). EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing*, *70*(13–15), 2194–2203. https://doi.org/10.1016/j.neucom.2006.02.024

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants With and Without a Sword: Self-Governance is Possible. *The American Political Science Review*, *86*(2), 404–417. JSTOR. https://doi.org/10.2307/1964229

Over, H. (2016). The origins of belonging: Social motivation in infants and young children. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686). https://doi.org/10.1098/rstb.2015.0072

Pacherie, E. (2011). Framing Joint Action. *Review of Philosophy and Psychology*, *2*(2), 173–192. https://doi.org/10.1007/s13164-011-0052-5

Pacherie, E. (2013). Intentional joint agency: Shared intention lite. *Synthese*, *190*(10), 1817–1839. https://doi.org/10.1007/s11229-013-0263-7

Pandey, A. K., & Alami, R. (2010). A framework towards a socially aware Mobile Robot motion in Human-Centered dynamic environment. *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5855–5860. https://doi.org/10.1109/IROS.2010.5649688

Paprzycka, K. (1998). Normative Expectations, Intentions, and Beliefs. *Southern Journal of Philosophy*, *37*(4), 629–652. https://doi.org/10.1111/j.2041-6962.1999.tb00886.x

Powell, H., & Michael, J. (2019). Feeling committed to a robot: Why, what, when and how? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1771), 20180039. https://doi.org/10.1098/rstb.2018.0039

Prinz, J. J. (2004). *Gut Reactions: A Perceptual Theory of the Emotions*. Oxford University Press.

Prinz, W. (1997). Perception and Action Planning. *European Journal of Cognitive Psychology*, *9*(2), 129–154. https://doi.org/10.1080/713752551

Ramberg, B. (2000). Post-ontological Philosophy of Mind: Rorty versus Davidson. In *Rorty and his critics*. Blackwell Publishers.

Ramenzoni, V. C., Riley, M. A., Shockley, K., & Davis, T. (2008). Carrying the height of the world on your ankles: Encumbering observers reduces estimates of how high an actor can jump. *Quarterly Journal of Experimental Psychology*, *61*(10), 1487–1495. https://doi.org/10.1080/17470210802100073

Reyes, M., Meza, I., & Pineda, L. A. (2016). The Positive Effect of Negative Feedback in HRI Using a Facial Expression Robot. In J. T. K. V. Koh, B. J. Dunstan, D. Silvera-Tawil, & M. Velonaki (Eds.), *Cultural Robotics* (pp. 44–54). Springer International Publishing. https://doi.org/10.1007/978-3-319-42945-8_4

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R. L., & Schmidt, R. C. (2007). Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human Movement Science*, *26*(6), 867–891. https://doi.org/10.1016/j.humov.2007.07.002

Riek, L. D., Rabinowitch, T.-C., Bremner, P., Pipe, A. G., Fraser, M., & Robinson, P. (2010). Cooperative gestures: Effective signaling for humanoid robots. *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 61–68. https://doi.org/10.1109/HRI.2010.5453266

Roth, A. S. (2004). Shared Agency and Contralateral Commitments. *The Philosophical Review*,

113(3), 359–410. JSTOR.

Sahaï, A., Pacherie, E., Grynszpan, O., & Berberian, B. (2017). Predictive Mechanisms Are Not Involved the Same Way during Human-Human vs. Human-Machine Interactions: A Review. *Frontiers in Neurorobotics*, *11*. https://doi.org/10.3389/fnbot.2017.00052

Sanders, T. L., Schafer, K. E., Volante, W., Reardon, A., & Hancock, P. A. (2016). Implicit Attitudes Toward Robots. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 1746–1749. https://doi.org/10.1177/1541931213601400

Satne, G. (2014). What binds us together: Normativity and the second person. *Philosophical Topics*, *42*(1), 43–61.

Scherer, K. R. (2005). What Are Emotions? And How Can They Be Measured? *Social Science Information*, *44*(4), 695–729. https://doi.org/10.1177/0539018405058216

Sciutti, A., Mara, M., Tagliasco, V., & Sandini, G. (2018). Humanizing Human-Robot Interaction: On the Importance of Mutual Understanding. *IEEE Technology and Society Magazine*, *37*(1), 22–29. https://doi.org/10.1109/MTS.2018.2795095

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76. https://doi.org/10.1016/j.tics.2005.12.009

Siposova, B., Tomasello, M., & Carpenter, M. (2018). Communicative eye contact signals a commitment to cooperate for young children. *Cognition*, *179*, 192–201. https://doi.org/10.1016/j.cognition.2018.06.010

Sisbot, E. A., Marin-Urias, L. F., Broquère, X., Sidobre, D., & Alami, R. (2010). Synthesizing Robot Motions Adapted to Human Presence. *International Journal of Social Robotics*, *2*(3), 329–343. https://doi.org/10.1007/s12369-010-0059-6

Spexard, T. P., Hanheide, M., Li, S., & Wrede, B. (2008). *Error Detection and Recovery for Advanced Human-Robot-Interaction*. 6.

Stanton, C., & Stevens, C. J. (2014). Robot Pressure: The Impact of Robot Eye Gaze and

Lifelike Bodily Movements upon Decision-Making and Trust. In M. Beetz, B. Johnston, & M.-A. Williams (Eds.), *Social Robotics* (pp. 330–339). Springer International Publishing. https://doi.org/10.1007/978-3-319-11973-1_34

Steckenfinger, S. A., & Ghazanfar, A. A. (2009). Monkey visual behavior falls into the uncanny valley. *Proceedings of the National Academy of Sciences*, *106*(43), 18362–18366. https://doi.org/10.1073/pnas.0910063106

Sugden, R. (2000). The Motivating Power of Expectations. In J. Nida-Rümelin & W. Spohn (Eds.), *Rationality, Rules, and Structure* (pp. 103–129). Springer Netherlands. https://doi.org/10.1007/978-94-015-9616-9_7

Tai, Y. F., Scherfler, C., Brooks, D. J., Sawamoto, N., & Castiello, U. (2004). The Human Premotor Cortex Is 'Mirror' Only for Biological Actions. *Current Biology*, *14*(2), 117–120. https://doi.org/10.1016/j.cub.2004.01.005

Tanevska, A., Rea, F., Sandini, G., Canamero, L., & Sciutti, A. (2019). A Cognitive Architecture for Socially Adaptable Robots. *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 195–200. https://doi.org/10.1109/DEVLRN.2019.8850688

Tanevska, A., Rea, F., Sandini, G., & Sciutti, A. (2018). Designing an Affective Cognitive Architecture for Human-Humanoid Interaction. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 253–254. https://doi.org/10.1145/3173386.3177035

Török, G., Pomiechowska, B., Csibra, G., & Sebanz, N. (2019). Rationality in Joint Action: Maximizing Coefficiency in Coordination. *Psychological Science*, *30*(6), 930–941. https://doi.org/10.1177/0956797619842550

Vaish, A. (2018). The prosocial functions of early social emotions: The case of guilt. *Current Opinion in Psychology*, *20*, 25–29. https://doi.org/10.1016/j.copsyc.2017.08.008

Vandemeulebroucke, T., Casterlé, B. D. de, & Gastmans, C. (2018). How do older adults

experience and perceive socially assistive robots in aged care: A systematic review of qualitative evidence. *Aging & Mental Health*, *22*(2), 149–167. https://doi.org/10.1080/13607863.2017.1286455

Vesper, C., Abramova, E., Bütepage, J., Ciardo, F., Crossey, B., Effenberg, A., Hristova, D., Karlinsky, A., McEllin, L., Nijssen, S. R. R., Schmitz, L., & Wahn, B. (2017). Joint Action: Mental Representations, Shared Information and General Mechanisms for Coordinating with Others. *Frontiers in Psychology*, *07*. https://doi.org/10.3389/fpsyg.2016.02039

Vesper, C., & Richardson, M. J. (2014). Strategic communication and behavioral coupling in asymmetric joint action. *Experimental Brain Research*, *232*(9), 2945–2956. https://doi.org/10.1007/s00221-014-3982-1

Vignolo, A., Sciutti, A., Rea, F., & Michael, J. (2019, November 26). *Effort reciprocation in child-robot interaction*. The Communication Challenges in Joint Action for Human-Robot Interaction (ICSR2019), Madrid.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Harvard University Press.

Wang, S., Lilienfeld, S. O., & Rochat, P. (2015). The Uncanny Valley: Existence and Explanations. *Review of General Psychology*, *19*(4), 393–407. https://doi.org/10.1037/gpr0000056

Wendt, C., Popp, M., Karg, M., & Kuhnlenz, K. (2008). Physiology and HRI: Recognition of over- and underchallenge. *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 448–452. https://doi.org/10.1109/ROMAN.2008.4600707

Wendt, C. S., & Berg, G. (2009). Nonverbal humor as a new dimension of HRI. *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 183–188. https://doi.org/10.1109/ROMAN.2009.5326230

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots As Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social. *Frontiers in Psychology*,

8. https://doi.org/10.3389/fpsyg.2017.01663

Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., Jacobstein, N., Kumar, V., McNutt, M., Merrifield, R., Nelson, B. J., Scassellati, B., Taddeo, M., Taylor, R., Veloso, M., Wang, Z. L., & Wood, R. (2018). The grand challenges of *Science Robotics*. *Science Robotics*, *3*(14), eaar7650. https://doi.org/10.1126/scirobotics.aar7650

Zawidzki, T. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.