Phenomenology and the Cognitive Sciences
https://doi.org/10.1007/s11097-022-09805-x

# Breaking the right way: a closer look at how we dissolve commitments

**Matthew Chennells**[1] · **John Michael**[2]

Accepted: 4 February 2022


## Abstract

Joint action enables us to achieve our goals more efficiently than we otherwise could, and in many cases to achieve goals that we could not otherwise achieve at all. It also presents us with the challenge of determining when and to what extent we should rely on others to make their contributions. Interpersonal commitments can help with this challenge – namely by reducing uncertainty about our own and our partner's future actions, particularly when tempting alternative options are available to one or more parties. How we know whether a commitment is in place need not, however, be based on an explicit, identifiable event; in many cases, joint action is stabilized by individuals' experience of an implicit sense of commitment, which is sensitive to subtle situational cues such as the effort costs invested by one or more agents. While an emerging body of work has investigated the conditions under which a sense of commitment may emerge and/or be strengthened, little attention has been paid to the conditions under which people are comfortable *dissolving* commitments. Specifically, what are the factors that modulate people's motivation and which determine whether circumstances merit the dissolution of a commitment? After evaluating and rejecting the answers to this question suggested by standard approaches to commitment, we develop a new approach. The core insight which we articulate and defend is that, when considering whether new information or changing circumstances merit the dissolution of a commitment, people virtually bargain with their partners, performing a simulation of a bargaining process with the other person, including imagining how the other will feel and act towards them, and what effect this will have on them. The output of this simulation is a consciously accessible, affective state which provides motivation either to dissolve the commitment or to persist in it. Overall, our account expands our understanding of the phenomenology of being motivated to act committed in joint activity, an area in which existing accounts of interpersonal commitment fall short.

**Keywords** Joint action · Commitment · Virtual bargaining · Social expectations · Norms

Extended author information available on the last page of the article

Published online: 14 April 2022

## 1 Introduction

Imagine the following scenario. James and his friend Giulia have agreed to take a walk in the park this Saturday morning. They have coordinated the necessary decision-making (James will bring breadcrumbs for the swans, Giulia brings the binoculars for birdwatching, etc.). But now, waking up on Saturday morning, James finds that the person he met in the bar last night is already up and preparing breakfast, and he feels very much inclined to stay home and spend a few more hours with this exciting new acquaintance. Or, in an alternative scenario, James wakes up to find the water pipe in his bathroom has sprung a leak and, while he could wait until later to deal with it, he is inclined to address it right away. Would it be permissible, in either of these scenarios, to cancel his engagement with Giulia? What if he can't use his phone (he can't find it after a wild night; his wet clothes damaged his phone) and therefore has no way of contacting Giulia?

In everyday life, we often find ourselves confronted with situations in which we would like to extricate ourselves from commitments that we have made – because our interests have changed, because we are tempted by some alternative that has arisen, because it is no longer feasible, because there is a conflicting commitment which we value more, etc. What factors or principles do we, or should we, appeal to in such situations, and what is the reasoning process we go through in doing so?

As we shall see, this type of scenario turns out to be a revealing test case for theories of commitment. It is easy to be lured into thinking that it is a straightforward matter: when we want to be released from commitments, we need only ask to be released. If the person to whom we are committed releases us, we are free; if not, then we remain committed. We will refer to this conception as the simple view. The simple view follows from standard theoretical approaches to commitment in the philosophical literature (e.g. Bratman, Gilbert, Searle, Shpall), which say much about how commitments are generated but little about how they are dissolved. More generally, these approaches also appear to lack an adequate explanation of what it is that motivates agents to meet, and not dissolve or renege on, their commitments. Our view on this follows closely in the spirit of recent work by Fernandez Castro and Pacherie (2020), who discuss what they call "the credibility problem for commitments"[1]. The problem these authors are concerned with is theoretical: existing approaches to commitment in the philosophy of joint action fail to provide a sufficiently robust account of the motivational basis that fully explains why it is that agents abide by their commitments and, so, why their commitments should be seen to be credible. In this sense, the shortcoming is in theory rather than in practice, in that we observe that people do, in reality, seem to meet commitments they have made, though we find it hard to explain why, if we take as a starting point existing theoretical accounts of joint action.

Like these authors, we believe that current accounts don't fully capture the myriad, and often subtle, ways in which we are motivated to meet our commitments. More specifically, we believe that the link between, on the one hand, a purely normative account of commitment in joint action and, on the other, a phenomenological account which pays attention to the ways in which we actually experience, and are motivated

[1] We are grateful to an anonymous reviewer for making clear the relevance of this research for our project.

to meet commitments, is weakly explicated in existing accounts. We aim to show this by presenting a descriptive account of commitment in joint action, drawing on research which tests the extent to which an agent's sense of commitment to their partner in a joint action - that is, the motivation they have to remain committed in a joint activity - is modulated by situational cues of that partner's expectations. The studies we describe, which test a range of situational cues, are based on a minimal conceptualisation of commitment, in which several features of commitment in a strict sense - that is, types of interpersonal relations with a more traditional, promise-like character - are absent. We contrast this with a simplified account of the dynamics of commitment dissolution inspired by existing normative theories of joint action.

We note that the minimal account we propose is not incompatible with, nor does it seek to replace, any existing theoretical conceptualisation of commitment. Just the opposite; existing theories provide us with our theoretical starting point by grounding what we regard as two of the most essential features of commitments: first, that they give rise to a particular form of obligation, which is to meet commitments we have made; second, that these obligations are not general, but are specifically directed towards those we have made commitments to. Further to this, with the minimal account we propose we are neither aiming to provide a comprehensive characterisation of interpersonal commitment nor an exhaustive list of all the factors that may motivate us to either meet or abandon commitments we make to others. One of our hopes, instead, is to highlight how our experiences of commitment are far richer and more nuanced than what might be expected were we to focus solely on what is contained within traditional philosophical approaches to joint action.

Paying attention to additional, often overlooked, situational features of commitment is not, however, all we are hoping to achieve in this paper. Rather, another of our motivations is to look at the implications for a more general understanding of the psychological processes at play in situations involving commitments in joint action. Though a fully-fleshed out account that captures the full range of such processes is beyond the scope of this paper, our aim is to provide evidence and support for a view of cognition in commitment that makes room for basic, proximal mechanisms which modulate an agent's motivation to generate, meet or dissolve commitments they make with joint interaction partners. We believe that recent theoretical and empirical work both point towards this need - and in this we take a further, tentative step in exploring what the normative implications of this view might be.

Returning to the case at hand, in which we have an agent confronted with situations in which they might prefer to dissolve their commitments, we noted that a simple view that sees commitment dissolution as a straightforward matter, in which we simply request release, tells us little about the different forces that motivate for and against this release. We think it is important to focus on contrasting this view with an alternative which sees possibly non-reflective processes as integral to requests like this. The other reason for thinking this contrast is important is that it is the simple view (which draws directly on purely normative accounts of commitment) which has informed much of the empirical research that has been undertaken so far concerning the dissolution of commitments.

As we shall demonstrate, however, this way of thinking does not provide a phenomenologically adequate explanation capturing the actual dynamics that unfold in

such situations (Sect. 2). Sometimes it would not be appropriate to ask for release, and sometimes it may be awkward to do so. Likewise, sometimes it is awkward or difficult to say "no" if one is asked to release someone else. And indeed, even if one does ask to be released, it is far from clear how the various costs and benefits should be weighed against each other in order to decide whether or not to release is appropriate. In short, there appears to be a gap between what we can glean from the simple view and a proper explanation of the motivation we have to meet or dissolve our commitments. This provides the motivation to take a closer look at these dynamics, as we proceed to do in Sects. 3–4. As we shall see, careful consideration of the dynamics of commitment dissolution also turns out to generate important insights about how we identify and assess the level of motivation in our commitments in the first place, and about what we actually care about when we care about commitments. Most importantly, careful examination of the dynamics of commitment dissolution will enable us to provide answers to the following four key questions:

1) What factors/principles do we appeal to in situations in which we may want to dissolve commitments?
2) What are the reasoning processes we go through when considering whether to request release from a commitment?
3) How do we identify and assess the level of motivation in our commitments we have to others in the first place?
4) What do we actually care about when we talk about caring for commitments?

## 2 The Simple View

In the philosophical literature, commitment is usually treated as a relation among one committed agent, one agent to whom the commitment has been made, and an action which the committed agent is obligated to perform in virtue of having given her assurance to the second agent that she would do so (Michael et al., 2016a; cf. Gilbert, 1990; Scanlon, 1998; Searle, 1969; Shpall, 2014). If we start out from this standard conception of commitment, then we are likely to arrive at a particular answer to the question about when it is appropriate to dissolve commitments. Specifically, we are likely to think that it is appropriate whenever the second agent agrees to relinquish their entitlement to expect the committed agent to perform the action. We shall call this the *simple view* of commitment dissolution:

> Simple View: we are likely to think that it is appropriate for a committed agent to dissolve their commitment whenever their partner agrees to relinquish their entitlement to expect the committed agent to perform the action.

To illustrate where this view comes from, let's briefly consider two leading theoretical accounts, from Margaret Gilbert and Michael Bratman, to show how commitment dissolution is approached when the simple view is adopted; namely, that it is largely ignored and, where addressed, generally underspecified in its explanation of why agents are motivated to act committed (see Michael & Pacherie, 2015, and Fernandez

Castro & Pacherie, 2020, for a more detailed discussion of the possible role of commitments in the accounts of these two authors).

For Margaret Gilbert, a structure of joint commitments can be explained in reference to the process required for members wanting to dissolve it. On her account, joint commitments involving two or more people can only be created and rescinded by those individuals together. Gilbert notes that a key feature of individual *personal* commitments "is that the one who personally formed or made the corresponding personal decision or intention is in a position unilaterally to expunge them as a matter of personal choice" (Gilbert, 2009). While one may rescind one's own personal commitment by simply changing one's mind, because joint commitments are not built up from personal commitments one cannot rescind a joint commitment in the same way. In the absence of special background understandings, unilaterally deciding to drop a joint commitment by, for example, choosing not to act in accordance with it without the concurrence of the other parties, is thus a violation of it and not its revocation. Unless concurrence on its release has been given, individuals have a mutual obligation to one another to the performance of their part in the joint activity. A concurrence condition is explicitly made and is core to her theoretical account: that is, "absent special understandings, the concurrence of all parties is required in order that a given shared intention be changed or rescinded, or that a given party be released from participating in it" (Gilbert, 2009). Her account thus involves a kind of social normativity at its heart, involving a form of obligation *sui generis* to shared intentional activity. Yet, she does not go further in discussing either the extent to which these obligations are expected to motivate individuals to meet them in the face of competing considerations (with their own normative grounding) or the factors that individuals should take into account when requesting a partner's concurrence. Indeed, a strict reading of her account inclines the reader towards believing that one's partner is not even able to refuse such concurrence when it is requested! A more charitable reading suggests, rather, that she takes this process for granted in providing a purely normative account of commitment in joint action. For our purposes here, we believe that Gilbert's account does not give us much in the way of understanding when it's appropriate or not to dissolve commitments.

In Michael Bratman's (2014) view, interpersonal commitments are generated when individuals share an intention to act jointly. Here, shared intention is a complex of interlocking intentions of the individuals which plays a basic role in helping coordinate the intentions and planning of all agents involved, allocating roles and responsibilities between them and tracking the goal they have of their joint activity. Unlike Gilbert, Bratman doesn't regard joint action as involving any special form of social normativity that creates special obligations between the agents involved. Rather, what commitments are present are those that are distinctively characteristic of intentions in the way he sees them within his planning theory of agency (Bratman, 1987). Specifically, individual intentions are commitments to act - and, as such, are subject to a general commitment to norms of practical rationality, including that they are stable, conduct controlling and prompt reasoning about means. Bratman argues these norms extend to the joint case, which necessarily involve "commitments to mutual compatibility of relevant sub-plans, commitments to mutual support, and joint-action tracking mutual responsiveness." (Michael & Pacherie, 2015). In rela-

tion to our question of when it is appropriate to resolve commitments, we note two features of Bratman's account. First, while intentions are governed by a norm of stability - that is, in the absence of relevant new information, an intention is rationally required to resist reconsideration - Bratman is not more specific about what constitutes new information nor what it means for it to be relevant. His assumption, that once we have formed an intention we see the matter of our acting as settled, therefore leaves little room for thinking about when, or how often, it is appropriate to revisit our commitments. Second, perhaps more fundamentally, commitments founded on norms of intention rationality don't give us much insight into how these normative constraints motivate us to act, particularly in relation to other attractive alternatives that we might, under a norm of self interest, be required to turn to. This concern is similar to that mentioned in relation to Gilbert's account and is reflective of a more general problem; namely, that existing normative accounts don't give us much direction in the way of thinking about when and what factors should be considered when thinking about whether to meet or dissolve our commitments.

We note briefly a later addition to Bratman's account where he addresses, in light of Gilbert's work, the question of some kind of social normativity in shared intentional activity. With respect to obligations, while Bratman is at pains to emphasise that, unlike Gilbert, he makes no explicit appeal to the necessity of obligations and entitlements, he acknowledges that typical cases of shared intentional activity (e.g. those not involving coercion or deception) are usually accompanied by certain kinds of interpersonal obligations. He outsources this component by drawing on Thomas Scanlon (1998) to discuss the moral requirement to meet expectations that one has voluntarily and intentionally created in another (and which they have come to rely on), such that "in the absence of special justification, A must do X unless B consents to X's not being done" (as Scanlon's 'Principle of Fidelity' concludes), or otherwise take "reasonable steps" to prevent or compensate for a partner's possible losses in cases where reasonable expectations are violated. Though Bratman says nothing (and Scanlon very little) about the factors and processes that characterise how A goes about this and what A considers when requesting consent, this provides us with an early indication of how we might use a normative account like Bratman's as a springboard for thinking about when individuals should meet their commitments; namely, when their partners have reasonable expectations that they will do so.

The simple view we presented earlier is based partly on our interpretation of the way commitments emerge from these two purely normative accounts of shared intentional activity. We have argued that we don't believe either account, in isolation, gives us much insight as to when it's appropriate or not to dissolve commitments. In particular, it is worth emphasizing two general points about the simple view conception of commitment.

First, this conception presents us with a particular explanation of why people should do the things they are committed to doing, and of why we are willing to rely on them to do so – namely, because commitments generate obligations and entitlements directed towards our interaction partners. More specifically, commitments enable us to take on obligations that we would not otherwise have and thereby to provide assurance to others that would otherwise be lacking. To illustrate: James always has the obligation to pay his taxes and to treat others with respect - even without making any

commitment to doing so. In contrast, he has the obligation to take a walk in the park with Giulia today because he has made a commitment to do so (whether to her or to a third party). The assurance thereby provided would be especially valuable to Giulia if she must forgo other opportunities in order to take the walk with James or if she has reason to be uncertain about his future willingness to take the walk (e.g. because he has attractive other options or because he is known to be impulsive). It is evident that such assurance would not be necessary in the absence of uncertainty; that is, if Giulia could perfectly predict her own and James' behaviours as well as the affordances and action-outcomes of their action environment. Thus, commitments enable us to further constrain our range of possible actions beyond the general constraints that exist simply by virtue of living in society. This is a valuable function: by reducing uncertainty about our future actions, commitments facilitate the planning and coordination of multifarious joint actions unfolding over arbitrarily long timescales (Michael & Pacherie, 2015). However, neither of these two normative accounts provides us with much insight into the underlying motivation that agents have to meet their commitments, an argument also made by Fernandez Castro & Pacherie (2020): to properly understand how commitments perform their function of reducing uncertainty in joint action, we need to understand what makes them credible - and to do this, we need to explain what motivates agents to act as committed. In their analysis, however, these authors conclude that attempting to map the normative reasons to act that these accounts propose with motivation to act is unsuccessful: "they cannot provide a sufficient motivational basis to fully explain why agents abide by their commitments and thus why their commitments are credible" (Fernandez Castro & Pacherie, 2020). Specifically, there are reasons to doubt that either norms of practical rationality (Bratman) or social normativity (Gilbert) provide the kind of motivation that is needed for an agent to remain committed and eschew more attractive alternative options that may be in their interest - though we acknowledge that neither Bratman nor Gilbert aim for this in their respective accounts.

Second, as implied by the simple view, commitment is treated in this literature as a binary notion: either the aforementioned conditions have been fulfilled (and there is a commitment) or they have not (and there is no commitment). Thus, it does not provide us with a basis for distinguishing among different degrees of motivation in commitment. For example, it does not enable us to say that James may have a higher degree of motivation given his commitment to taking a walk with Giulia if he knows that she has driven one hour to reach the park or if he knows that she has turned down the alternative option of having brunch with her sister. We might think, though, that a useful conceptualisation would illuminate the graded nature of motivation within commitments, and explain how agents calibrate their motivation to meet commitments.

In summary, the simple view, and the accounts on which it is based, do not provide a full explanation of agents' motivation to meet their commitments, nor of the graded nature of commitment. Yet something like the simple view has shaped many empirical studies that have so far been undertaken to investigate the psychology of commitment (and in particular commitment dissolution). For example, one recent study by Kachel and colleagues (2019) probed children's responses to scenarios in which a puppet playmate abandoned a joint action. In one condition, the puppet simply

stopped playing, in a second condition it requested to be released from the commitment to play together, and in a third condition it announced that it would leave the game. The main finding was that even three-year-old children did differentiate among these conditions, indicating that children as young as three understand that it is possible to be released from commitments by asking for permission.

The interpretation of these findings suggested by the simple conjecture is that children acquire the concept of commitment by around three. But consider a study conducted by Mant & Perner (1988), in which children were presented with vignettes describing two children on their way home from school, Peter and Fiona, who discuss whether to meet up and go swimming later on. In one condition, they make a joint commitment to meet at a certain time and place, but Peter decides not to go after all, and Fiona winds up alone and disappointed. In the other condition, they do not make a joint commitment, because Fiona believes that her parents will not let her. She is then surprised that her parents do give her permission, and she goes to the swimming pool to meet Peter. In this condition, too, however, Peter decides not to go after all, so again Fiona winds up alone and disappointed. The children in the study, ranging from 5 to 10 years of age, were then asked to rate how naughty each character was. The finding was that only the oldest children (with a mean age of 9.5) judged Peter to be more naughty in the commitment condition than in the no-commitment condition. This may seem late, but it is, in fact, consistent with the findings of a study by Astington (1988), who reported that children under 9 fail to understand the conditions under which the speech act of promising gives rise to commitments. If we take these results at face value, it suggests that the development of children's understanding of commitment is protracted. Whatever it was that Gräfenhain and colleagues' (2009) study was tapping into in three-year-olds, it was not full mastery of the concept of commitment in the strict sense. This indicates that we need some other explanation of the pattern observed with these younger children.

More generally, the simple conjecture does not provide us with any guidance in generating predictions about what components of the concept of commitment may emerge first, or about what behavioral tendencies may emerge first (waiting for a partner, checking on her, helping her, persisting until all parties are satisfied that the goal has been reached, protesting if a partner abandons a joint action, etc.). In other words, the simple conjecture presents a complex concept and a suite of behaviors licensed by the concept as a single package. But these components may come apart, and some may be more basic than others. The simple conjecture does not tell us in what order these components should emerge, which components are most basic, or how the developmental process should unfold.

Having summarised the simple view and briefly discussed two of its general limitations, let us now return to the main thread by considering the answers which the simple view provides to each of the four key questions identified above:

With respect to the first question (What factors/principles do we appeal to in situations of commitment dissolution?), the simple view suggests that, when we desire to be released from a commitment, provided we have a good reason for doing so, we simply ask.

With respect to the second question (What are the reasoning processes we go through when considering whether to request release from a commitment?), the sim-

ple view proposes that we consider whether there are any obligations which outweigh the obligation associated with the commitment in question.

With respect to the third question (How do we identify and assess the level of commitments we have to others in the first place?), the simple view states that we keep track of our commitments by remembering having entered into them. It provides no basis for distinguishing among levels of commitment.

With respect to the fourth question (What do we actually care about when we talk about caring for commitments?), the simple view states that we care about meeting our obligations.

Despite its simplicity and its intuitive appeal, we therefore believe that the simple view is inadequate. Most importantly, it does not explain why we may sometimes deem it inappropriate or awkward to request release. More specifically, it tells us nothing about the principles/factors that are relevant to consider in cases in which we consider asking for release (see the answer to question 1 above), nor about the psychological processes that underpin our judgments in such cases (see the answer to question 2 above). Because of this, it also fails to explain why sometimes, even when release from a commitment is expressly granted, we nevertheless feel as though we had violated a commitment, and there can nevertheless be damage to the relationship.

To address these shortcomings, we believe that a different approach is needed. In the following section, we sketch a recently developed theory of the psychological underpinnings of commitment which constitutes the starting point for an alternative approach to commitment dissolution. On this view, agents may develop and experience a sense of commitment towards a partner – even in the absence of explicit communication or in cases agents are uncertain of whether an obligation (or a specific 'type' of obligation) is present. Crucially, this sense of commitment both explains an important source of our motivation to act committed and explains how such motivation may be felt in degrees, such that agents are more or less motivated to meet expectations their partners may have of their future action performance.

This helps us isolate and address what we care about when we care about a commitment: that is, notwithstanding how a commitment is established, we care about it to the extent that we sense a commitment and are motivated to act in the direction of its fulfilment. We then present a psychological description of commitment dissolution in which agents simulate bargaining with a partner before deciding whether to request release or refrain from fulfilling expected actions. The virtual bargaining account we draw on meets several important criteria we believe such a model of social interaction should meet.

## 3 The Sense of Commitment

Recently, in the psychological literature, Michael, Sebanz & Knoblich (2016a) have proposed an alternative approach which treats motivation in commitment as a graded phenomenon: an agent can be more or less motivated to perform an action that a second agent expects, and may feel more or less guilty if she does not perform the action. To capture this, they introduce the notion of a 'sense of commitment,' which

admits of degrees. Following Michael and colleagues, we will adopt the following definition:

> Sense of Commitment (SoC): A has a sense of being committed to performing X to the extent that A is motivated by her belief that B expects her to contribute X and may be relying on that expectation.

This approach differs in several respects from the simple view presented in the previous section.[2] Three of these are worth emphasising here. First, while the simple view entails a binary conception of commitment, this approach provides us with a graded conception: insofar as motivations and expectations come in degrees, so does the sense of commitment. To borrow an example from Michael, Sebanz & Knoblich (2016a; itself adapted from Gilbert, 2009):

> Polly and Pam are in the habit of smoking a cigarette and talking together on the balcony during their afternoon coffee break. They have never explicitly agreed to do this, but Polly is aware that Pam expects her to show up today, like every other day. The sequence is broken when one day Pam waits for Polly but the latter doesn't arrive. This may be experienced by Polly and Pam as a violation of a commitment. Moreover, the extent to which this is the case will depend on further details about the case. For example, if Polly and Pam have smoked and talked together every day for 2 or 3 weeks, Polly might feel only slightly obligated to offer an explanation, but she would likely feel more strongly obligated if the pattern had been repeated for 2 or 3 years. Thus, we can see that in everyday cases like this, the sense of commitment comes in degrees.

Second, while the standard account is tailored to cases of explicit commitment (i.e., when an assurance has been given verbally, in the form of a promise or otherwise), this is not true of the sense of commitment framework: many situational factors can modulate expectations and motivations in the absence of any explicit verbal assurance. The example of Polly and Pam also provides preliminary motivation for this thought by illustrating the intuition that mere repetition can give rise to an implicit sense of commitment (and see Bonalumi et al., 2019 for evidence that people in general share this intuition). Similarly, one agent's investment of effort or other costs in a joint action may also give rise to an implicit sense of commitment on the part of a second agent. If Pam, for example, must walk up five flights of stairs to reach the balcony where she and Polly habitually smoke together, Polly's implicit sense of commitment may be greater than if Pam only had to walk down the hall. And indeed, this hypothesis has been supported by evidence from recent empirical research. Székely & Michael (2018), for example, reported that the perception of a partner's effort increases people's sense of commitment to joint actions, leading to increased effort, persistence and performance on boring and effortful tasks. Using the same stimuli

[2] See Michael, Sebanz & Knoblich (2016a) for a more detailed characterisation of the sense of commitment, in particular the requirement that X be an outcome, or goal, that B desires to come about and which requires the contribution of A to be successful.

as in Székely and Michael's (2018) study, Chennells & Michael (2018) found that participants were willing to invest more effort and also earned greater joint rewards when they perceived what they believed were cues of a partner's high effort than when they perceived cues which they were led to interpret as indicating a low degree of effort. Finally, research (Michaelet al., 2016b) has also shown that coordination in joint action can generate or enhance a sense of commitment. The rationale for this is that, when two agents coordinate their contributions to a joint action, they form and implement interdependent, i.e., mutually contingent, action plans. Each agent must therefore have – and rely upon – expectations about what the other agent is going to do. Indeed, the higher the degree of coordination, the more spatiotemporally exact must those expectations be. One important consequence is that an agent's performance of her contribution within a highly coordinated joint action expresses her expectations about the other agent's upcoming actions, as well as her reliance upon those expectations. This may generate social pressure on the other agent to perform her contribution in order to avoid disappointing the other's expectation and wasting her efforts.

Third, and more generally, on this account what motivates us to honour commitments is not a sensitivity to obligations per se but, rather, a desire to meet the (reasonable) expectations that others have of us, in particular insofar as they may be relying on those expectations (Dana et al., 2006; Heintz et al., 2015; Molnár & Heintz, 2016). While obligations may provide a focal point for what those expectations might be, they need not be the ultimate source of our motivation. Rather, expectations can be both a proximal, independent source of motivation *and* provide cues as to the possibility that a (directed) obligation is in place.

Support for this view comes from work in recent years, across domains as diverse as evolutionary theory and experimental economics and psychology, investigating the evolutionary origins of human cooperation (e.g. Henrich & Henrich, 2007; Nowak, 2012; Tomasello, 2009; Skyrms, 2004; West, Griffin, & Gardner, 2007). This has led to significant progress in specifying evolutionary mechanisms that are likely to have supported the evolution of cooperation in humans, including research into possible cognitive and motivational mechanisms that proximally support cooperation. For example, theoretical work on indirect reciprocity (Nowak & Sigmund, 2005) and on competitive altruism (Roberts, 1998) has inspired research devoted to illuminating the mechanisms by which people manage their reputations (Nowak & Sigmund, 2005; Fehr & Gächter, 2002; Andreoni & Bernheim, 2009; Rege & Telle, 2004). This research has provided evidence that reputation management may be subserved by prosocial preferences, such as a preference for fairness (Andreoni, 1990), an aversion to inequity (Fehr & Schmidt, 1999) and an aversion to disappointing others' expectations (Dana et al., 2006; Heintz et al., 2015). Reputation management need not, however, be the only evolutionarily cooperative reason for a person's desire to meet expectations others have of their future behaviour. For example, such expectations may also act as a cue that one is likely to interact with that agent in the future, encouraging directly reciprocal cooperative behaviour (Trivers, 1971), or that each has a stake in the other's wellbeing, as per Roberts' (2005) interdependence

hypothesis.[3] It's interesting to note that an appeal to a requirement to meet reasonable expectations also takes us right back to, and finds support in, Bratman's account of joint activity and the work of Thomas Scanlon, mentioned earlier, to which he refers when discussing possible sources of interpersonal obligation often present in cases of shared activity. Unsurprisingly, a focus on expectations also thus provides insight into the kinds of things philosophers seem to care about when they discuss obligations that arise in collective activity.

Overall, our broad view is that research into evolutionary mechanisms provides us with a good reason to think that an interaction partner's expectations of our future activity provides us with a proximal motivation that boosts our willingness to cooperate with them by guiding us as to what we should do - that is, to meet these expectations. In addition, though we present a descriptive rather than a normative account of the psychological sense of commitment in joint action, it does in fact have implications for a normative characterisation of the phenomenon of commitment. Specifically, the aforementioned hypothesis concerning proximal motivations to honor commitment provides a reason to expect people to honour commitments -- and thus also a justification for relying on them to do so.

The reason why the three differences between the simple view and the minimal account of a sense of commitment - namely, graded motivation, the role of situational cues, the desire to meet expectations - are particularly relevant to our discussion of commitment dissolution is that they undermine the importance of the act of release from a commitment. If one does not think of the act of giving an assurance as being decisive for generating a commitment, fully capturing everything that is expected to be fulfilled in meeting a commitment, or covering the myriad ways in which a commitment can be established, then it is also natural to think that the act of granting release is not decisive for dissolving commitments. Instead, this account suggests that, when we desire to be released from a commitment, we consider to what extent the other agent is expecting and relying on us to perform X. Any factors which imply a high degree of expectation and/or a high degree of reliance speak against requesting dissolution. Expectations come in degrees insofar as they can be associated with subjective probabilities. Reliance can be quantified as the sum of the costs that are incurred by the other agent if one fails to honour the commitment, and the opportunity costs irrespective of whether one honours the commitment. But what psychological processes underpin our judgments in such situations? While we have thus far a more psychological account of commitment that we think helps us explain behaviour better than the simple view, we still need to explain how, given that you have a commitment, you determine whether or not to dissolve them. To answer this,

[3] Another interesting direction of research posits a possibly proximal, more basic need to belong (Fernandez Castro & Pacherie, 2020) that grounds agents' motivation to act committed and makes their commitments credible. This builds on previous research exploring a possible role of social motivation in joint action - a source of motivation stemming from acting socially, with others, which is independent from the instrumental benefits expected to accrue from acting together (Godman, 2013; Godman et al., 2014). We believe minimal accounts like these and like ours offer an interesting and rich area for future research into commitments, their conceptualisation and how they motivate agents. We thank an anonymous reviewer for pointing us in the direction of this research.

we will need to take a step back and introduce a further bit of theoretical cognitive machinery -- namely, the concept of virtual bargaining.

## 4 Applying the sense of commitment framework to commitment dissolution

On the simple view, agents who wish to dissolve a commitment simply ask their partners to do so. However, while it is a general feature of these accounts (and the normative ethics they involve) that agents must justify why they should be released, they stop short of saying either exactly what a good reason for release might be, or how it is that agents decide whether they should ask to justify their request for release. Our focus here is on the latter question: What is the process by which an agent reasons whether or not to ask a partner to whom they are, or feel, committed if the commitment can be dissolved? Here we describe one game-theoretic possibility for how this process unfolds based on a novel theory of social interaction. Crucially, we believe the rational-choice model proposed meets several additional criteria we believe are required for the model to be credible given the application of the sense of commitment framework developed in the previous section.

### 4.1 A Virtual Bargaining approach

To understand how the psychological process underlying commitment dissolution works, we propose that one way involves people simulating their bargaining situations in order to guide their decision-making. A recent body of research has introduced and begun empirically testing a theory of *Virtual Bargaining* (Melkonyan et al., 2018; Misyak et al., 2014; Misyak & Chater, 2014) which provides us with a template for imagined or simulated interactions. Virtual Bargaining (VB) describes a rational-choice, psychological model of social interaction to explain and predict equilibrium states, given available actions and expected action-outcomes, of jointly interacting agents. In a nutshell, the model predicts that agents "should prefer the equilibrium that they would select if able to openly bargain" (Misyak & Chater, 2014) though, crucially, it is assumed that they do not actually communicate – that is, bargaining is, in this sense, virtual. Though a full explication of the theory of VB extends beyond the scope of this paper here, we are of the view that as an account of agents' psychology in joint action it holds great promise, for several reasons we detail below. Interestingly, it may also provide a possible new conceptual account of shared intentionality (Chater et al., 2021) which addresses several issues identified in existing accounts of the phenomenon.

Applying VB to the case at hand, we can generate a proposal for understanding the psychological processes involved when agents decide whether to seek commitment dissolution.

> Virtual bargaining proposal: we imagine virtually bargaining with our partner over whether or not to honour the commitment, and a decision is made based on the equilibrium state that would obtain.

This VB account has core features of most rational choice models applied to situations such as this, which involve multiple agents attempting to pursue their own goals in the knowledge that the outcomes of their actions may be mutually interdependent. It presents a theory of decision-making that includes an explanation of how agents incorporate predictions of others' actions - namely, by accommodating a game-like approach that is able to incorporate judgments of a partner's best response under conditions of uncertainty (J. F. Nash, 1950; J. Nash, 1951), which includes considerations for own and others' interests, but goes beyond to include affective states, desires to meet conventions, save face, act fairly, punish or reward partner behaviour, reciprocate positively or negatively in light of perceived intentions and outcomes, frame outcomes in terms of the team versus the individual, etc… (the literature is vast, but see for example Charness & Rabin, 2002; Charness & Levine, 2003; McCabe et al., 2003; Fehr & Schmidt, 2006; Schelling, 1980; Bacharach, 2006). This VB account also meets several important criteria we believe that such a proposal should meet, based on the minimal conceptualisation of commitment - the sense of commitment - presented in the previous section.

First, a VB account is sensitive to whether there is some exchange through which the commitment is generated in the first place. As we emphasised in the previous section, there are a great many instances in which we experience a sense of commitment without there ever having been any such exchange. A VB account incorporates what agents believe about their own and others' attitudes and, crucially, we can see whether or not a commitment has been generated in the first place as itself a possible result of a process of virtual bargaining between those who may be involved; i.e. as a practical matter, would we agree that one agent has effectively committed to another agent to the performance of a future action, though this commitment may not have been made explicit, or, perhaps more weakly, whether one agent has given another agent a reason to expect them to perform said action (e.g. by investing costs - including opportunity costs, as in Giulia may have passed up on the opportunity to take a daytrip with John to New York), and is relying on them to do so. VB also makes room for such expectations to be generated through previous interactions, seeing the virtual bargains agents make as the result of previous, overlapping real or virtual bargains (including e.g. conventions). Further, VB is able to incorporate important additional factors that may also be present and which could affect the equilibrium outcome, notably any power differential between those involved or differences in the extent to which each is relying on the other, or both.

Second, it makes a difference whether we can contact the other person or not. In other words, to simply renege without notifying the other person can lead them to suffer inconveniences or losses that they would not suffer if we did warn them (e.g. if Giulia winds up having to wait around in the park because James has failed to alert her that he is not going to make it for their planned stroll). As mechanisms for uncertainty reduction, commitments are more useful when our ability to monitor and sanction partners is limited; when our investments are done with incomplete knowledge of our partners' actions, for example. Often, then, reasoning over whether to dissolve a commitment or not is not done by engaging directly with our partners - VB is a theory of social interaction that is precisely focused on the sort of simulated bargaining that is required in such contexts.

Third, a VB account can include a simulation of what a real bargaining process might be experienced like. We believe this is a vital part of the phenomenology of approaching commitment dissolution in joint action: it is a feature of many social interactions that, just by bringing something up or asking for something, we reveal something to our interaction partners about what we want, what we find acceptable, what we think the other finds acceptable, etc. We would not, for example, raise a request to dissolve a commitment unless we felt it was justified (or unless the act itself is a signal). In bargaining over whether or not to dissolve a commitment, we must be cognisant of these pragmatics in virtue of which communicating a desire to engage in such bargaining can communicate meaning beyond the semantic content of what is said. For example, by asking to be released from a commitment, or engaging in a discussion about possible release, a committed agent may signal a lack of intrinsic preference for or interest in the activity they are expected to perform. If both James and Giulia assume that the other is in favour of their walking together, it may come as a surprise to Giulia should James ask her to abandon the activity. She may conclude that James wasn't, in fact, as keen on the joint activity as she had thought him to be. Taking this further, requesting release may, likewise, signal the existence of a divergence, or greater-than-expected divergence, between interests or 'types' of those involved. Giulia may, in future, avoid making walking plans with James given her judgement of his enthusiasm for walking. Finally, in certain contexts, requesting such a release may signal differences in or boost uncertainty around which social norms – including those often-unwritten rules which guide much social behaviour – are likely shared. Should James' request arrive with little time to spare, or once Giulia has invested significant costs (material, emotional or otherwise) in embarking on the joint activity, or in cases where such a request would be strange or frowned upon, Giulia might decide, for example that James' behaviour and impulsivity will likely extend beyond walking plans and into other plans they may have made, or might otherwise have made, for the future. There are multiple ways in which the act of requesting release or choosing to engage in a real or imagined bargain may undermine trust and commitment of those involved in a joint activity.

### 4.2 Virtual bargaining and commitment dissolution

Our VB proposal thus meets several important criteria that we believe should feature in a psychological model of agents involved in joint action. Returning to our virtual bargaining proposal, it will be useful to consider the responses it suggests to the four key questions we identified at the outset. Let's take each question in turn.

*What factors/principles do we appeal to in situations in which we may want to dissolve commitments?* Drawing upon the sense of commitment framework, the proposal is that we tend to feel committed (and thus reluctant to dissolve commitments) to the extent that we perceive the other agent to be expecting and relying on us to do our part. This means that any cue to the other agent's expectation and reliance will tend to speak against dissolution. For example, if James is aware that Giulia has gone to great trouble to polish her hiking boots in preparation for their outing to the park, this is likely to make him feel reluctant to dissolve the commitment. Note also that this response also indirectly addresses the concern raised above, in relation

to proposal (1), that an agent's accrual of opportunity costs after entering into the commitment makes a difference. Specifically, the reason for this is that the agent's willingness to incur opportunity costs indicates her expectation and constitutes an act of reliance. It also explains why it makes a difference whether we have been able to warn the other agent that we will not follow through on the commitment: this is because, if we cannot warn them, they will continue to expect and rely on us to perform our part, possibly accruing further opportunity costs.

*What are the reasoning processes we go through when considering whether to request release from a commitment?* First and foremost, the proposal implies an act of imagination by which we simulate the experience of bargaining with the other agent. This may be either a simulated interaction in which we inform the other agent that we will not follow through on the commitment (or ask them to release us), or it may be a simulated interaction in which we meet them and apologise for not having followed through on the commitment. Indeed, we may of course also simulate multiple interactions with the other agent. In any event, the simulation of future interactions may incorporate other processes, such as the application of theory of mind in order to predict how the other agent will respond. Similarly, it may involve affective forecasting (Wilson & Gilbert, 2003, 2005), e.g. in order to predict how one will feel about having failed to follow through on the commitment.

*How do we identify and assess the level of motivation in our commitments we have to others in the first place?* Insofar as the proposal draws upon the sense of commitment framework sketched above in Sect. 3, it suggests that we identify and assess the level of our motivation in our commitments by tracking others' expectations and reliance. Moreover, it also suggests that in practice this is often achieved by registering and responding to situational cues, such as an agent's investment of effort or other costs. In other words, it does not always occur through verbal exchanges.

*What do we actually care about when we talk about caring for commitments?* The current proposal, like the sense of commitment framework from which it draws, implies that what we actually care about when negotiating commitments is, among others, to maintain meaningful relationships with others and a solid reputation for ourselves. This is in contrast to the simple view which we started out from, according to which we care about meeting our obligations. Of course, our proposal does not deny that we often care about meeting our obligations, but it implies that we care about this *as a means of maintaining important relationships and our reputation*. Moreover, it also implies that we sometimes feel and act committed to do X in the absence of an obligation to do X -- namely, when doing X is important because some other agent is relying on us to do so (in particular insofar as our relationship with that other agent is important to us). And indeed, there is evidence that the sense of commitment can be decoupled from judgments about obligations (Michael, Sebanz, & Knoblich, 2016b).

### 4.3 (In-)Commensurability of costs and benefits

Something we have taken for granted thus far in analysing both the simple view and our own proposal is the assumption that individuals have the ability to identify, measure and compare – accurately or otherwise – different utility costs and benefits asso-

ciated with available actions. This raises the question of commensurability, given that various types of costs and benefits need to be compared with each other. Incommensurability, the absence of a common standard of measurement or judgement, though an issue for psychological and economic models of decision-making more generally and not unique to the context here (see for example Vlaev et al., 2011), nonetheless poses a problem for our view, which extends beyond the simple view by including a possibly diverse and wide range of factors that affect individual judgements about whether or not to meet commitments. It is not immediately clear how, for example, James would weigh up the enjoyment he gains from spending a few more hours in bed against that of going for a walk, let alone when contextual factors, like weather or sleep deprivation, are taken into account.

In situations of collective activity, the problem of incommensurability also arises in another form; namely, it is not obvious how an agent compares her own costs and benefits not only with each other but also with those of her partner. There is limited empirical work in this area, which has produced mixed findings (see for example Apps et al., 2016; Michael et al., 2020). In the absence of relevant research, we must note that it is unclear how James should or would compare Giulia's enjoyment of companionship on a walk with his own preference for remaining at home.

A deeper investigation of this issue, while valuable given our research question, is beyond the remit of this paper. For discussion purposes here, it suffices to say that both intra- and inter-personal benefit and cost utility comparisons – across multiple action-options and indeterminate action-outcomes – poses a problem for any psychological model that involves individuals making judgements about which course of action is optimal. It is therefore not straightforward how we evaluate relevant factors and arrive at a correct judgement[4]. Yet, as the the discussion in Sect. 3 showed, research suggests that an agent's commitment towards a partner does, in fact, appear to be influenced by relevant factors such as costs previously invested in a joint activity, how reasonable a partner's expectations are, the level of coordination between interacting agents, and the extent to which a partner is relying on an agent.

One possible response to the problem of incommensurability, a response which still allows for the fact that agents' actual and perceived commitment do not appear to be independent of certain relevant factors (and which the simple view thus says nothing about), is to make room for a kind of metacognitive judgement that is not metarepresentational (Proust, 2007, 2010). On this view, agents take relevant factors into account as inputs to a simulation, and experience an emotional response for each action option. They then base their decision upon these emotional responses, responses which are, importantly, comparable across different action-options. Emotions may thus act as a 'common currency' when comparing different action options, an idea that parallels a view of emotions recently proposed as a solution to cases of incommensurability when individuals are required to weigh up various costs – such as effort versus monetary costs – when making decisions. In such cases, emotions attached to outcomes are converted to reward in the brain and the amount of reward associated with the combination of costs and benefits is what informs the decision

[4] As an anonymous reviewer pointed out, it's plausible that in many situations agents are actually unable to arrive at a judgement about whether or not to honour a commitment.

between different actions. Emotions thus act as a neural common currency for choice (see Levy & Glimcher, 2012; Sescousse et al., 2015).

## 5 Conclusions and Implications

At the outset, we posed the question: what happens psychologically when we consider whether or not to follow through on commitments in instances in which we find ourselves tempted to abandon them? In particular, we set out to develop an account which would incorporate answers to the following four key questions:

1) What factors/principles do we appeal to in situations in which we may want to dissolve commitments?

2) What are the reasoning processes we go through when considering whether to request release from a commitment?

3) How do we identify and assess the level of motivation in our commitments we have to others in the first place?

4) What do we actually care about when we talk about caring for commitments?

We started out by considering what we called 'the simple view': when we want to be released from commitments, we need only ask to be released. If the person to whom we are committed releases us, we are free; if not, then we remain committed. The simple view follows from standard approaches to commitment in the philosophical literature (Bratman, 2014; Gilbert, 2009; Searle, 1969; Shpall, 2014), which say much about how commitments are generated but little about how they motivate agents or how they are dissolved, and it is this view which has informed the limited empirical research that has been undertaken so far concerning the dissolution of commitments (Kachel & Tomasello, 2019).

Having identified several problems with this simple view, we developed our own proposal, based on the sense of commitment framework. This proposal suggests that, when we desire to be released from an interpersonal commitment, we consider to what extent the other agent is expecting and relying on us to perform our part. Any factors which imply a high degree of expectation and/or a high degree of reliance speak against requesting dissolution and for fulfilling the commitment. Expectations come in degrees insofar as they can be associated with subjective probabilities. Reliance can be quantified as the sum of the net costs that are incurred by the other agent if one fails to honour the commitment, and the opportunity costs irrespective of whether one honours the commitment.

We then showed how our proposal can provide a basis for answering the four key questions identified above. In answer to the first question, our proposal is that the factors/ principles we appeal to when considering commitment dissolution are those which affect our sense of commitment towards a partner, those influencing how committed we feel towards the other agent - and thus our reluctance to dissolve the commitment - to the extent that the other agent is expecting and relying on us to do our part. This means that, notwithstanding the presence or absence of explicit obligations, any cue to the other agent's expectation and reliance will tend to speak against dissolution.

In response to the second question, our proposal implies an act of imagination by which we simulate the experience of interacting with the other agent. Agents may, indeed, simulate multiple different interactions with the other agent, including, for example, following through on the commitment, requesting release, apologising or not following through, etc. Simulations of future interaction may incorporate other processes as well, such as the application of theory of mind to predict another agent's response, or affective forecasting to predict what different interactions might be experienced like.

Answering the third question, our proposal, drawing on the sense of commitment framework, suggests that we identify and assess the level of our motivation in our commitments by tracking others' expectations and reliance. Moreover, this need not always occur through verbal exchanges; in practice, this may be achieved by, for example, registering and responding to situational cues, such as an agent's investment of effort or other costs.

Finally, our proposal implies that what we actually care about when negotiating commitments is to maintain meaningful relationships with others and a solid reputation for ourselves. This contrasts with the simple view from which we began, according to which we care about meeting our obligations. While we do not deny that we often care about meeting our obligations, our proposal is that we care about this to the extent that this helps us to maintain our relationships and our reputation. Indeed, our proposal implies that our sense of commitment might, in fact, be decoupled from judgements about obligations, such that we sometimes feel and act committed to do X in the absence of an obligation to do X -- namely, when doing X is important because some other agent is relying on us to do so (in particular insofar as our relationship with that other agent is important to us).

The approach our proposal takes - in which the extent to which we sense we are committed is dependent on and graded by the extent to which we perceive the other agent to be expecting and relying on us - invites us to think of dynamically changing, imperfectly aligned (between ourselves and other agents) interests, as being the norm, and therefore also to think that we constantly monitor and re-evaluate our commitments in light of changing environments. As John le Carré put it, in reference to Karla's resolution never again to communicate via radio after the debacle in Delhi: 'Like most promises, it was subject to review' (Carre, 2002; pg. 305). The importance of reassessing commitments, traditionally characterised as promise-like structures, in light of changing environments and changing preferences is something that the sense of commitment framework motivates us to think is important: this framework places our sensitivity to each other's expectations at center stage, and expectations change dynamically. This is in contrast to existing accounts, which are focused on agreements and obligations which, once made, remain in place and unchanged until dissolved. More generally, the sense of commitment framework gives us reason to be skeptical about the central role which these accounts accord to obligations. In particular, by doing so, they elide distinctions among cases in which commitments matter a great deal to the individual and cases in which they do not. Thinking in terms of obligations does not enable us to see what we actually care about when we care about commitments, nor why we are more motivated to follow through on our commitments in some cases than in other cases. In sum, the question of how we dissolve

commitments, or whether we ask to be released, reveals something about the psychological and phenomenological complexity of these situations, which is not addressed by traditional accounts.

In this connection, it bears emphasising that the account developed here is primarily descriptive rather than normative. However, the account does in fact have implications for a normative characterisation of the phenomenon of commitment. The reason for this is that, by spelling out why people tend to honor their commitments (namely to avoid disappointing others' expectations), we have also identified the reasons why it is sometimes justified to expect and rely upon people to honour their commitments.

The current proposal also provides new impulses for research on the development of the understanding of, and sensitivity to, commitment in children. It would be valuable, for example, to probe at what age children develop a proficiency in distinguishing between good and bad reasons for abandoning commitments (Bonalumi et al., n.d.; Kachel & Tomasello, 2019; Michael & Székely, 2018). Moreover, it provides a platform for investigating how individuals with pathologies of social cognition, such as borderline personality disorder, may differ in their assessment of cases in which someone does or does not follow through on a commitment (Ooi et al., 2018). And indeed, it also generates novel, testable predictions about what happens when healthy adults consider whether or not to follow through on their commitments. Specifically, it predicts that people might engage in future-directed mental time travel to simulate their next encounter with the other party to the commitment -- presumably all the more vividly in difficult cases.

A further virtue of this account is that it builds in space for cultural differences. Instead of attempting to lay out specific principles governing the dissolution of commitments, based on fixed ideas of the types of obligations which are generated and the circumstances under which they are maintained, we instead sketch a procedure in which different principles and factors may figure according to cultural context (and those principles and factors are likely to be weighted differently depending on the cultural context). Identifying these differences, and linking them to more general cultural differences, is an important avenue for further research.

In sum, we learn a great deal about the substance of commitment by looking at instances in which we come to reconsider whether or not to honour our commitments. In such instances, as we have seen, leading accounts fail to provide a phenomenologically adequate explanation of the processes that unfold when we decide whether it's appropriate to break a commitment and when we imagine how to break it in the right way.

**DECLARATIONS**.

**Author contributions** Both authors contributed equally in all aspects to the submission.

**Funding** This research was supported by a Starting Grant awarded to John Michael by the European.

**Research Council (Nr. 679092, Sense of Commitment).**

**Availability of data and material** Not applicable.

**Code Availability** Not applicable.

## Declarations

**Conflict of interest** There authors declare no conflicts of interest.



## References

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100(401), 464–477

Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5), 1607–1636

Apps, M. A. J., Rushworth, M. F. S., & Chang, S. W. C. (2016). The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others. *Neuron*, 90(4), 692–707. https://doi.org/10.1016/j.neuron.2016.04.018

Astington, J. W. (1988). Children's understanding of the speech act of promising. *Journal of Child Language*, 15(1), 157–173

Bacharach, M. (2006). *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton University Press

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

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

Bonalumi, F., Isella, M., & Michael, J. (2019). Cueing Implicit Commitment. *Review of Philosophy and Psychology*, 10(4), 669–688. https://doi.org/10.1007/s13164-018-0425-0

Bonalumi, F., Siposova, B., Christensen, W., & Michael, J. (n.d.) (Eds.). Should I stay or should I go? Three-year-olds' sensitivity to appropriate motives to break a commitment. *Under Review*

Carre, J. (2002). le. *Smiley's People*. Simon and Schuster

Charness, G., & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, 117(3), 817–869. https://doi.org/10.1162/003355302760193904

Charness, G., & Levine, D. I. (2003). The Road to Hell: An Experimental Study of Intentions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.437870

Chater, N., Zeitoun, H., & Melkonyan, T. (2021). *The paradox of social interaction: shared intentionality, we-reasoning and virtual bargaining*. Draft Manuscript

Chennells, M., & Michael, J. (2018). Effort and performance in a cooperative activity are boosted by perception of a partner's effort. *Scientific Reports*, 8(1), 15692. https://doi.org/10.1038/s41598-018-34096-1

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, 100(2), 193–201. https://doi.org/10.1016/j.obhdp.2005.10.001

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868), 137–140

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817–868

Fehr, E., & Schmidt, K. M. (2006). The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories. In S.-C. Kolm & J. M. Ythier (Eds.), *Handbook of the Economics of Giving, Altruism and Reciprocity* (Vol. 1)

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

Gilbert, M. (1990). Walking Together: A Paradigmatic Social Phenomenon. *Midwest Studies In Philosophy*, 15(1), 1–14. https://doi.org/10.1111/j.1475-4975.1990.tb00202.x

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

Godman, M. (2013). Why we do things together: the social motivation hypothesis. *Philosophical Psychology*, 26(4), 588–603

Godman, M., Nagatsu, M., & Salmela, M. (2014). The social motivation hypothesis for prosocial behaviour. *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

Heintz, C., Celse, J., Giardini, F., & Max, S. (2015). Facing expectations: Those that we prefer to fulfil and those that we disregard. *Judgment and Decision Making*, 10(5), 14

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

Kachel, U., & Tomasello, M. (2019). 3- and 5-year-old children's adherence to explicit and implicit joint commitments. *Developmental Psychology*, 55(1), 80–88. https://doi.org/10.1037/dev0000632

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, 22(6), 1027–1038. https://doi.org/10.1016/j.conb.2012.06.001

Mant, C. M., & Perner, J. (1988). The child's understanding of commitment. *Developmental Psychology*, 24(3), 343–351

McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization*, 52(2), 267–275. https://doi.org/10.1016/S0167-2681(03)00003-9

Melkonyan, T., Zeitoun, H., & Chater, N. (2018). Virtual Bargaining as a Formal Account of Tacit Agreements. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3168669

Michael, J., Gutoreva, A., Lee, M. H., Tan, P. N., Bruce, E. M., Székely, M. … Ludvig, E. A. (2020). Decision-makers use social information to update their preferences but choose for others as they do for themselves. *Journal of Behavioral Decision Making*, 33(3), 270–286. https://doi.org/10.1002/bdm.2163

Michael, J., & Pacherie, E. (2015). On commitments and other uncertainty reduction tools in joint action. *Journal of Social Ontology*, 1(1),89–120https://doi.org/10.1515/jso-2014-0021

Michael, J., Sebanz, N., & Knoblich, G. (2016a). The Sense of Commitment: A Minimal Approach. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01968

Michael, J., Sebanz, N., & Knoblich, G. (2016b). Observing joint action: Coordination creates commitment. *Cognition*, 157, 106–113. https://doi.org/10.1016/j.cognition.2016.08.024

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

Misyak, J., & Chater, N. (2014). Virtual bargaining: A theory of social decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655), 20130487–20130487. https://doi.org/10.1098/rstb.2013.0487

Misyak, J., Melkonyan, T., Zeitoun, H., & Chater, N. (2014). Unwritten rules: Virtual bargaining underpins social interaction, culture, and society. *Trends in Cognitive Sciences*, 18(10), 512–519. https://doi.org/10.1016/j.tics.2014.05.010

Molnár, A., & Heintz, C. (2016). Beliefs about people's prosociality: Eliciting predictions in dictator games. *CEU: Department of Economics - Working Paper*, 19

Nash, J. (1951). Non-Cooperative Games. *Annals of Mathematics*, 54(2), 286–295. https://doi.org/10.2307/1969529. JSTOR

Nash, J. F. (1950). Equilibrium Points in N-Person Games. *Proceedings of the National Academy of Sciences of the United States of America*, 36(1), 48–49

Nowak, M. A. (2012). Evolving cooperation. *Journal of Theoretical Biology*, 299, 1–8. https://doi.org/10.1016/j.jtbi.2012.01.014

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature: Reviews Vol*, 437, 1–8

Ooi, J., Francová, A., Székely, M., & Michael, J. (2018). The Sense of Commitment in Individuals With Borderline Personality Traits in a Non-clinical Population. *Frontiers in Psychiatry*, 9, 519. https://doi.org/10.3389/fpsyt.2018.00519

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for metacognition? *Synthese*, 159(2), 271–295. https://doi.org/10.1007/s11229-007-9208-3
Proust, J. (2010). Metacognition. *Philosophy Compass*, 5(11), 989–998. https://doi.org/10.1111/j.1747-9991.2010.00340.x
Rege, M., & Telle, K. (2004). The impact of social approval and framing on cooperation in public good situations. *Journal of public Economics*, 88(7), 1625–1644
Roberts, G. (1998). Competitive altruism: from reciprocity to the handicap principle. Proceedings of the Royal Society of London. Series B: Biological Sciences, Vol. 265, Issue 1394, 427–431
Roberts, G. (2005). Cooperation through interdependence. *Animal Behaviour*, 70(4), 901–908
Scanlon, T. (1998). *What we owe to each other*. Belknap Press of Harvard University Press
Schelling, T. C. (1980). *The Strategy of Conflict*. Harvard University Press
Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (626 vol.). Cambridge University Press
Sescousse, G., Li, Y., & Dreher, J. C. (2015). A common currency for the computation of motivational values in the human striatum. *Social Cognitive and Affective Neuroscience*, 10(4), 467–473. https://doi.org/10.1093/scan/nsu074
Shpall, S. (2014). Moral and Rational Commitment. *Philosophy and Phenomenological Research*, 88(1), 146–172. https://doi.org/10.1111/j.1933-1592.2012.00618.x
Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge, UK: Cambridge University Press
Székely, M. & Michael, J. (2018). Investing in commitment: Persistence in a joint action is enhanced be the perception of a partner's effort. Cognition 174, 37–42
Tomasello, M. (2009). *Why we cooperate*. Cambridge, MA: MIT Press
Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly review of biology*, 46(1), 35–57
Vlaev, I., Chater, N., Stewart, N., & Brown, G. D. A. (2011). Does the brain calculate value? *Trends in Cognitive Sciences*, 15(11), 546–554. https://doi.org/10.1016/j.tics.2011.09.008
West, S. A., Griffin, A. S., & Gardner, A. (2007). Evolutionary explanations for cooperation. *Current Biology*, 17, R661–R672. https://doi.org/10.1016/j.cub.2007.06.004
Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (35 vol., pp. 345–411). Elsevier Academic Press
Wilson, T. D., & Gilbert, D. T. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science*, 14(3), 131–134. https://doi.org/10.1111/j.0963-7214.2005.00355.x



## Authors and Affiliations

**Matthew Chennells[1] · John Michael[2]**

✉ Matthew Chennells
m.chennells@warwick.ac.uk

John Michael
johnmichael.cogsci@gmail.com

[1] Department of Philosophy, University of Warwick, Coventry, United Kingdom

[2] Affiliated with Department of Cognitive Science, Central European University, Vienna, Austria