Joint actions, commitments and the need to belong

Víctor Fernández Castro, Elisabeth Pacherie

Draft version. For purposes of quotation please consult the published version:
Fernández Castro, V. & Pacherie, E. (2021) Joint actions, commitments and the need to belong. *Synthese*, 198, 8: 7597–762. doi:10.1007/s11229-020-02535-0

Joint Actions, Commitments and the Need to Belong

Víctor Fernández Castro[12] & Elisabeth Pacherie[1]

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

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

Correspondance to:

Víctor Fernández Castro

E-mail: vfernandezcastro@gmail.com

Acknowledgements: We would like to thank John Michael and two anonymous referees for their valuable comments and suggestions on earlier drafts of this paper. We would also like to thank the participants to the workshop "Layers of Collective Intentionality" held in Vienna in August 2018, the participants to the workshop "Human-Robot Joint Action: Refining the understanding of joint action through an interdisciplinary perspective" held in Paris in September 2018, and the members of the Philosophy Department at the University of Granada.

This research was supported by the Agence Nationale de la Recherche [grant number ANR-16-CE33-0017] and by the EUR Frontiers in Cognition [grant number ANR-17-EURE-0017].

Joint Actions, Commitments and the Need to Belong

Abstract

This paper concerns the credibility problem for commitments. Commitments play an important role in cooperative human interactions and can dramatically improve the performance of joint actions by stabilizing expectations, reducing the uncertainty of the interaction, providing reasons to cooperate or improving action coordination. However, commitments can only serve these functions if they are credible in the first place. What is it then that insures the credibility of commitments? To answer this question, we need to provide an account of what motivates us to abide by our commitments.

We first discuss two conceptions of the nature of the commitments present in joint action and of the norms that govern them. We contend that while normative considerations may have some motivational force, there are reasons to doubt that they, by themselves, could provide a sufficient motivational basis to fully explain why agents abide by their commitments and thus why their commitments are credible. In the next two sections, we discuss two proposals regarding further sources of motivation, reputation management and social emotions. We argue that while reputation management and social emotions certainly play a role in motivating us to act as committed, there are both theoretical and empirical reasons to think that neither captures the most basic motivational force at work in sustaining commitments. We propose instead that the need to belong, i.e., the need to affiliate with others and form long-lasting bonds with them, is what primarily motivates us to interact and engage with those around us and act so as to preserve and reinforce the bonds we have forged with them. We argue that the need to belong is a more basic proximate motivation for conforming to commitments, in the sense both that affiliative behaviors are evidenced much earlier in human development than either reputation management or social emotions and that the need to belong is at least part of an explanation of why we care for our reputation and why we care about others' assessments of our behavior.

Keywords: joint action; commitment; credibility; practical rationality; social normativity; reputation; social emotions; need to belong.

1. Introduction

Humans spend a significant amount of their time engaged in social situations carrying out cooperative projects and interacting with each other. Not surprisingly then, an increasing body of literature in philosophy of mind and psychology is devoted to a notion that encompasses an important number of these social encounters, namely, *joint action* (Butterfill and Sebanz 2011; Bratman, 1992; 2009a; Brownell, 2011; Gilbert, 1992; Pacherie, 2011; Sebanz et al. 2006; Tollefsen, 2005; Tomasello and Rakoczy, 2003; Vesper et al. 2010). In its widest sense, the notion of joint action refers to any form of social interaction where two or more individuals coordinate their actions in pursuit of a common goal. While certain forms of joint action are also observed in other animals, humans seem to have a higher degree of flexibility and proficiency in performing these collective behaviors.

This social flexibility and proficiency at acting jointly would not be possible without a special set of skills and cognitive mechanisms to deal with cooperative interactions and solve the specific difficulties raised by inter-agent coordination. Among this myriad of mechanisms involved in joint actions, *commitments* appear to play a key role. Commitments can dramatically improve the performance of joint actions by stabilizing expectations, reducing the uncertainty of the interaction, providing reasons to cooperate or improving action coordination. This power of commitments for facilitating social interactions lies in their reliability or credibility (Michael and Pacherie, 2015). A commitment can only reduce uncertainty, stabilize expectations or provide reasons for acting jointly if it is credible. However, the credibility of commitments proves not to be a straightforward matter, especially when one notes that credibility depends upon the motivation of the committed agent to honor her commitment in

many situations where alternative options that maximize her interests are available and conflicting motivations are present (Michael and Pacherie, 2015: 101). Let us call this challenge *the credibility problem*.

What is needed to answer this challenge is an account of what motivates agents to abide by their commitments in the first place. Classical philosophical accounts of joint actions and of the commitments they involve have failed to clearly confront this issue. They have offered normative considerations why people should act as committed but have had very little to say regarding the psychological mechanisms that may explain why people are actually motivated to act as they should. In the last decades, however, there has been growing interest in the phylogenetic and ontogenetic roots of human prosociality in an array of disciplines spanning biology, psychology and the social sciences. These investigations have yielded a wealth of proposals regarding the evolutionary origins of human cooperativeness and the proximal psychological mechanisms that sustain it. In recent years, philosophers have proposed more empirically informed answers to the credibility problem, appealing to psychological mechanisms such as reputation management or social emotions. While we agree that both reputation management and social emotions have a role to play in motivating conformity to commitments, we do not think either constitute the most basic motivation at work in sustaining commitments. We propose instead that the need to belong, i.e., the need to affiliate with others and form long-lasting bonds with them, is what primarily motivates us to interact and engage with those around us and act so as to preserve and reinforce the bonds we have forged with them. We argue that the need to belong is a more basic proximate motivation for conforming to commitments, in the sense both that affiliative behaviors are evidenced much earlier in human development than either reputation management or social emotions and that the need to belong is at least part of an explanation of why we care for our reputation and why we care about others' assessments of our behavior.

Our focus will be on the developmental emergence of commitments and the proximal mechanisms they engage. This strategy of focusing on development can help us to gain insight into the cognitive structures behind our proficiency in tracking and responding to commitments and the motivational mechanisms underpinning the tendency to honor the expectations generated by commitments. Accounting for the development of such competencies can improve our understanding of how our expectations about others' contributions to shared goals become reliable. If, as it seems, the human capacity to engage in joint action involving commitments is an uncommon form of sociality in the animal kingdom, paying attention to the ontogenetic paths of such a socio-cognitive capacity seems to be a worthwhile methodological procedure for discerning how humans tackle the problem of credibility in joint actions.

This paper is organized as follows. In section 2, we characterize commitments and their role in joint action and introduce the credibility problem for commitments. In section 3, we delineate two conceptions of the nature of the commitments present in joint action and of the norms that govern them – inspired respectively by Michael Bratman's and by Margaret Gilbert's accounts of shared intentions. We consider ways in which one might try to connect up normative reasons for compliance with commitments with motivation to act as committed. We contend that such normatively derived motivations do not provide a sufficient motivational basis to fully explain why agents abide by their commitments and thus why their commitments are credible. In sections 4 and 5, we discuss two proposals regarding further sources of motivation, the first appealing to reputation and reputation management and the second to social emotions. We argue that while reputation management and social emotions certainly play a role in motivating us to act as committed, there are both theoretical and empirical reasons to think that neither captures our most basic motivation to engage in share intentional action and follow through on our commitments. In section 6, we develop our own proposal that the need to belong constitutes a more basic motivational mechanism than either reputation or social

emotions. In particular, we argue that our proposal fits developmental data about the emergence of commitments better than either of these approaches and that the need to belong may contribute to explaining the effectiveness of these later developing motivational mechanisms. We also discuss how our proposal relates to another proposal, the social motivation hypothesis, recently put forward by Godman and colleagues as a basic explanation of the appeal of prosocial behavior (Godman 2013; Godman et al. 2014). In section 7, we discuss possible objections and contrast our view with a recent proposal by Michael and Székely (2018) on the developmental origins of the sense of commitment.

2. What are commitments and what are they for?

In the philosophical literature, the notion of commitments has been closely associated with the notion of joint action (Bratman, 2009a; Gilbert, 1992; Roth, 2004). For instance, Gilbert (1997: 13) claims that shared intentions essentially involve joint commitments, thus putting joint commitments at the very heart of her theory of joint action. While Bratman (2009a) denies that joint actions necessary involve joint commitments in Gilbert's sense, he nevertheless claims that joint actions require that each participating agent be committed to acting jointly with others, that is, at least, committed to the mutual compatibility of their relevant sub-plans and committed to help others fulfill their role if needed (mutual support). To this philosophical literature, we can add a growing body of empirical research in developmental and cognitive psychology exploring the relation between joint actions and commitments and normative understanding (Gräfenhain, et al., 2009; Carpenter, 2009; Rakoczy 2006; Rakoczy, et al, 2008; Siposova et al. 2018; Tomasello and Carpenter 2007; Tomasello and Rakoczy 2007, Székely & Michael 2018); for instance, the role that commitments and their verbal and gestural elicitation could play in children developmental capacities to cooperate has been investigated (Siposova, et al. 2018).

But what is a commitment in the first place and how is it established? On a standard philosophical conception, a commitment in the strict sense is, as Michael and Salice (2017) put it, "a triadic relation among two agents and an action, where one of the agents is obligated to perform the action as a result of having given an assurance to the other agent[that she would do so, and of the other agent's having acknowledged that assurance under conditions of common knowledge" (2017: 756).[1] To give an example, if Sarah promises to help Andrew repair his bike, she will feel obligated to help him on the basis of her promise and of Andrew's acknowledgement of her promise, under the condition where both recognize that the other knows about the intention of Sarah to help to repair the bike. 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 by intentionally communicating that one intends to x through a promise or other speech acts (Austin, 1975; 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 (Ledyard, 1995; Sally, 1995; Scalon, 1998; Siposova et al. 2018).

In addition, as Michael et al. 2016 (see also Michael and Salice, 2016; Lo Presti, 2013) have claimed, in certain conditions an agent might experience a sense of commitment even in the absence of verbal or non-verbal communication. Situational affordances or other contextual factors, for example, can make an agent experience a sense of commitment that puts pressure on her to act correspondingly. For instance, we may feel committed to push the open button of the elevator when we see someone trying to get in and the doors are closing. More generally,

[1] [1] Note that, as a limiting case, the two agents can be one and the same. For instance, when Bratman (1987) argues a future-directed intention involves a characteristic commitment to future action, the agent who forms the future-directed intention is both the author and the recipient of the commitment it involves.

one may feel committed to contribute to another agent's goal simply by identifying this goal and realizing that the contribution of another agent is crucial to their achieving this goal (Michael, Sebanz & Knoblich 2016, Michael and Székely 2018) Similarly, the mere repetition of patterns of interactions or the perception that one's partner is investing effort may generate a sense of commitment (Székely and Michael 2018).

Michael and Székely (2018) further argue that the sense of commitment is a broader and less complex phenomenon than commitments in the strict sense and that this is reflected in the developmental timeline through which children progressively gain proficiency with commitments. According to them, children do not acquire proficiency with commitments by first acquiring the concept of commitments in the strict sense and then exhibiting a suite of behaviors licensed by the concept. Rather, they first acquire a sense of commitment that is then "gradually calibrated through social experience to give rise to a mature proficiency in managing commitments" (Michael & Székely, 2018: 112). In their view then, in childhood the first step towards the emergence of an understanding of commitments involves the development of a sense of commitment. Suffice it to say for the present that while we have some reservations with some aspects of their account, which we will discuss in section 7, we fully agree with them that children's sensitivity to commitments predates their mastery of the concept of commitment.

As Michael and Pacherie (2015) and Michael and Salice (2017) have argued, commitments play an important role in cooperative human interactions at large and joint actions in particular in that they make agents' actions predictable in the face of fluctuations in their desires and interests. As a result, they may enable agents to have more reliable expectations about each other's actions than would otherwise be possible, thus facilitating cooperation and coordination. In particular, having reliable expectations about others'

contributions to shared goals may facilitate the planning of joint actions with mutually interdependent sub-plans and facilitate in turn the online coordination of co-agents.

However, these benefits only accrue if commitments are credible in the first place, that is, if the authors of commitments do more often than not act in accordance with their commitments and if their recipients trust the authors to act as committed. The credibility problem is the problem of explaining what motivates agents to abide by their commitments. The credibility of commitments is not a straightforward matter. On the one hand, it may be irrational to engage in and follow through on commitments, to the extent that they foreclose options which may arise and which may be more attractive than the action to which one is committed. This is vividly illustrated by this example that Frank borrows from Schelling (1960): "A kidnapper who suddenly gets cold feet […] wants to set his victim free, but is afraid he will go to the police. In return for his freedom, the victim gladly promises not to do so. The problem, however, is that both realize it will no longer be in the victim's interest to keep this promise once he is free. And so, the kidnapper reluctantly concludes that he must kill him" (Frank, 1988: 4). The impending tragic outcome results from the lack of credibility of the victim's commitment to keep their mouth closed. On the other hand, human agents are prone to act irrationally. Thus, even in cases where it would be in an agent's best interests to abide by their commitments, the agent may be led astray by momentary temptation. Thus, rationality may not always require us to abide by our commitments and, even when it does, the motivational force of rational considerations may not by itself be sufficient to counteract non-rational motivations that pull in the opposite direction. To offer a solution to the credibility problem is to explain what motivates us to act as committed. This involves identifying the motivational forces that, sometimes together with rational assessments but sometimes also against them, lead us to act as committed. Of course the particular motivational explanation we may give for why we acted as committed in a given situation may differ in part from the

explanation we would give for another situation. Our hope, though, is to show that the Need to Belong is a major source of motivation at work, either directly or indirectly, in making commitments credible. While it goes beyond the scope of the present paper to discuss group agents and corporate agency, we also think that the Need to Belong has an important role to play in explaining why individuals are keen to join social groups in the first place, why they are ready to pay sometimes hefty personal costs to be accepted by a group and show their loyalty to it.

Before we turn to the Need to Belong hypothesis, let us consider first other approaches to the credibility problem.

3. Normative approaches to commitments and the credibility problem

In the philosophical literature on joint action, two broad conceptions of the nature of the norms that govern commitments in shared intention can be discerned: one approach appeals to considerations of practical rationality, while the other appeals to the social normativity of commitments. These two approaches are perhaps best exemplified by the accounts of Michael Bratman and Margaret Gilbert, respectively. As such neither theory directly tackles the credibility problem, Rather, their main aim is to characterize the norms at play when people form shared intentions and their role in explaining social coordination. In so doing, they explain why, normatively speaking, people should comply with their commitments. What is needed, however, to solve the credibility problem is an account what actually motivates us to accept these norms and thus abide by our commitments rather than an account of why we should do so. We delineate Bratman's and Gilbert's respective accounts and consider ways in which one might try and connect up normative reasons to act with motivation to act in order to derive

solutions to the credibility problem from these normative accounts.[2] We argue that such attempts are ultimately unsuccessful as they cannot provide a sufficient motivational basis to fully explain why agents abide by their commitments and thus why their commitments are credible.

3.1 Bratman on commitments

On Bratman's theory of planning agency (Bratman, 1987, 2014), intentions are distinctive elements of human planning agency that go beyond the ordinary desires and beliefs characteristic of simple purposive agency. In particular, future-directed intention involves a characteristic commitment to future action and it is this feature of intentions that allows us to become temporally extended agents as well as social agents. Having a capacity for intention frees us from the confines of the present, allowing us to coordinate our present self with our future selves, while at the same time freeing us from the confines of our own self and allowing us to coordinate with others. By thus extending our agency, intentions contribute in the long run to our securing greater desire-satisfaction than simple purposive agency would. In order to accrue these benefits, however, intentions must be subject to norms of practical rationality. As Bratman insists, "Primary among these norms are norms of consistency, agglomeration, means-end coherence, and stability: intentions are to internally consistent, and consistent with one's beliefs; and it should be possible to agglomerate one's various intentions into a larger intention that is consistent in these ways." (2014: 15). The norm of stability concerns the reconsideration of intentions already formed: they are rationally required to resist reconsideration and be stable, as their instability would defeat the very purpose of planning agency.

[2] Note that we are not suggesting that this is a project Bratman or Gilbert themselves are engaged in or would condone, only that it is a possibility one could in principle wish to explore.

In his book *Shared Agency* (2014), Bratman defends a continuity thesis, arguing that the step from individual planning agency to shared agency need not involve fundamentally new conceptual, metaphysical, or normative elements. He develops a constructivist approach to shared intentions that exploits the conceptual and normative resources of his planning theory of individual agency. As he puts it, "the idea, roughly, is that the social-norm-assessable social functioning characteristic of shared intention emerges from the individual-norm-assessable and individual-norm-guided functioning of relevant structures of interrelated intentions of the individuals, as those intentions of individuals are understood by the planning theory" (2014: 32) His "basic thesis" is that one can capture the interconnections among agents characteristic of shared agency by construing shared intentions as complexes of interlocking and interdependent intentions and other attitudes of individual agents. Bratman also argues that the social normativity characteristic of shared agency derives from the normativity already associated with individual planning agency. Intentions of individual participants, when they are interconnected in the way specified by the basic thesis, will normally, in responding to these norms of individual practical rationality, lead to the emergence of corresponding norms of social consistency, social agglomeration, social coherence and social stability. Finally he holds that this structure of interrelated intentions will normally support and guide planning and shared deliberation, but also emphasizes the importance for such shared deliberations of "shared commitments to weights", that is shared commitments concerning what to treat as mattering in our deliberations and planning, the presence of which "distinguishes shared deliberation from ordinary bargaining" (2014: 133).

On Bratman's view, to the extent that an agent is practically rational, her intentions, whether personal or shared, are subject to a norm of stability. The agent has committed to act in a certain way and is rationally required to act as committed. To try and answer the credibility problem, we must find a way to connect up normative reasons for abiding by one's

commitments with motivation to do so. While Bratman does not provide a theory of what motivates us to comply with our commitments, his view of the normative force or significance of norms of intention rationality suggests ways in which we might try to build a connection to motivation. Bratman contends that norms of intention rationality have both instrumental and non-instrumental normative significance. With regards to the instrumental significance of these norms, Bratman's idea is that "being guided by one's acceptance of these norms is an important element in how [the characteristic coordinating, organizing and settling roles of planning] are normally realized, and that it is important to us that these roles indeed be realized" (2014: 17). In other words, to the extent that we care about our planning agency and the benefits it yields, you should care about these norms as adherence to them makes effective agency more likely. In addition, Bratman proposes that these norms also have non-instrumental significance. We value them not just as means towards other things but also in themselves insofar as they are constitutive of self-governance and self-governance is something we care intrinsically about (Bratman 2009b).

This suggests two ways we might try and connect up normative reasons and motivation. The first sees consistency, coherence and stability as tools for effectively pursuing our intended ends and considers that to the extent that we are motivated to achieve certain ends we will normally be motivated to comply with norms of consistency, coherence and stability as means for achieving these ends. The second is that insofar as agents are motivated to govern their own life, they will be motivated to comply with these norms since such compliance is constitutive of self-governance. The question then is whether theories of motivation built along such lines would be sufficiently robust to yield a solution to the credibility problem.

Consider first the instrumental approach to motivation and the norm that is most directly relevant to the credibility problem, namely stability. Would our (instrumental) motivation to

comply with this norm be sufficiently strong to yield sufficiently robust solution to the credibility problem? There are at least two types of reasons to be doubtful. First, as Bratman himself points out, practical rationality does not demand that we never reconsider once we have formed an intention, but rather that we do not reconsider unless we have valid reasons to do so. In other words, practical rationality requires that we carefully navigate the, sometimes, narrow straits between pusillanimity and foolish stubbornness. The problem, then, is that there is no guarantee that the agent will not modify her intentions if new information comes to light or her interests change. Indeed, practical rationality may demand that she re-consider in certain circumstances. While this possibility may be thought to constitute a minor threat to the stability of intentions in individual agency, the threat may be amplified when we turn to shared agency. According to Bratman, shared intentions are structures of interrelated intentions in favor of shared intentional activity, where there is persistence interdependence between the intentions of the co-agents: "the *persistence* of one's intention that we J supports the *continued persistence* of the other's intention that we J, and vice versa" (Bratman, 2014: 68). At the same time, Bratman emphasizes that "shared intention in favor of shared action need not involve commonality of reasons for participating in the sharing" (2014: 145). Thus, the stability of my intention is at risk not just if I am led to reconsider the reasons I had for engaging in the shared activity, but also, in virtue of the persistence interdependence of our intentions, if any of the other participants to the shared activity is led to reconsider their own reasons for participating in it. In other words, persistence interdependence may create a domino effect and lead to the unraveling of the whole structure of interrelated intentions, without this involving irrationality. Second, even in cases where reconsideration is not rationally warranted, one's motivation to comply with the stability norm may be in competition with motivations pulling in other directions and outweighed by these competing motivations. Both common experience and empirical research tell us that this is, indeed, often the case. For instance, humans often yield

to temptation, are liable to hyperbolic discounting, creating temporary preference reversals, and display a host of further rationality failings (e.g., Kahneman & Egan 2011). The second route to a motivation theory, building on the constitutive relation between self-governance and adherence to norms of intention rationality, gives rise to similar misgivings. First, as Bratman acknowledges (2009b: 443), agents may not all care intrinsically about governing their own lives, and, second, those who so care may also have competing motivations.

To solve the credibility problem, it is not enough to simply claim that normative reasons can motivate us to act. Rather, a much stronger claim would have to be made, namely that the motivation associated with normative reasons is reliably stronger than other competing motivations. On the face of it, this is an implausibility strong claim and certainly a claim that Bratman does not explicitly endorse. Where does that leave us? What Bratman has to offer is a theory of why agents should, in normal circumstances, comply with their commitments. It appears reasonable to demand that normative reasons for action be able to connect up with motivations of action and Bratman's reflections on the normative force of norms of intention rationality suggest ways of building such connections. However, what we need to solve the credibility problem is a robust theory of what actually motivates agents to comply with their commitments and such a theory will have to appeal to more than just these normatively derived motivations.

3.2. Gilbert on commitments

In contrast to Bratman, Gilbert (2009) takes it that there is a deep conceptual, metaphysical and normative discontinuity between individual and social agency. Her account of shared intentions essentially involves the notion of joint commitments:

> Persons X, Y, and whatever particular others share an intention to do A if and only if X, Y, and these particular others are jointly committed to intend as a body to do A. (Gilbert 2009: 179)

Importantly, she insists that joint commitments are not concatenations of personal commitments. Rather, the author of a joint commitment comprises those who have jointly committed themselves by their concordant expressions. Together they constitute the plural subject of the commitment. In forming a joint commitment, the parties to the commitment together impose obligations on each other to act in conformity with the commitment, and concomitant rights to demand of one another that they so act. In addition, since a joint commitment can only be rescinded with the consent of all the parties involved (the plural subject), absent such consent, agents remain obligated to act in conformity with the shared intention even if their interests have changed and they do no longer have matching personal intentions. For Gilbert, the idea of a joint commitment is a primitive social notion that does not admit of further reductive analysis. Similarly, the obligations and entitlements a joint commitment grounds cannot be understood as moral in kind or as emerging from the norms associated with individual planning agency. Rather, they engage a *sui generis* kind of social normativity.

According to Gilbert, an appeal to joint commitments and to the normative force of the obligations and entitlements they generate provides, in comparison to Bratman's appeal to the practical rationality of individual agents, "a more stable framework for bargaining and negotiation and, relatedly, a more felicitous means of coordinating the personal intentions of individuals, and keeping them on the track of the shared intention" (2009: 185).

On Gilbert's approach, the normative reason why people should abide by their commitments is that joint commitments give rise to obligations and thus that they have an

obligation to act as committed. Although her view of the norms central to commitments is quite different from Bratman's, her theory, like Bratman's, is a theory of why agents should abide by their commitments and not a theory of what actually motivates them to act as they should. To provide a solution to the credibility problem germane to Gilbert's stance on the social normativity of commitments, one would need a theory of motivation that explains why people are motivated to act as their obligations dictate. Again, while it appears reasonable to demand that normative reasons for action be able to connect up with motivations of action, it would be unreasonable to insist that the connection is so tight that the motivation supplied by the recognition of obligations is reliably stronger than other competing motivations. Since Gilbert insists that the mutual obligations inherent in joints commitments can remain in force even in the absence of correlative personal intentions, the motivation to conform to these obligations would have to be strong enough to counteract not only competing 'non-rational' motivations but also motivations associated with other types of normative reasons (e.g., the reasons that led the agent to reconsider and give up his correlative personal intentions). There is no reason to think that Gilbert would draw such a tight connection between social normative reasons and motivation, but then this means that she has no answer to the question what motivates us to act as our obligations dictate. Thus, it seems that we can no more derive a solution to the credibility problem from her view than from Bratman's.

To recap, while we do not want to deny that normative considerations, whether linked to practical rationality demands or to social normativity, have motivational force, we think there are reasons to doubt that they, by themselves, could provide a sufficient motivational basis to fully explain why agents abide by their commitments and thus why their commitments are credible. In addition to normatively derived motivations, we need to appeal to further sources of motivation to block forms of practical irrationality or to counteract selfish motives that would otherwise threaten the credibility of our commitments.

In the next two sections, we review two important proposals that highlight further sources of motivations to abide by our commitments. The first proposal, in line with the practical rationality approach to commitments, appeals to the idea of reputation and reputation management as a further source of motivation. The second attempts to explain why we are motivated to act as obligated by appealing to the motivational force of social emotions.

4. Reputation

How can the practical rationality approach to credibility be reinforced? Why would an agent remain motivated to contribute to a joint action and to honor her commitment when her interests have changed and her personal reasons against contributing to *the joint action* are now stronger than her reasons in favor of contributing to *it*? One important answer is that agents care for their reputation and that their concern for their reputation may counterbalance their inclination to renege on their commitments.

Several evolutionary theories of human cooperation have proposed that reputation is a key mechanism in stabilizing cooperation and insuring that people cooperate and honor their commitments in situation where they may be tempted not to. Theories of indirect reciprocity (Alexander, 1987; Nowak & Sigmund, 2005), of competitive altruism (Barclay & Willer, 2007; Van Vugt, Roberts, & Hardy, 2007) and costly signaling (Zahavi & Zahavi, 1997) all concur on the idea that reputation facilitate cooperation. They suggest that people cooperate to maintain a good reputation in their social environment, where this reputation in turn attracts valuable partners and allies, thus positively affecting their future benefits. Like Bratman's theory of planning agency, theories of reputation-based cooperation emphasize the temporally-extended dimension of human agency, but they appear to further expand the shadow of the future by putting into the equation not just current plans for the future and their impact on

intrapersonal and interpersonal coordination, but also potential future interactions with potential future partners.

This approach then offers an answer to the credibility problem in terms of reputation management. Even if an agent has lost his initial motivation to engage in joint action, he may still care about his reputation as a cooperator and trustworthy partner. Michael et al. (2016) conjecture that "a tendency to be motivated to fulfill others' expectations about one's contributions to their goals or to outcomes which they desire (i.e., a preference for expectation fulfillment) has the status of a default in humans" (p. 6) and that this default tendency to fulfill expectations may have evolved as a mechanism for reputation management (see Heintz, et al., 2016). That is, we tend to act in conformity with our explicit or implicit commitments, at least in part, in order to maintain our reputation.

Although the importance of reputation in the stabilization and replication of cooperative and joint action is undeniable (e.g. Trivers 1971; Alexander 1987; Haley & Fessler 2005; Nowak & Sigmund 2005), it is less clear that it constitutes our primary motivation for honoring our commitments in the sense relevant to this paper. As we have noted in Section 1, our central concern revolves around identifying our most basic motivations for conforming to commitments, that is, motivations that are evidenced early in ontogeny and that can contribute to scaffold later emerging, more complex motivational mechanisms. In this sense, one may object, children acquire a sense of commitment and at least some understanding of how implicit commitments work much before they develop a capacity for reputation management. In a set of studies, Gräfenhain et al. (2009) investigated children's emerging understanding of commitments. In their first study, they tested whether children reacted differently when an experimenter with whom they were engaged in a simple joint action abruptly disengaged, depending on whether the experimenter had made an explicit commitment to the joint action

or simply entered into the action without making any commitment. Their found that 3-year-olds, but not 2-year-olds, protested significantly more when a commitment had been violated than when there had been no commitment. In their second study with 3- and 4-year-old children, they tested the children's understanding of their own obligation to a committed joint activity. They found that when they were enticed away from a joint activity with an adult, children in both age groups acknowledged their leaving significantly more often when they had made a joint commitment to act together than when they had not. From these results, Gräfenhain et al. (2009) concluded that by the age of three children have acquired an understanding of the nature of commitments in joint activity and of the obligations they carry for themselves and for their partners. Interestingly, Michael and Székely (2018) propose an alternative explanation for the findings of Gräfenhain et al.'s first study. They point out that in both experimental conditions, the 2-year-olds reacted to the interruption at a level as high as that of the 3-year-olds in the joint commitment. This, they argue, suggests that by age 2 children already have a default sense of entitlement that inspires their protest over an unfulfilled expectation, but that what changes during their third year is that "children learn they are not always entitled to expect contributions to their goals" (Michael & Székely 2018: 111).

These findings then suggest that an understanding of commitments has emerged in children by the age of three or, if we follow Michael and Székely, that a sense of commitment is already in place by two or earlier. This developmental timeline does not sit well with the suggestion that the preference for expectation fulfillment is primarily motivated by concern for one's reputation. Although still scarce, developmental studies on reputation management suggest that children start exhibiting actions aimed at promoting their own reputation by the age of five (see Silver and Shawn, 2018 for a review). For instance, 5-year-old children are more generous when their behavior is perceived by their partner (Leimgruber, et al. 2012). Further, 5- to 6-year-olds avoid cheating when they believe they are observed by another person

(Piazza, Bering, & Ingram, 2011), even when this person is an imaginary character (e.g. Princess Alice). The capacity to manage reputation is not restricted to strategies promoting a positive self-perception. Engelman et al. (2016) found that 5-year-old (but not 3-year-old) children communicate evaluative information to partners (gossip) about a third party's proclivity to cooperate. By the age of 6, children understand explicitly the importance of reputation. As the study of Shawn and Olson (2015) suggest, children dislike plagiarism for reasons regarding reputation. For instance, children do not consider that a girl who is receiving positive feedback from someone else's story is doing something wrong as far as she gives credit to the source, and thus, improves the source's reputation. Thus, children start exhibiting sensitivity to the importance of reputation around the age of 5, whereas their understanding of commitments emerges at least two years earlier. This developmental trajectory casts doubt on the idea that reputation management is our primary motivation for complying with our commitments.

Apart from empirical considerations, the reputation view also faces some theoretical difficulties. First, the approach is afflicted by a version of the so-called open-question argument (see Moore, 1903; Strandberg, 2004). According to this argument, equating motivation to fulfill commitments with our motivation to prompt our own reputation is uninformative, since it simply replaces the question of why we care about honoring our commitments with the question of why we care about our own reputation. To put it differently, we need a motivational explanation for why we are being moved to promote others' positive evaluations of our actions in the same way that we need a motivational explanation for why we tend to fulfill our commitments. Thus, appealing to reputation management just seems to take us one step back rather than solve the problem. A plausible reply could be that motivations for reputation management are irrelevant because reputation improves evolutionary fitness *per se*. However, given that reputation management is not an inborn capacity, an account of the developmental

emergence of the competence is still necessary. Second, it is questionable whether reputation management can cover all relevant cases where we engage in joint and cooperative actions. For engagement in such actions to increase our reputation and evolutionary fitness, collaborators, witnesses of the action or people they are acquainted with must be potential cooperators of subsequent interactions. However, as economic findings with one-shot public good games demonstrate, humans are ready to cooperate even when they do not have reasons to expect further interactions (see Ledyard 1995, Chaudhuri 2011 for reviews). Thus, humans engage in cooperative actions even when there is no incentive to increase their reputation. The scenarios where we behave altruistic or cooperatively without expecting further interactions are not restricted to experimental situations. We often behave altruistically with strangers in everyday interactions; for instance, when we comply with conventional norms in situations involving complete strangers such as giving up our seat to older people on public transportation in a foreign city.

In a nutshell, although reputation management may sometimes explain or contribute to explaining why we honor our commitments, it does not seem to constitute a plausible general solution to the credibility problem. Empirical findings suggest that human understanding of implicit commitments appears before the capacity for reputation management. Further, there are theoretical considerations that jeopardize the idea that reputation management is the most basic motivation at work in sustaining commitments. After all, even if we are moved to comply with our commitments to improve our reputation, one could always wonder why our reputation matters to us anyway.

5. Social Emotions

Like the practical rationality account, the deontic account demands an explanation of what motivates people to act as they are obligated to. Why would an agent feel compelled to act as

her obligations dictate, to contribute her part to a joint action or to comply with their commitments? A plausible move to complement Gilbert's account could appeal to emotions as such a driving force. On this approach, our tendency to fulfill our obligations would be a result of our inclination to avoid negative emotions and seek positive ones. For instance, authors of commitments would tend to satisfy their obligations and meet the expectations of their partners because not doing so can give rise to negative emotions such as embarrassment, guilt, fear or aversion. An appeal to emotions to solve the credibility problem of commitments doesn't seem far-fetched, given that emotions have already been claimed to support joint action in a variety of ways. For instance, the avoidance of, or preference for, certain emotions might serve to control selfish impulses (Vaish, 2018). It has also been proposed that emotions can function as both motivating and justifying reasons for joint action per se (Salmela & Nagatsu, 2016). In addition, emotional responses can work as coordination smoothers (Michael, 2011) or inform others about our inner states, giving them reliable clues about how to interact with us (Frank, 1988; Zahn-Waxler et al, 1992). As Frank (1988) emphasizes: "Being known to experience certain emotions enables us to make commitments that would otherwise not be credible" (p. 5). Similarly, then, it is reasonable to assume that avoidance of negative emotions or search of positive emotions can be what motivates us to fulfill our obligations towards our co-agents and comply with our commitments.

What kinds of emotions can play such a role? Social emotions, as a specific subset of emotions including guilt, embarrassment, or pride, can be seen as plausible candidates (see Tomasello, 2009; Chang et al., 2011). While all emotions can be affected by social factors and thus are social in some sense, what make social emotions social in a special sense is, according to Hareli and Parkinson (2008), that "they necessarily depend on other people's thoughts, feelings or actions, as experienced, recalled, anticipated or imagined at first hand, or instantiated in more generalized consideration of social norms or conventions." (2008: 131).

Social emotions are the motivational source of other types of actions in social situations, such as reparative and regulative behavior when one causes a harm to another agent (Baumeister et al., 1994; De Hooge et al., 2007; Ketelaar & Au, 2003; Nelissen et al., 2009). Likewise, one could hypothesize that avoiding negative social emotions or seeking to experience positive ones could prompt compliance with commitments. Avoiding feeling guilty or seeking to feel pride may prompt one to behave pro-socially and to act in conformity with one's commitments. For example, someone could act in compliance with her commitment to attend her best friend's birthday party despite having a lot of work because she would feel guilty otherwise.

Social emotions can certainly play a role in motivating us to act as committed. However, as was the case with reputation, the idea that social emotions could be the core motivation behind compliance with commitments is hard to sustain in the light of the developmental findings available. Vaish et al. (2016) have recently suggested that guilt is an early form of social emotion that emerges by three years of age as a way of repairing social bonds when harm is inflicted on others. In their studies, the experimenters tested the children in four conditions, varying whether or not a mishap caused harm to someone and whether children themselves caused that mishap or not. They found that 2-year-olds exhibited less reparative behavior in general and that, although they repaired more in the harm conditions, they did so irrespective of whether or not they were themselves the cause of the harm, pointing to sympathy, rather than guilt, as a plausible motivation for their reparative actions. In contrast, 3-year-olds exhibited more reparative behavior when they were the causal agent of the harm, suggesting that guilt as a social emotion motivating pro-social behavior emerges around age 3. The problem is, then, that according to Michael and Székely's interpretation of Gräfenhain et al's (2009) studies, indicators of an implicit understanding of commitment are already present around age 2. Thus, developmental findings suggest that an understanding of commitments is manifested before children exhibit social emotions. Things get worse, however, since in

contrast to the guilt case investigated by Vaish, where children are in a position to experience at first hand their partner's reaction to the mishap, the social emotions we need to appeal to in order to explain adherence to commitments typically depend on a capacity to anticipate or imagine what the thoughts, feelings and actions of our partners would be, were we to renege on them. However, developmental evidence suggests that the ability to imagine situations in which social emotions might be experienced does not appear until around seven years of age (Harris et al. 1987).

There are further reasons to forego social emotions as basic motivators of compliance with commitments. Social emotions are sophisticated abilities that require the previous acquisition of other cognitive capacities, including capacities for self-representation and mind-reading. As Hareli & Parkinson's characterization of social emotions suggests, to experience moral emotions, one must be capable of imagining other people's mental states, including their assessment of our own behavior. To put it more colloquially, one must be able to see oneself through the eyes of others. The development of these capacities relies on an extensive history of social interactions that, in principle, might require the same kind of motivation that our preference for honoring commitments requires. To see why, notice that social emotions like guilt or shame seem to require the capacity to make self-evaluations (Lewis et al. 1989; Mills, 2005), that is, showing approval or disapproval toward a particular aspect of oneself. Arguably, the capacity of self-recognition and evaluation requires social expertise to deal with social contexts (Rochat et al. 2012). Thus, individual differences in shame expressions seem to correlate with parental evaluative feedback, which indicates that these emotional feelings appear as a byproduct of social regulative behaviors (Mills, et al 2010; Parisette-Sparks, et al 2017). So, the acquisition of social emotions presupposes a history of social interactions, and thus, a motivation to engage in social situations. Further, if as Vaish et al. suggest, guilt's main social function is to repair social bonds when harm is inflicted on others, then one must assume

that children must have developed the capacity to evaluate positively such social bonds and engage in building them before such social emotions arise. The capacities to evaluate social relations and evaluating oneself in accordance to social patterns seems to necessitate a substantial accumulation of social interactions which can hardly take place without previous prosocial preferences and motivations.

Further, it is not clear that an appeal to social emotions can avoid the open-question argument presented against the reputation management view (section 4). As in the case of reputation, postulating social emotions as the motivational underpinning of compliance with commitments seems to just replace the question why do we stick to our commitments with the question of why should we care about what others think or feel about us?[3] The open-question argument presses on the idea that experiencing guilt or embarrassment presuppose other social motivations, for instance, empathic concern or socio-affiliative tendencies. Without assuming that others concern us in some way or another, it is hard to see why one could feel such guilt. As a result, social emotions seem to presuppose a more basic form of social motivation which indeed could explain our tendency to abide by our commitments.

To be clear, we are not denying that social and cooperative behavior can be backed by different psychological devices, including social emotions or reputational management. In fact, as we argue later on, we believe that there is a variety of plausible proximate mechanisms that include, but are not restricted to, emotions and reputational engagement. What we find difficult to accept is that any of these mechanisms is basic enough to account for the emergence of compliance with commitments (see Godman et al., 2014: 577-580). Of course, as Michael & Székely (2018) suggests, one may consider a plurality of mechanisms. However, as we argue

[3] See Godman et al. (2014) for a similar point regarding Robert Sugden (2000)'s resentment hypothesis according to which we are motivated to meet the expectations of others because we are averse to their resentment. As Godman et al.point out, "it raises the question why others' resentment should matter to us anyway" (p. 569).

in section 6, we believe all these mechanisms are scaffolded by a general human need, namely, the need to belong. Hypothesizing such a need, we believe, can explain why children develop certain emotional and non-emotional responses to social behaviors including empathy, negative reactions to anti-social behaviors or normative compliance. Further, we argue that an appeal to the need to belong can explain some specific phenomena that we find in recent studies regarding commitments.

6. The need to belong as a fundamental social motivation

Part of the rationale behind postulating motivational mechanisms to supplement purely normative approaches to commitments is the idea that the recipient of a commitment trusts the author because, in general, people exhibit a tendency to honor their commitments. Our social interactions involving commitments are successful because agents' behavior conforms to the expectations generated by their commitments. So, the explanandum of the theory must be the motivational component that encourages the author to act in accordance with such expectations. Although we believe the reputation management view and the emotions view are problematic as accounts of the fundamental motivation behind commitment compliance, we share their basic rationale that the key to solving the credibility problem is to offer an account of the fundamental motivation that drives us to act as committed. This fundamental motivation is, we propose, the need to belong (Baumeister & Leary, 1995; Over, 2016).

The need to belong is conceptualized as the need 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). Notice that the need to belong is not merely a desire or inclination to interact or cooperate with others or to share their goals. Instead, the need to belong is categorized as a need in order to emphasize its relation to wellbeing manifested in social long-term bonds, so it refers to durable and systematic relations with other

agents. In this sense, our view contrast with other proposals that appeal to closely related motivational factors such as *prosociality*. In a series of recent papers, Godman, Nagatsu and Salmela (Godman 2013; Godman et al. 2014; Salmela & Nagatsu 2016) have proposed what they call the social motivation hypothesis, according to which: "There is a particular psychological disposition whose role is to orient us toward affiliative stimuli, which yields social reward (affect) and enables the formation of social bonds." (Godman et al., 2014: 575). In particular, they argue that agents find acting with others rewarding in its own right and that 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. They also argue, more generally, that the social motivation hypothesis represents a basic explanation of the appeal of pro-social behavior (in terms of anticipated social rewards) and provides a plausible scaffold for other more sophisticated motivations.

The need to belong hypothesis (NTB hypothesis for short) we put forward here can be seen as a more constrained version of the social motivation hypothesis. The NTB hypothesis shares with the motivational hypothesis the predictions that humans tend to give attentional priority to social cues, that they experience social interactions as rewarding, and that they exhibit a preference for promoting actions that maintain and strengthen social relations and engagements (Chevalier et al. 2012; Leary and Allen, 2011). However, the NTB hypothesis, but not the more generic social motivation hypothesis *per se*, also predicts: (1) that people should strive to achieve a certain minimum quantity and quality of social bonds but that, once this level is surpassed, their motivation should diminish; (2) that interactions with a constantly changing sequence of partners will be less satisfactory than repeated interactions with the same persons; and (3) that people should be willing to devote more energy to preserving and consolidating existing bonds than to interacting with strangers and that interactions with strangers should be appealing mainly as potential first-steps towards long-term contact. Thus,

according to the NTB hypothesis, while humans are highly prosocial, their prosociality is neither indiscriminate nor unbounded but rather manifests selectivity.

Although the idea that NTB may be a mechanism underpinning the management of commitment requires empirical confirmation, a number of studies in the developmental literature provide support for the NTB hypothesis rather than the more general hypothesis of pro-social motivation. In particular, the specific predictions of the NTB hypothesis seem to be supported by a range of empirical findings. For instance, while, as early as 8-week-old, infants smile and engage in proto-conversations with caregivers and other agents (Rochat, et al. 1999; Trevarthen & Aitken, 2001), 6 months-old infants strongly prefer to interact with people who engage in contingent interactions with them (Hay et al., 1983, 2004; Jacobson, 1981). Selective preference for imitating, engaging, attending or helping those agents who look warmer and friendlier or prosocial is robust (Hamlin and Wynn, 2011; Hamlin et al, 2007, 2010; Lakin and Chartrand. 2003; Nielsen, 2006; 2009; Over and Carpenter, 2009). For instance, Hamlin and her colleagues have shown that after being presented with scenarios in which a character, attempting to reach the top of a steep hill was alternately pushed up the hill by a "Helper" and pushed down the hill by a "Hinderer", 6- and 10-month-old infants robustly preferred to reach for the helper (Hamlin et al. 2007; Hamlin, 2015). Similarly, Nielsen (2006) found that eighteen-month-old children differed in their copying skills depending on whether the models demonstrating the actions act socially or are aloof. While children focused on copying the outcome of the demonstrated action when the model was aloof, they were as likely to focus on copying actions as outcomes when the model behaved socially.

These findings support a central prediction of the NTB. They show that children do not only interact with caregivers, which could suggest a preference for those who improve their survival, or with people in general, which could suggest a general prosocial preference, but

rather preferentially interact with those who seem ready or more apt to maintain interactions with them. This evidence regarding selective preference over interaction partners speaks, Over (2016) suggests, in favor of the idea that children seek to affiliate with people, preferring those who exhibit behaviors and features that make them more appropriate for maintaining systematic and long terms relations. Such an interpretation is also favored by naturalistic studies demonstrating how preschoolers form stable patterns of friendship involving positive interactions (Howes, 1996; Gifford-Smith & Brownell 2003; Newcomb & Bagwell, 1995). Even during the first year of life, infants exhibit a preference for interacting with unfamiliar peers rather than with unfamiliar adults (Brooks & Lewis, 1976) but they also interact differently with familiar peers than unfamiliar ones (Stefani & Camaioni, 1983; Young & Lewis, 1979). For instance, Stefani and Camaioni observed that after a familiarization period consisting of consecutive meetings of the pairs during three weeks, 8-10 months-old infants exhibited more positive interactions with the peers they interacted with during the familiarization period than with unfamiliar peers who were being raised at home.

Apart from the findings in developmental psychology that support NTB prediction, vis-à-vis the pro-social motivation hypothesis, there is an important source of indirect evidence for the idea that NTB may be a central source of motivation for pro-social and cooperative behavior. In a recent set of experiments, Rusch and Luege (2016) attempted to test the hypothesis that the evolution of cooperation is linked to the cooptation of behavioral strategies evolved to solve problems of coordination to solve problems of cooperation with a greater incentive to defect. In the experiments, subjects played a sequence of Stag Hunt (coordination task) and Prisoner's dilemma games (cooperation task). They devised three types of sequences of 20 games each. The first was composed of 15 Stag Hunt games followed by 5 Prisoner's Dilemma games, the other two were mixed sequences with either a preponderance of Stag Hunt games or a preponderance of Stag Hunt games. The second factor they manipulated was the

matching between the participants, where the participants could play an entire sequence of 20 games with the same partner ("partner-matching protocol") or be matched with new partner after each game ("stranger-matching protocol"). The study found that subjects' cooperation rates were significantly increased compared to baseline when participants played with a fixed partner and Prisoner Dilemma's games were embedded in a sequence of Stag Hunt games but that this effect was absent when players were randomly rematched after each round. In other words, people tend to cooperate more in a context with a big incentive to defect when they have previously engaged in coordination with the same partner. In our view, NTB as a motivation to reinforce our social bonds with certain partners can help us understand why coordination can lead to cooperation in particular circumstances. The fact that agents find interactions with some stable partners rewarding boost more pro-social and cooperative behavior, and thus, it decreases the motivation for defeating[4].

Certainly, evidence that the NTB is a central motivational mechanism in human behavior is not yet proof that it, rather than other motivational forces, plays a pivotal role in complying with commitments. However, even though the claim that NTB plays such a pivotal role has yet to be directly tested, in the absence of direct corroborating evidence, there are some studies that seem to point in that direction. Given the prediction that humans strive to build long-standing relations with others, we can hypothesize that they would be more motivated to conform to commitments in contexts where the others are potential participants in long-

[4] Although Rusch's and Luege's results seem to suggest that there are differences between inclinations to cooperate with partners and to cooperate with strangers, these results are in conflict with other studies involving variables of the same type (see Andreoni and Croson, 2008 for a review). A possible explanation of the contradictory results could be due to the fact that in Rusch and Luege's experiments, the agents are not more cooperative with those they perceive as partners for reasons involving preferences or motivations, but what is being manipulated are the agents' expectations (thanks to an anonymous referee for pointing this out to us). One possible answer, though extremely speculative, might be that even in such a case, the NTB can modulate the force with which the expectations of others affect our decisions in this type of game. However, in the absence of further studies along these lines, we can only indicate that the support this evidence provides for our hypothesis is weak.

standing relations or when social cues that inform of the aptness of the participant are available. This hypothesis is partially confirmed by the everyday observation that repetition can give rise to a sense of commitment (Michael et. al., 2016: 3; Michael & Salice, 2017: 756). To see how, Michael & Salice introduce an example adapted from Gilbert (2006: 9) where two factory workers, Polly and Pam, are in the habit of smoking a cigarette together during their coffee break every day. Intuitively, one might consider that there is an implicit commitment between the two partners to show up at the coffee break. So if Pam, for instance, doesn't show up, one might say that she is violating an implicit commitment and has some obligation to offer an explanation. Further, the sense of obligation seems to increase with time. They can feel less obligated if they have carried out their ritual for a couple of weeks than if they have done it for a year. Thus, Michael and his colleagues conclude, the repetition of social encounters can give rise to commitments. Such a conclusion speaks in favor of the idea that the need to belong, as a bias towards seeking long-standing relationships, supports the sense of commitment, accounting for why we feel more strongly obligated to participate in a joint action when the social ritual with a particular person is maintained over time.

A similar conclusion can be drawn from the experimental studies that Székely and Michael (2018b) have conducted regarding the effect of perceived effort on commitment in joint action. In these experiments, the subjects had to play a 2-player modified version of the 'snake game' in which the participants controlled one axis of the game (left-right) while an algorithm controlled the up-down axis. In experiment 1, the participants were led to believe that the other axis was controlled by someone they had met in the waiting room and who, before each round of the snake game, had to perform a cognitive task to unlock the round. The cognitive task consisted in deciphering a captcha, which could be either difficult (High Effort condition) or easy (Low Effort condition). After that, the subjects were told to play the game, which progressively slowed down thus becoming increasingly boring, and that they could end

the game by pressing a finish button when they judged it was time to move to the next round. Experiment 2 was identical to Experiment 1 except that the subjects were told that their partner was an algorithm. In Experiment 3, the subjects were instructed to perform a cognitive task themselves to unlock each round. Székely and Michael found that participants in Experiment 1 persisted longer at an increasingly boring game when they believed they were playing with a human partner and their partner had had to perform a difficult cognitive task to unlock the round. This effect was not observed when they knew that their partner was an algorithm (Experiment 2) or when they themselves had had to perform the cognitive task (Experiment 3). These findings suggest that our perception of our partner's effort increases our commitment to a particular joint action. Again, the NTB hypothesis can account for such findings. If part of our motivation to engage in a joint action depends on our need to engage in long-standing relations, then a partner's perceived effort provides an important cue to their aptness as a potential partner in a long-standing relation and this should motivate us to collaborate with them.

In addition, the NTB hypothesis seems to avoid the central concerns of the reputation management and the social emotions views. First, the developmental findings presented above suggest that the disposition to engage with others manifests before children understand implicit commitments. As we have seen, even 6-month-old children exhibit selective preference for those they have observed helping another (Hamlin et al. 2007) or those who engage in contingent interactions with them (Hay et al. 1983, 2004; Jacobson 1981). Even children a few weeks of age exhibit behaviors we can associate with the need to belong (Rochat et al. 1999). Such findings do not only cohere with the appropriate developmental timeline but suggest that, unlike social emotions and reputation management, the need to belong does not presuppose sophisticated cognitive abilities (self-recognition, self-representation, capacities for mindreading and for anticipating the mental states of others). Rather, the need to belong

involves a basic "set of psychological dispositions and biological mechanisms biasing the individual to preferentially orient to the social world (social orienting), to seek and take pleasure in social interactions (social reward), and to work to foster and maintain social bonds (social maintaining)". (Chevallier et al. 2012: 231). Thus, the NTB hypothesis can avoid the empirical problems its contenders confront. NTB engages psychological mechanisms that operate early in ontogeny and can be part of an explanation for basic forms of commitments (e.g. sense of commitment or implicit commitments) that emerges in early childhood.

Second, contrary to social emotions and reputation views, the NTB hypothesis does not seem to be subject to the open-question argument. While the question of why do we care about others' distress or our own reputation is relevant when considering the motivations behind our commitments, asking why do we care about a need seems to be an odd question. As a need, the need to belong does not require further psychological mechanisms explaining how it motivates conformity with commitments. The satisfaction of a need, as a requirement for the maintenance an agent's well-being, is a basic and primitive force with an intrinsic motivational value[5]. Thus, an explanation of compliance with commitments in terms of the need to belong does not raise new questions regarding the psychological origins of the motives.

In fact, the need to belong can explain the motivational force behind social emotions and reputation that the open-question argument points out. Social emotions such as guilt serve to repair social bonds when harm is inflicted on others (Vaish et al. 2018). However, as we argue in section 4, children must have developed the capacity to evaluate positively such social relations and engage in them before emotions such as guilt can arise. Such preference for engaging in social relations is explained by the need to belong, so we typically experience guilt

[5] Certainly, one may wonder why humans have a need to belong that (some) other animal species lack. However, this question seems to fall beyond the scope of psychological explanation and ontogenetic development. Rather, like asking why cold-blooded animals need to warm up under the sun, asking for the origin of the need to belong calls for explanations in terms of the evolutionary history.

when we harm someone with whom we have social bonds or an individual who is a (potential) member of our group. Such an interpretation is reinforced by the evidence presented above regarding friendship and the capacity of children to evaluate the prosocial behavior of others. It is also reinforced by evidence suggesting that children exhibit in-group favoritism and bias from early ages (see Everett et al. 2015; Skinner & Meltzoff, 2019 for a review). For instance, 7–8-year-old children exhibit greater generosity towards ingroup members than outgroup members across a series of economic games (Fehr et al., 2008). Thus, in contexts where they had to decide between allocating 1 sweet for themselves and 1 for the partner or just talking 2 for themselves, children promote equality more often when the partner is an ingroup member. Similarly, the need to belong can explain why we are concerned about our reputation. Being motivated to engage in systematic social relations prompts our tendency to behave according to our ingroup reputation standards. Thus, while they are more complex motivations that depend on more sophisticated cognitive abilities, reputation and social emotions are scaffolded at least in part by the need to belong and the behavior they motivate contributes to the satisfaction of this need.

7. Answers to possible objections

We have proposed a solution to the credibility problem that aims at identifying the basic psychological device behind human motivation to honor commitments: the need to belong. Contrary to the reputation management approach and the social emotion approach, the NTB hypothesis coheres with the developmental timeline which situate children's implicit understanding of commitments at the age of two. Further, the need to belong is a primitive motivational factor in the sense both that it operates early in ontogeny and that it can provide a plausible basis for more sophisticated motivations. In this section, we address two possible objections to the NTB hypothesis. First, one may think the NTB hypothesis can be dispensed with because one takes low-level devices such as conditional learning capacities to be sufficient

to account for the basic motivational factors underpinning our tendency to honor commitments. Second, one may object that the NTB hypothesis faces obvious counter-examples: people often honor their commitments in situations where they do not care about others or they do not want to form long standing relations with the recipient of their commitments.

One may resist embracing the NTB hypothesis while avoiding the concerns presented in section 4 and 5 by identifying a simpler route to the acquisition of relevant motivations. For instance, our preference for fulfilling commitments may be elicited through conditional learning when the relevant courses of actions are paired with rewards or sanctions of certain types. Likewise, the motivational endowment underpinning the obligation to fulfill commitments could rely on learning mechanisms that, in principle, require less socialization or sophisticated cognitive capacities to operate. Those willing to exploit such an alternative explanatory route need, nevertheless, to be more specific about the factors that could reinforce or sanction the given responses.

A plausible move in this direction could appeal to external factors, for instance, the co-actors' tendency to punish individuals who fail to behave as their commitments dictate or reward those who do. On this approach, the authors of commitments tend to fulfill their commitments as the result of a history of conditioned reinforcement through sanctioning or rewarding responses by their co-actors. Such an account does not need to appeal to sophisticated capacities to explain how we develop a tendency to meet our commitments, and thus, it does sit well with the developmental trajectory presented in the previous sections. However, evidence from developmental psychology suggests that external rewards do not improve prosocial behavior (Warneken et al. 2007; Warneken & Tomasello, 2008). For instance, experiments with 20-month-old children show that they are subject to the so-called over-justification effect; children who in principle are motivated to behave prosocially are less

motivated to continue in a test phase when they have received rewards during a treatment phase (Warneken & Tomasello, 2008). Also, although several studies suggest that third-party punishment can promote cooperative behavior in children (e.g. Lergetporer et al. 2014), some recent findings indicate that punishment could even diminish the benefit of reciprocity or other social behavior (Dreber et al. 2008; Fehr & Rockenbach, 2003), which indicates that prosocial behavior is not necessarily enforced by sanctions (Bicchieri (2006: 8). Thus, to the extent that conditional learning may rely on external mechanisms, the proposal conflicts with empirical findings that cast into question the role of punishment and external rewards in fueling prosociality.

Another possibility is to appeal to internal factors, for instance, basic emotions that reinforce the relevant pro-social tendencies[6]. On this approach, the development of a preference for honoring commitments would be based on the interaction between conditional learning capacities and basic emotions that reward or sanction a given course of action. For instance, the agent may learn that not acting as expected may produce distress in the recipient, so courses of action incompatible with their commitments are sanctioned by the experience of aversion to others' distress (Michael & Szekély, 2018: 116). Again, such a story would cohere with the relevant developmental timelines. In fact, interpersonal harm aversion or aversion to others' distress seem to appears early in development (Decety and Cowell, 2017). For instance, several

[6] Another endogenous mechanism may appeal to a low-level notion of reputation management. Although Silver and Shawn (2018) argue that managing one's reputation requires high-level capacities (an awareness of the distinction between self and others' evaluations; and a motivation to achieve positive evaluations from others and assess others' accurately), one might argue that agents could develop different mechanisms to increase others' positive beliefs about them without themselves being aware that others could have such beliefs. In principle, one agent could detect a correlation between following the strategies of fulfilling others' expectations and an increment in her success in certain cooperative contexts without realizing that this is due to the increment in their belief that she is a reputable partner. Now the question is what kind of mechanism can increase reputation without one's awareness of it? Given the high cultural variation of what is considered reputable and the empirical evidence presented above, an inborn capacity for reputation seems to be an implausible solution. Such mechanisms could only rely on some sort of emotional endowment of the type analyzed in section 5 or 7. Thus, this route seems to collapse into some of the other alternatives presented in this paper.

studies suggest that neonates are sensitive to others' manifestation of pain, anger or other expressions of distress (Cheng et al. 2012; De Haan et al. 2004; Marchant, 2014). With just a few days of life, babies already possess a neuronal mechanism for discriminating affective vocal reactions (Cheng et al. 2012) and 7-month-old children exhibit more neural activity and look longer when they are presented with fearful faces than when they are presented with happy faces (De Haan et al. 2004). In this sense, such emotional states seem to appear before the capacity to protest the violation of commitments and thus, basic emotions may seem to be a better candidate than the other views to explain our preference for compliance.

While this approach is more coherent with the developmental timeline than some of its rivals, it is dubious, however, whether it can work without presupposing that children already value social relations and care about their preservation. For others' distress, resentment or comfort to play a sanctioning or rewarding function, it seems children would have to care about others' wellbeing, which presupposes certain dispositions towards affiliative stimuli. Of course, one could plausible reply that one's reactions to others' distress or resentment can be the manifestation of a desire to minimize one's own discomfort when observing someone in pain or distress rather than of empathic concerns or socio-affiliative tendencies (Cialdini et al., 1987). To our knowledge, there is no extant empirical evidence in developmental psychology that could help us to decide between these two hypotheses. However, we have some indirect evidence with adults that seem to support the affiliative rather than the personal distress hypothesis. FeldmanHall et al (2014) conducted an experiment where participants were required to decide between financial self-benefit and ensuring the physical welfare of another (they could spend some of their money endowment to spare another person a painful electric shock). At the end of the task, empathic concern and personal distress were measured. It was found that individuals' preference for helping others correlated with their level of empathic concern rather than their personal distress. Further, MRI scanning of the subjects revealed the

activation of the ventral tegmental area, caudate and subgenual anterior cingulate, brain regions associated to social attachment. These data suggest that aversion to others' distress is linked to our capacity for empathetic concern or socio-affiliative tendencies. In other words, finding others' expressions of discomfort or pleasure sanctioning or rewarding presupposes a disposition to find social interactions rewarding in itself, which is a central prediction of the proposal we put forward here.

A second objection to the NTB hypothesis appeals to counterexamples where people fulfill their commitments in joint actions despite not caring about their partners or despite there being no opportunity to promote a long-standing relation with them. The NTB approach seems unable to account for these cases, as the motivation to maintain a long-standing interaction with the other participant cannot be the explanation for such cooperative behavior. Further, humans often feel obligated to follow social norms (see Bicchieri, 2006) mediating cooperative interactions even in contexts where other participants are not necessarily perceived as potential social partners. For instance, people may donate blood, return a lost wallet, or tip while traveling abroad even though they have no expectations of ever forming long-lasting social bonds with the beneficiaries of their actions. In such situations, the objection goes, compliance with norms cannot be explained by a need to belong; rather to explain it we need to appeal to reputational concerns, social emotions or other complex social motivations.

However, our contention is not that social emotions or reputation management have no role to play in an explanation of why we honor our commitments. Rather, we claim that reputation and social emotions are scaffolded at least in part by the need to belong. The need to belong is a need to maintain social bonds with friends and ingroup members that translates into implicit obligations and commitments that help stabilize joint actions and cooperative interactions. Later in development, our understanding of such implicit obligations may scaffold

the appearance of more sophisticated motivations when we realize that breaching such obligations can result in a deterioration of our reputation or cause others harm or distress and make us feel guilt or embarrassment. Once social emotions and reputation management come into play, they can work as independent motivations and, in this sense, they can incorporate different practical reasons and social obligations as immediate components of our tendency to fulfill commitments in a large range of cases. However, to answer the questions of why we care about our reputation or why we care about how others feel about us, we must appeal to a more basic need to engage in social interactions.

Furthermore, we do not rule out the possibility that NTB could play a crucial role in the emergence of other types of complex motivational factors. For instance, as Bicchieri (1997, 2006) has emphasized, human norm-abiding behavior is often not the result of rational choice, cost/benefits calculation or anticipation of punishment. Instead, humans are subject to a process of internalization whereby they consistently learn to behave in conformity to the norms of the group (e.g. the cooperative norm) even when facing a new situation. Such a default strategy appears because, when faced with new situations, individuals search for heuristics that worked well in the past as a way of economizing their effort. In this sense, social norms spread on a population as cognitively cheap heuristics that facilitate regulation of behavior when dealing with new contexts (Bicchieri, 2006: 55-99). Such a dynamic of internalization is premised on the idea that norms appear in small groups where ongoing interactions are the rule, interactions that presuppose a pro-social motivational factor like NTB. The necessary in-group cohesion and long-standing relations that secure the appropriate dynamic of internalization demand the existence of a widespread disposition to engage in social interactions on the population. Thus, like reputation management and social emotions, the tendency to regulate our behavior in accordance to norms presuppose a basic motivational factor such as the one the NTB approach hypothesizes.

Such a basic impulse scaffolds more complex motivations in development but it is also a pervasive immediate motive to engage in joint action and fulfill our commitments in itself. As Godman (2013) has emphasized, one may not necessarily be inclined to engage in a joint action because another agent expects one to do so or because we anticipate their distress; rather one may engage in a joint action because we find acting together pleasant or rewarding in itself. For instance, two children can play a game with one another, even when the game does not have a specific purpose or clear objective (Warneken et al., 2006). We can help a friend assemble their new IKEA furniture not because she really needs our help or because we care about the piece of furniture, but just for the pleasure of hanging out with her. In these scenarios, once we engage in the joint action, we often experience a sense of commitment or even make our commitment explicit without experiencing others' distress or caring about our reputation. The need to belong can explain both why in these situations we find engaging with others rewarding in itself and why we can feel committed to the performance of the joint action even when we were not obligated to carry it out in the first place.

8. Concluding Remarks

Solving the credibility problem requires us, we believe, to provide an account of the motivational forces that lead us to behave in accordance with the expectations generated by our commitments. While normative considerations, whether linked to practical rationality demands or to social normativity, have motivational force, we think there are reasons to doubt that they, by themselves, provide a sufficient motivational basis to fully explain why agents abide by their commitments and thus why their commitments are credible. In addition to normatively derived motivations, we need to appeal to further sources of motivation to block forms of practical irrationality or to counteract selfish motives that would other otherwise threaten the credibility of our commitments. We have surveyed two plausible moves that formulate a solution in terms of reputation management or social emotions. However, the two proposals

are incompatible with developmental findings which situate the understanding of implicit commitments before children start caring about their own reputation or anticipating social emotions. Further, such proposals are also incomplete insofar as they leave pending two central questions: Why do we care about others' assessments of us and our behavior? Why do we care about our reputation?

According to the proposal put forward here, the answer of such questions is also the key to solving the credibility problem. The importance we assign to our reputation, to others' assessments of us and to our credibility are manifestations of a more basic prosocial disposition to engage in long-standing systematic relations with others, that is, the need to belong. The need to belong leads us to engage in certain social interactions and stick to our commitments, but also scaffolds more complex social motivations like reputation management, social emotions and the internalization of group norms that also contribute to explaining adherence to commitment. Taken together, the need to belong and the more complex social motivations it scaffolds lead us to act as committed in a sufficient number of cases to make our commitments credible, and thus, potentially reduce the uncertainties that could jeopardize joint actions.

References

Alexander, R. D. (1987). *The Biology of Moral Systems*. New York: Aldine de Gruyter

Andreoni, J. and Croson, R. (2008) Partners versus Strangers: Random Rematching in Public Goods Experiments. In: C.R. Plott and V.L. Smitt, editors, *Handbook of Experimental Economics Results* (776-783), Volume 1. Amsterdam: North-Holland.

Austin, J. (1962). *How to do Things with Word*s. Cambridge, MA: Harvard University Press.

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, 274, 749–753.

Baumeister R.F., Stillwell, A.M., Heatherton, T.F. (1994) Guilt: An interpersonal approach. *Psychological Bulletin*, 115, 243-267.

Baumeister R.F., Leary, M.R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117, 497 – 529.

Bicchieri, C. (1997). Learning to Cooperate. in C. Bicchieri, R. C. Jeffrey, and B. Skyrms, *The Dynamics of Norms* (pp. 17–46)., Cambridge: Cambridge University Press.

Bicchieri, C. (2006) *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge, UK: Cambridge University Press.

Bicchieri, C., & Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making*, *23*(2), 161-178.

Bicchieri, C. & Xiao, E. (2009). Do the right thing: but only if others do so. Journal of Behavioral *Decision Making*, 22, 191–208.

Bratman, M. E. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Bratman, M.E. (1992). Shared cooperative activity. *The Philosophical Review,* 101(2), 327–341.

Bratman, M.E. (2009a). Shared agency. In C. Mantzavinos (Ed.), *Philosophy of the social sciences: Philosophical theory and scientific practice* (pp. 41–59). Cambridge: Cambridge University Press

Braman, M.E. (2009b). Intention, practical rationality, and self-governance. *Ethics*, 119: 411-443.

Bratman, M. E. (2014). *Shared Agency: A Planning Theory of Acting Together*. Oxford: Oxford University Press.

Brinck, I. (2015). Understanding social norms and constitutive rules: Perspectives from developmental psychology and philosophy. *Phenomenology and the Cognitive Sciences,* 14, 699–718.

Brooks, J., & Lewis, M. (1976). Infants' responses to strangers: Midget, adult, and child. *Child Development*, 47, 323–332.

Brownell, C. A. (2011). Early developments in joint action. *Review of Philosophy and Psychology,* 2(2), 193–211.

Butterfill, S. A., & Sebanz, N. (2011). Joint action: What is shared?. *Review of Philosophy and Psychology*, *2*(2), 137-146.

Carpenter, M. (2009). Just how joint is joint action in infancy?. *Topics in Cognitive Science*, 1(2), 380-392.

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, *70*(3), 560-572.

Chaudhuri, A. (2011). "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature." *Experimental Economics,* 14 (1), 47-83.

Cheng, Y., Lee, S. Y., Chen, H. Y., Wang, P., & Decety, J. (2012). Voice and emotion processing in the human neonatal brain. *Journal of Cognitive Neuroscience*, 24, 1411–1419.

Chevallier C., Kohls G., Troiani V., Brodkin E.S., & Schultz R.T. (2012) The Social Motivation Theory of Autism. *Trends in Cognitive Sciences*.16(4), 231-239.

Cialdini, R.B., Schaller, M., Houlihan, D., Arps, K., Fultz, J., Beaman, A.L. (1987). Empathy-based helping: is it selflessly or selfishly motivated? *Journal of Personality and Social Psychology*, 52 (4), 749–758.

De Haan, M., Belsky, J., Reid, V., Volein, A., & Johnson, M. H. (2004). Maternal personality and infants' neural and visual responsivity to facial expressions of emotion. *Journal of Child Psychology and Psychiatry*, *45*(7), 1209-1218.

De Hooge, I. E., Zeelenberg, M., & Breugelmans, S. M. (2007). Moral sentiments and cooperation: Differential influences of shame and guilt. *Cognition and emotion*, 21(5), 1025-1042.

Dreber A., Rand D. G., Fudenberg D., & Nowak M A. (2008). Winners don't Punish. *Nature* 452, 348–351.

Everett, J. A., Faber, N. S., & Crockett, M. (2015). Preferences and beliefs in ingroup favoritism. *Frontiers in behavioral neuroscience*, *9*, 15.

Frank, R. (1988). *Passions within reason*. New York: Norton.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature*, *454*(7208), 107

Fehr E., & Rockenbach B. (2003). Detrimental Effects of Sanctions on Human Altruism. *Nature* 422, 137–140.

FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *NeuroImage*, 105, 347–356.

Gifford-Smith M., & Brownell C. (2003). Childhood peer relationships: social acceptance, friendships & peer networks. *Journal of School Psychology*, 41, 235-284.

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

Gilbert, M. (2006). Rationality in Collective Action. *Philosophy of the Social Sciences* 36 (1), 3–17.

Gilbert, M. (2009). Shared intention and personal intentions. *Philosophical Studies,* 144, 167–187.

Godman, M. (2013). Why we do things together: The social motivation for joint action. *Philosophical Psychology*, 26(4), 588-603.

Godman, M., Nagatsu, M., & Salmela, M. (2014). The social motivation hypothesis for prosocial behavior. *Philosophy of the Social Sciences*, 44(5), 563-587.

Gräfenhain, M., Behne, T., Carpenter, M., & Tomasello, M. (2009). Young children's understanding of joint commitments. Developmental Psychology 45, 1430–1443.

Haley, K. J., & Fessler, D. M. T. (2005). Nobody's watching? Subtle cues affect generosity an anonymous economic game. *Evolution and Human Behavior*, 26, 245–256.

Hamlin, J. K. (2015). The infantile origins of our moral brains. In J. Decety & T. Wheatley (Eds.), *The moral brain—Multidisciplinary perspectives* (pp. 105–122). Cambridge, MA: MIT Press.

Hamlin J.K., & Wynn K. (2011). Young infants prefer prosocial to antisocial others. Cognitive Development, 26(1), 30–39.

Hamlin J.K., Wynn K., & Bloom, P. (2007) Social evaluation by preverbal infants. *Nature,* 450, 557 – 559.

Hareli, S., & Parkinson, B. (2008). What's social about social emotions?. *Journal for the Theory of Social Behaviour*, *38*(2), 131-156.

Harris, P. L., Olthof, T., Terwogt, M. M., & Hardman, C. E. (1987). Children's knowledge of the situations that provoke emotion. *International Journal of Behavioral Development*, *10*(3), 319-343.

Hay, D.F., Nash, A., & Pedersen, J. (1983). Interaction between six-month-old peers. *Child Development*, 54, 557–562.

Hay, D. F., Payne, A., & Chadwick, A. (2004). Peer relations in childhood. *Journal of child psychology and psychiatry*, 45(1), 84-108.

Heintz, C., Karabegovic, M., & Molnar, A. (2016). The co-evolution of honesty and strategic vigilance. *Frontiers in psychology*, 7, 1503.

Howes C. (1996). The earliest friendships. In W.M. Bukowski, A.F. Newcomb, W.W. Hartup (Eds.), *The company they keep: friendship in childhood and adolescence (*pp. 66 – 86). Cambridge, UK: Cambridge University Press.

Jacobson, J.L. (1981). The role of inanimate objects in early peer interaction. *Child Development*, 52, 618– 626

Kahneman, D., & Egan, P. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Ketelaar T., & Au, W.T. (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as information interpretation of the role of emotion in social interaction. *Cognition and Emotion*, 17,429-453.

Lakin J.L., & Chartrand T.L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Sciences* 14, 334 – 339.

Leary J.L. and Allen, A. B. (2011) Belonging Motivation: Establishing, Maintaining, and

Repairing Relational Value. In D. Dunning (Ed.), *Social Motivation* (pp. 37-56). Psychology Press.

Ledyard, J. (1995). Public goods: a survey of experimental research. In A. Roth & J. Kagel (Eds.), n *Handbook of Experimental Economics* (pp. 111-194). Princeton: Princeton University Press.

Lergetporer, P., Angerer, S., Glätzle-Rützler, D. & Suttera, M. (2014) Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. *Proceedings of the National Academy of Sciences*, 111(19), 6916–6921.

Leimgruber, K. L., Shaw, A., Santos, L. R., & Olson, K. R. (2012). Young children are more generous when others are aware of their actions. *PloS one*, 7(10), e48292.

Lewis, M., Sullivan, M. W., Stanger, C., & Weiss, M. (1989). Self development and self-conscious emotions. *Child Development*, 60, 146–156.

Ledyard, O. (1995). Public goods: some experimental results. In J. Kagel & A. Roth (Eds.), *Handbook of experimental economics*. Princeton: Princeton University Press

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.

Marchant, A. (2014). Neonates do not feel pain: A critical review of the evidence. *Bioscience Horizon*, 7, 1–9.

Mele, A., 1987, *Irrationality*, New York: Oxford University Press.

Michael, J. (2011). Shared emotions and joint action. *Review of Philosophy and Psychology*, 2(2), 355-373.

Michael, J., and Pacherie, E. (2014). On commitments and other uncertainty reduction devices. *Journal of Social Ontology*,1, 1–34.

Michael, J., & Salice, A. (2017). The Sense of Commitment in Human–Robot Interaction. *International Journal of Social Robotics*, 9(5), 755-763.

Michael, J., Sebanz, N. & Knoblich, G. (2016). The sense of commitment: a minimal approach. *Frontiers in Psychology*, 6, 1–11.

Michael, J., & Székely, M. (2018). The developmental origins of commitment. *Journal of Social Philosophy*, 49(1), 106-123.

Mills, R. S. (2005). Taking stock of the developmental literature on shame. *Developmental Review*, 25(1), 26-63.

Mills, R. S., Arbeau, K. A., Lall, D. I., & De Jaeger, A. E. (2010). Parenting and child characteristics in the prediction of shame in early and middle childhood. *Merrill-Palmer Quarterly* (1982-), 500-528.

Moore, G. A. (1903). *Principia Ethica*. Cambridge: Cambridge University Press.

Nelissen, R. M. A., & Zeelenberg, M. (2009). When guilt evokes self-punishment: Evidence for the existence of a Dobby effect. *Emotion*, 9, 118–122.

Newcomb A.F., and Bagwell C.L. (1995). Children's friendship relations: a meta-analytic review. *Psychological Bulletin*. 117, 306 – 347. (doi:10.1037/0033- 2909.117.2.306)

Nielsen M. (2006). Copying actions and copying outcomes: social learning through the second year. *Developmental Psychology*, 42, 555 – 565.

Nielsen M. (2009). The imitative behavior of children and chimpanzees: a window on the transmission of cultural traditions. *Revue de primatologie*, 1, 254.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437, 1291–1298.

Over H. (2016). The origins of belonging: social motivation in infants and young children. Philosophical Transaction of the Royal Society: Biology 371: 20150072. http://dx.doi.org/10.1098/rstb.2015.0072

Over H., & Carpenter M. (2009). Priming third-party ostracism increases affiliative imitation in children. Developmental Sciences 12, F1 – F8. (doi:10.1111/j.1467-7687. 2008.00820.x)

Pacherie, E. (2011). Framing Joint Action. *Review of Philosophy and Psychology,* 2 (2), 173-92.

Parisette-Sparks, A., Bufferd, S. J., & Klein, D. N. (2017). Parental predictors of children's shame and guilt at age 6 in a multimethod, longitudinal study. *Journal of Clinical Child & Adolescent Psychology*, 46(5), 721-731.

Rakoczy, H. (2006). Pretend play and the development of collective intentionality. *Cognitive Systems Research*, 7, 113-127.

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: young children's awareness of the normative structure of games. *Developmental Psychology*, 44(3), 875.

Rochat, P., Broesch, T., & Jayne, K. (2012). Social awareness and early self-recognition. *Consciousness and cognition*, 21(3), 1491-1497.

Rochat P., Querido J.G., & Striano T. (1999). Emerging sensitivity to the timing and structure of protoconversation in early infancy. *Developmental Psychology,* 35, 950–957.

Roth, A. S.. (2004).Shared Agency and Contralateral Commitments. *The Philosophical Review,* 113 (3), 359–410.

Rusch, H., & Luetge, C. (2016). Spillovers from coordination to cooperation: Evidence for the interdependence hypothesis?. *Evolutionary Behavioral Sciences*, *10*(4), 284-296.

Salmela, M., & Nagatsu, M. (2016). Collective emotions and joint action. *Journal of Social Ontology*, 2(1), 33-57.

Sally, D. (1995). Conversation and cooperation in social dilemmas a meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, 7(1), 58–92.

Scanlon, T. (1998). *What we Owe to Each Other*. Cambridge, MA: Harvard University Press.

Schelling, T.C. (1960). *The strategy of conflict*. Cambridge, MA.: Harvard University Press

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Science*, 10, 70–76.

Shaw, A. & Olson, K. (2015). Whose idea is it anyway? The importance of reputation in acknowledgement. *Developmental Science*, 18, 502–509

Shpall, S. (2014). Moral and rational commitment. *Philosophy and Phenomenological Research*, 88, 146–172.

Silver, I. M., & Shaw, A. (2018). Pint-Sized Public Relations: The Development of Reputation Management. *Trends in cognitive sciences*, 22(4), 277-279.

Siposova B., Tomasello M., & Carpenter. M. (2018) Communicative eye contact signals a commitment to cooperate for young children. *Cognition*, 179, 192-201.

Skinner, A. L., & Meltzoff, A. N. (2019). Childhood experiences and intergroup biases among children. *Social Issues and Policy Review*, *13*(1), 211-240.

Stefani, L.H., & Camaioni, L. (1983). Effects of familiarity on peer interaction in the first year of life. *Early Child Development and Care*, 11, 45–54.

Strandberg, C. (2004). In defence of the open question argument. *The Journal of ethics*, 8(2), 179-196.

Sugden, R. (2000). “The Motivating Power of Expectations.” In *Rationality, Rules, and Structure*, edited by J. Nida-Rümelin and W. Spohn, 103-29. Dordrecht: Springer.

Székely, M. & Michael, J. (2018). Investment in Commitment: Persistence in a Joint Action is Enhance by the Perception of a Partner's Effort. *Cognition*, 174, 37-42.

Trevarthen C., & Aitken K.J. (2001). Infant intersubjectivity: research, theory, and clinical applications. *Journal of Child Psychology and Psychiatry*, 42, 3 – 48.

Tollefsen, D. (2005). Let’s pretend! Children and joint action. *Philosophy of the Social Sciences,* 35(1), 75–97.

Tomasello, M. (2009). *Why we cooperate*. Cambridge: MIT press

Tomasello M., & Carpenter, M. (2007). Shared intentionality. *Developmental Sciences*, 10, 121 – 125.

Tomasello, M., & Rakoczy, H. (2003). What makes human cognition unique? From individual to shared to collective intentionality. *Mind and Languag*e, 18(2), 121–147.

Trivers, R.L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46 (1), 35-57.

Vaish, A. (2018). The prosocial functions of early social emotions: the case of guilt. *Current opinion in psychology*, 20, 25-29.

Vaish, A., Carpenter, M., & Tomasello, M. (2016) The Early Emergence of Guilt-Motivated Prosocial Behavior. *Child Development*, 87 (6), 1772–1782.

Vesper, C., Butterfill, S., Sebanz, N., & Knoblich, G. (2010). A minimal architecture for joint action. *Neural Networks*, 23, 998–1003.

Van Vugt, M., Roberts, G., & Hardy, C. (2007). Competitive altruism: Development of reputation-based cooperation in groups. In R. Dunbar & L. Barrett (Eds.), *Handbook of Evolutionary Psychology* (pp.531–540). Oxford, England: Oxford University Press

Warneken, F., & M. Tomasello. (2007). Helping and cooperation at 14 months of age. *Infancy* 11: 271–294.

Warneken, F., F. Chen, & M. Tomasello. (2006). Cooperative activities in young children and chimpanzees. *Child Development,* 77(3), 640–663.

Warneken F., & Tomasello M. (2008). Extrinsic rewards undermine altruistic tendencies in 20-month-olds. *Developmental Psychology*, 44, 1785 – 1788.

Young, G., & Lewis, M. (1979). Effects of familiarity and maternal attention on infant peer relations. *Merrill Palmer Quarterly*, 25, 105–119.

Zahavi, A., & Zahavi, A. (1997). *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford: Oxford University Press.

Zahn-Waxler, C., Radke-Yarrow, M., & Wagner, E. (1992). Development of concern for others. *Developmental Psychology*, 28, 126–136.