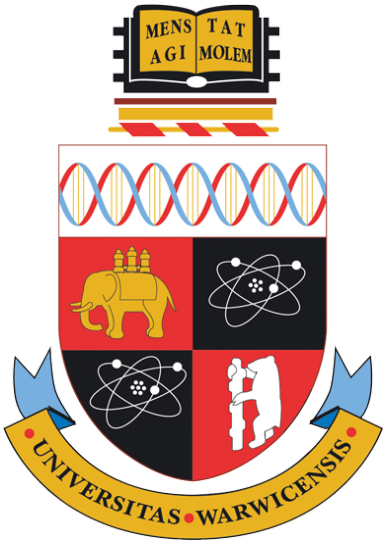
WARWICK
THE UNIVERSITY OF WARWICK



(In)compatible:

Shared Intention, Ordinary Uncertainty and Social Commitment

Matthew Chennells

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Philosophy

University of Warwick,

Department of Philosophy

29 April 2023

Contents

Acknowledgments

The main reason my thesis is complete is thanks to my family, and my greatest gratitude is towards them. My mum and dad's support, generosity, love and care is always immense and unconditional, and it never changed throughout this long project. I am privileged in many ways, but knowing they're there in whatever way and whenever I need it is the biggest one. My brother, for the instant connection, love, humour and thoughtfulness that we share and will hopefully continue to deepen and explore. And my aunts, uncles and cousins, both in SA but importantly in the UK, for ensuring I'm loved, looked after and always at home.

I am also grateful to my supervisors. To Steve Butterfill first and foremost, for his patience, incisive mind, unflappable support and constant care for my work and wellbeing. To Nick Chater, for his constant ideation and encouragement to be curious about things we take to to be obvious. And to John Michael, who taught me many things, always treated me as an equal and who, for a time, was the most enabling person with whom I've ever worked. A big debt of thanks goes to Tina Kiefer, whose jobs kept the wolves at bay, who trusted me with new skills and areas of research and with whom it's always a pleasure to interact.

I want to thank the faculty at Warwick Philosophy, for continued departmental funding and support and for taking a chance on me in the first place, to the initial crew who warmly welcomed me there—Alex, Jack, Tristan and Maria, and the SoC lab members—and certain special people who walked different stages of my Warwick journey with me—Silba, Lea, Francesca and Barbora. I'm also thankful for widening my family to include new homies—Lorenzo and Giulia—and those who embraced me in a lil' fam—Brigid, Simon, Will, Ida and Isa. Several more wonderful peers have given me so much—Arianna, Amul, Melissa, Michael, Simon and even a new witches coven—and several faraway friends who are always in my heart—Sammy, Anton, Dov, Leachy and Cara. To the Ramblers—Tom, Henry, Ref plus the rest of the Perkins family—for unbounded generosity, joy and connection. And to Rabi, my partner on many journeys, past present and undoubtedly future.

Finally, to Giulia, my favourite person and Peewee-tripper: thank you for all the things. I can't express here how much joy you bring to every moment spent together and how crucial your support has been to me—but I hope I can show you every day.

dum vivimus, vivamus

Declaration of Authorship

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The thesis was conducted under the supervision of Prof. Steve Butterfill and Prof. Nick Chater.

Parts of **Chapter 7** have been published in re-worked form in:

Chennells, M., & Michael, J. (2022). Breaking the right way: A closer look at how we dissolve commitments. *Phenomenology and the Cognitive Sciences—Special Issue: The Phenomenology of Joint Action: Structure, Mechanisms and Functions.* https://doi.org/10.1007/s11097-022-09805-x

Abstract

This thesis examines if traditional accounts of shared intention can explain how it works when there's doubt about partner intentions to contribute. Can I plan an activity with you and rely on you in ways required to share intentions when I cannot take for granted that you're predisposed to cooperate? This question, which reflects some everyday instances of social interaction, raises issues about whether and how shared intention can be possible when there is uncertainty about partner motivations and intentions. I address two potential objections to the idea. First, that the lack of common knowledge precludes sharing intentions. I propose a role for common knowledge in settling matters in shared intention and argue that despite its absence it's still possible that this function is fulfilled. Second, that this overly weakens a belief requirement on shared intention. I draw on research on individual agency to argue that uncertainty about joint action success needn't undermine shared intention. Neither challenge therefore precludes shared intention under motivational uncertainty. I address a new concern that under this uncertainty a joint settling condition is not met. I use two authors' accounts of shared intention to highlight the importance of this requirement, and show that, despite very different methodological approaches, both rely on similar background assumptions to explain how the requirement is met. However, I argue that the presence of attractive alternatives means there's no prima facie reason why these assumptions hold, presenting a theoretical problem. To solve it, I propose using the notion of mutual, social commitment—a popular tool to reduce motivational uncertainty between agents in joint activity—to ground reliance on others and explain how agents jointly settle matters in these contexts. This would, though, impact how we conceptualise the general connection between intentions and commitments. Despite this potential, I argue that social commitments lack credibility in traditional accounts of shared intention, a particularly acute problem in situations with motivational uncertainty. Finally, I outline a view of interpersonal commitment which responds to certain of the issues faced by traditional theories of commitment. I present a framework for a minimal psychological sense of commitment which arises in joint activities and explain how it allows individuals to settle matters in shared activity when there is uncertainty about intentions. I provide empirical support for the theory and address two concerns it faces. I conclude by showing how the sense of commitment can be a basis for solving the problem of how agents can jointly settle matters under conditions of substantial motivational uncertainty.

Preface

This thesis is a defence of the broad scope of shared intention and its generalisability, exploring contexts often overlooked in analysis of the phenomenon. In particular, it is a defence of shared intention in situations where it's not guaranteed that each of us will be motivated to play our part or make our contribution; where one or more of us is uncertain about what our partners intend because we cannot take for granted that we all share beliefs, desires and intentions, or that we are all predisposed to cooperation and collaboration. With this in mind, it also defends one form of social normativity which helps explain shared intention in contexts with this motivational uncertainty: namely, an interpersonal commitment which provides the basis for mutual reliance, settling and planning. The thesis is, simultaneously, a critique of certain accounts of shared intention in terms of whether they can credibly explain where these commitments come from and why individuals are motivated to meet commitments they have made. And finally, it ends in a proposal for a theory of a psychological sense of commitment, one which addresses some of the shortcomings of traditional accounts while providing a richer understanding of how we experience social commitments.

The thesis also presents a view of how individuals experience sharing intentions. First, it emphasises that doing things with others is a distinctly first-personal phenomenon, and that theoretical accounts of the subject should reckon with this. If we take seriously the idea that the way we come to know what others believe and intend is different to how we arrive at our own beliefs and intentions, then we need to account for the processes and mechanisms by which we come to form beliefs about and rely on our partners. Moreover, this means we need to acknowledge that there may be socially-relevant reasons which make it awkward, difficult or even undesirable to know for sure what others believe and intend. Second, by focusing on agents' perceptions that their partners might *not* intend to perform their part, it pinpoints what it is that gives certain collective activities their uniquely 'shared' character. It may seem strange to explore situations of potential breakdown in order to understand why we often act and work together so well and so easily. But it's in reflecting on why we are motivated to participate when it's not in our apparent interests to do so that, I think, we better understand what it is that binds us, and which makes the fact of our coordination and cooperation more fascinating.

CHAPTER 1

Introducing Uncertainty About Intentions in the Context of Shared Intention

It's a remarkable aspect of human society that we're able to coordinate and cooperate across a wide range of environments, in the context of diverse individual interests and often with only limited information of one another. One answer as to why shared activities are often successful focuses on explaining how those who are part of a group 'share intentions' in ways which support the success of their joint activities. How several individuals engage in a shared activity is thus partly understood with reference to how they share intentions to do so. Analogous to the role performed by intention in the case of solo activity, when acting together shared intention provides the normative and psychological infrastructure necessary for supporting shared activity. In short, by sharing intentions, individuals can act together intentionally and work together to realise collective goals.

What it means to act together intentionally is, naturally, heavily debated. We can, however, get a general idea of what leading accounts require of agents who act on a shared intention. Pacherie (2013), summarising Butterfill (2012), proposes four criteria:

1) Agents must be aware that they are not acting individually and think of their action as their contribution to a joint goal,

2) agents must be aware of (at least some of) the other contributing agents as intentional agents,

3) agents must act in part because of their awareness of others' agency and of the joint-ness of their goal, and

4) agents must be aware of at least some of the other agents' attitudes concerning the joint action.

What each person intends and believes, along with how this is arranged in relation to the intentions and beliefs of other group members, helps coordinate the group's ongoing

activities, facilitate their planning for future group activities and provide a platform for their bargaining and negotiation over who will do what, when and how. Importantly, it establishes that their actions are performed together—and it is this sense of acting in concert that theoretical accounts of shared intention try to capture. Taken together, these criteria, if met, help explain what sets shared intentional action apart from other types of collective activity, including those in which actions merely happen to be performed in parallel while agents work independently towards a shared goal, or performed together only with the intention of using others as a means to desired ends. These criteria are noticeably broad. They are also relatively non-controversial given the types of collective activity authors tend to pick to describe and examine instances of shared intention. The circumstances in the examples used generally don't explicitly defend background assumptions of cooperativity; namely, they assume that agents who might form an intention to act together are motivated only by the goal of acting together. These cases thus naturally lend themselves to the satisfaction of these conditions. But we should not take this background assumption for granted.

1.1 Collective activities and uncertainty about partner intentions

Given most authors' focus on shared intention for collaboration, one is forgiven for thinking that situations involving mutual interest encompass all that we do together. Yet there are many cases in which you and I might be in the process of doing something together, or anticipate doing something together, where I have quite a bit of uncertainty about your intentions. In many situations, for example, I might be fully aware that a variety of tempting alternative options are available to you, temptations which may wax and wane, adjusting how tempting things are to you and how tempted you might be to abandon our joint activity in their favour. Consider the following example I'm calling BEACH:

> *Mya and Iva each intend that they go to the beach this afternoon (say they phoned each other in the morning and set a time to meet in mid-afternoon, near a lifeguard tower that they both know). Mya said he would bring beach bats for them to play and Iva mentioned she would bring diving gear for the both of them. Imagine, however, that shortly before leaving to catch the bus to the beach, Mya sees on Instagram that Iva's favourite team is playing a football match whose kickoff time coincides with their agreed meetup time. Mya is struck by the fact that it's very likely that Iva won't join him at the beach, given what he knows about the latter's football obsession. Mya nonetheless catches the bus to the beach and walks to the lifeguard tower, arriving a*

few minutes early. Their agreed time arrives and goes, but just before Mya is about to give up, he sees Iva running towards him with their gear, apologising for being late. They meet up and enjoy a wonderful, sunny day at the beach together.

I believe that BEACH illustrates a feature—uncertainty about an ostensive partner's intentions—that is often present, to a greater or lesser degree, when we aim to engage in joint activities with friends, romantic partners, work colleagues and strangers. Of course, what we are uncertain about and the extent to which we are uncertain will depend on a range of factors, but it remains the case that sometimes we intend to act together with others even though we may have good reason to be uncertain that they intend likewise, irrespective of whether or not they do actually intend as we do. It would therefore seem possible for there to be successful shared activity in the presence of uncertainty about a partner's intentions. But can this shared activity be intentional? Can individuals share an intention when one or more of them is uncertain about what their partner intends now or will come to intend in the future? A cursory glance at the criteria above suggests this is not a straightforward answer. In BEACH, the first two are met, and so is the third if it's the case that both Mya and Iva continued to see their going to the beach as an intentionally joint exercise. The fourth criterion, though, is only partially met: Mya does not have perfect information of Iva's intentions, so whether it's met will depend on what is meant by agents being 'aware' of others' attitudes.

This case is an interesting one to examine. Up front, I note that I'm not insisting that there is shared intention in BEACH and this is not part of the premise. BEACH is therefore not supposed to be a counterexample to various accounts of shared intention; that is, to show what they lack or what must be modified to change them. An obvious issue, otherwise, is that a reader may not accept that we have a good reason to think there should be shared intention in cases like BEACH, that the uncertainty about another's intention simply precludes the possibility of there being shared intention at all. We would, in this case, merely be trading intuitions about what count as cases of shared intention—and we would get nowhere by simply throwing more examples of uncertainty about intentions into the picture.

However, I want to insist that you cannot rule out from the beginning the possibility that there is shared intention in the face of uncertainty about intentions. There are several reasons I'm taking this stance.

First, these cases look like they have fairly rich forms of social interaction despite the uncertainty present. More specifically, in briefly reflecting on the four criteria above we saw that BEACH looks like it has a lot, though not all, of the features of shared intention. Moreover, we can imagine situations very similar to BEACH in which there clearly is shared intention, except for that they don't involve this kind of uncertainty. It's useful, therefore, to see whether or not we can use existing theories to characterise the cases in question, or whether we need a different kind of theory—especially given the wealth of analysis of and insight into social interaction in the literature on shared intention.

Second, having an account of shared intention with the generality to include cases where there is uncertainty would be a positive thing, given how theories of shared intention are applied. For example, an author wanting to explore the function of the planning capacities of our lives should find it strange that as soon as uncertainty about partner intentions enters then these capacities no longer play a similar role—especially given that uncertainty about the future is a realistic feature of our lives. So, on the face of it, we shouldn't say in advance that there isn't shared intention in BEACH (though we're not saying that there is). An account that allows us to deal with situations involving uncertainty would have the kind of greater generality that seems to be beneficial, given the sorts of aims that authors have when they're constructing theories of shared intention.

Third, there are good reasons to think that uncertainty about partner intentions is a relevant and interesting question. An initial response to the problem posed in BEACH might be that resolving the uncertainty is a straightforward matter. For example, why doesn't Mya simply pick up the phone to see whether Iva is still on her way? The simple matter of allowing for communication would, it might be argued, address these issues right away, and allow us to either avoid the problem in the first place or see situations where such communication does not take place as simply inefficient instances of social interaction, either uninteresting or irrelevant for thinking about a general and minimal characterisation. This, though, would ignore the fact that there are often practical reasons why, in many cases, we can't simply resolve questions about what others intend or believe. We may lack the time or ability to communicate. We may also lack the financial means to do so, if there are material costs involved (data or airtime is, for example, expensive in many parts of the world).

It would also ignore the possibility of psychological costs involved in trying to reduce this uncertainty. This last point is the primary focus of this thesis and on which I'll expand in

chapters to come. For now, it's not hard to imagine that there are times when communication is either awkward—admitting that I felt lazy or sad and stayed in bed rather than join you at the beach; or admitting anxieties about whether you will stand me up—risky—unsure whether you will respond with fury or disappointment to my change of plans, or what you will think of my character going forward—or otherwise emotionally costly. I might also find it undesirable to resolve your uncertainty about what I intend if the state of imperfect information works in my favour. I might prefer, for example, to keep you as a fallback option in case my first-choice date goes poorly. Or, as theories of indirect speech point out, I may prefer being ambiguous if it reduces the chance of my being sanctioned should I speak too plainly (consider, for example, sexual innuendo or the subtle offer of a bribe).

In short, the point that it is *motivational* rather than practical uncertainty makes it more relevant for thinking about shared intention. It is sometimes better to live with rather than to try to resolve all the uncertainties that emerge in our shared activities because the latter is costly to do. This prompts us to accept that uncertainty is not, after all, an undesirable nor temporary feature of many instances of social interaction.

Ultimately, there is a rich literature on shared intention which is worth exploring in contexts in which there's uncertainty about intentions—and how it does so is an open question for the thesis. This will involve looking at the ways in which one or another theory might motivate including or excluding cases where there's motivational uncertainty, seeing whether or not we get a good theoretical justification for carving up theories of shared intention into cases where there is and isn't, respectively, this kind of uncertainty. What we need to do is to see whether we actually have a principled reason from accounts of shared intention to rule out cases involving motivational uncertainty, and when we look initially at Pacherie / Butterfill's four points above, it looks like it's just one clause of four that is, on the face of it, missing. And the reason that existing approaches exclude—often only implicitly—cases where there's uncertainty doesn't seem to justify actually excluding them; instead, it seems like an artefact of the theory and the biases of people structuring the theory rather than reflection of the true nature of the phenomenon.

1.2 Outlining the problem posed by uncertainty about intentions

How agents can intend a joint activity though they might be uncertain is not well established or explained in the literature. In the chapters to come I'll argue this in detail, but for the

moment we can get to the crux of the problem by first briefly reflecting on the role intention plays in practical agency. Practical reasoning, broadly understood, refers to "an inferential process through which new intentions are formed or old ones modified. According to this view, we resolve through reasoning the question of what we are going to do" (Broome 2013, in Wallace, 2020). Three key features stand out in this definition: first, that our practical reasoning involves *deliberation*; second, that, through deliberation, we *resolve* what to do; and third, that we are moved to *form the intention* to act on that which we have settled. This is typically taken to be guided by specific norms of rationality, such as principles of instrumentality—we should intend the means necessary to achieve our intended ends—and belief coherence and consistency. These are basic sources of normativity, justified by, for example, reflecting on how reasoning within these constraints typically helps us get what we want—this is the standard by which we judge them.

That one necessarily *resolves* the question about what to do highlights the special normative question that lies at the heart of practical reasoning: namely, that this form of reasoning tries to answer what I should *do*, or what is best for me to do, given a set of available alternative actions (Wallace, 2020). But it's not obvious that having resolved what's best, I decide to actually do it. As John Broome (1999) discussed, principles like these needn't generate reasons for action: one could instead abandon the ends rather than following through with the means judged necessary. The problem then is that practical reasoning for action looks very much like theoretical reasoning for belief: in other words, absent the capacity to modify intentions, practical reason looks like it's "practical only in subject matter, but not in its issue" (Wallace, 2020). We therefore need a bridge between the output of practical reasoning and acting. One idea is to assume additional pressures to conform to, for example, the instrumental principle and the other norms mentioned. In later work, Broome (2013), for example, proposes another requirement of rationality, *Enkrasia*, which "requires of you that, if you believe you ought to F, you F" (Pauer-Studer, 2014). This is a much larger topic than I have space to discuss here. For my purposes, I take these basic norms of practical rationality as non-controversial, though I do address them in the context of shared agency in the chapters to come.

It follows from the *Enkrasia* requirement that I intend to do what I believe I ought to do. Practical reasoning leading to intention formation thus resolves not only the question of what I should do but what I *will* do. This comes with ensuing implications, including some

plausible principles and conditions of intention that will be familiar to the reader, which I'll explicate in the upcoming chapters but for the moment take at face value:

- *Settling principle (S–P):* my intention resolves deliberative questions about what I will do.
- *Control principle (C–P)*: I regard my intended action as up to me to perform when the time comes.
- *Own Action principle (OA–P):* I can only settle matters that are up to me to settle or which I have control over.

In the context of shared intention, we have to something like the following:

- *Reliability condition (R–C)*: I am in a position reliably to predict that you do or will continue to intend like I do (that is, in favour of our shared activity), and you are in a similar position with respect to me.

A condition like R–C usually forms part of a response to a question of how the first three principles apply in the context of shared intention: how, when you and I share an intention, can I see my intention settling what we do while seeing what we do as partly up to you, so violating the OA–P? As we will see, certain authors have differing views, guided by their particular methodology, on how to address this problem.

In BEACH we have a prima facie reason to think that R–C does not hold because, simply put, we cannot take it for granted that everyone involved is in a position to predict that their partners will intend in favour of the shared activity. More specifically, Mya has reason to doubt Iva will join him at the beach, for despite Iva still intending this, Mya is not in a position to be sure[1]. And if the R–C does not hold then we don't yet have an answer to how, if the observation that Mya and Iva are still able to share the intention that they go to the beach together is correct (and, in fact, successfully end up doing so), the S–P and C–P principles are met; that is, how both of them can see the matter as settled.

It's true that whether or not it's possible for Mya and Iva to share an intention in these circumstances likely depends on how uncertain Mya is about whether Iva's intention has changed. Of course, this in turn suggests that we have an intuition about how confident Mya

[1] To flesh this out properly, we should say that Mya has good evidence that Iva may not intend their *J*-ing, so, provided Mya is rational, he has reason to believe that there is a significant chance that Iva doesn't intend that they *J*, and so he believes that there is a significant chance that Iva doesn't, in fact, intend that they *J*.

must be in his belief about Iva's intention for there to be shared intention. This approach presents its own challenges that mean it cannot explain the whole story. Why does the problem rear its head in contexts like BEACH? There are two distinct but related reasons. First, accounts of shared intention usually include a requirement that there is common-knowledge of beliefs and intentions as part of a set of minimal conditions necessary or sufficient for shared intention. Generally speaking, it means that all parties' intentions are public and out in the open, which means there's no place at all for uncertainty about what one's partners intend. Second, the R–C condition is taken to hold because it is underpinned by additional background assumptions made, typically that people we interact with are generally predictable and disposed to be cooperative. And these assumptions rest, in turn, on a view of shared activities as frequent, everyday, multifaceted experiences that licence this kind of ordinary predictability. Our social interactions are therefore so common that we don't struggle to work out what potential interaction partners think but, rather, can take for granted that they have the intentions we take them to have.

These two reasons are connected, though most authors don't do so explicitly. Their connection comes from a commitment to a particular standpoint of shared activity that sees agents as inherently inclined in favour of the joint activity. Agents don't have a reason to hide their intentions or mislead their partners, hence, whatever process of communication establishes common knowledge is usually not controversial. Though presuming ordinary cooperativity might generally make sense, these assumptions do not fit or are not suited to the situations of non-aligned interests we have in view. As I'll argue in upcoming chapters, that these background assumptions may be undermined means existing accounts of shared intention may struggle to explain how, in contexts of substantial uncertainty about intentions, agents can share intentions. This is the problem we face if we think it's possible that they can.

1.3 What is uncertainty about intentions and why is it relevant?

If we accept that investigating uncertainty about intentions within existing accounts of the phenomenon is not an irrelevant or uninteresting task, I want to make some general points about the kind of uncertainty I have in mind. This is necessary for understanding the analyses that follow while helping frame the constraints of the problem identified.

Why is uncertainty about intentions relevant?

I'm focused on shared intention's functional role in leading to joint action. Many authors on the topic try to characterise a set of necessary or sufficient conditions that plausibly allow shared intention, or a collection of attitudes conceptually resembling this, to play this functional role. For instance, Raimo Tuomela (2017) describes this as showing how "joint intentions serve their main purpose of leading to joint action" (fn. 10). Michael Bratman (2014) argues that shared intention explains shared agency (his 'connection condition'): "shared agency involves an appropriate explanatory role of relevant shared intentions. Our painting together is a shared intentional activity, roughly, when we paint together because we share an intention so to act" (pg. 10). Authors differ, of course, in what conditions they take to be minimal and how they think shared intention guides practical reasoning to enable agents to act together. Nonetheless, this implies that we're judging whether some feature is relevant and useful for an account of shared intention if it usually promotes shared intention translating into appropriate and successful action. This is the normative standard I'll be using throughout this thesis, and I'll use the terms 'promote', 'support', 'enable' and 'lead to', rather than 'ensure' or 'result in', to emphasise that individuals may not successfully act together despite sharing an intention to do so[2]. This normative standard will come under fire in a number of upcoming chapters, but remains, I think, a good benchmark for my project.

As we will see throughout this thesis, one way shared intention leads to shared intentional activity is by supporting agents' expectations about what their partners are or will be doing. Expectations that one's partner intends to *C* (e.g., contribute, coordinate, cooperate) emerge in different ways, depending on the account of shared intention. What uncertainty about intentions does is it undermines these expectations and so compromises the part they play in supporting either the sharing of intentions, successful joint action, or both. This is why uncertainty about intentions has a bearing on thinking about a generalised theory of shared intention.

Uncertainty about intentions for present or future action?

It might already be apparent that my investigation leans towards shared intention for future rather than present collective action. Questions about whether one's partner intends to play

[2] My emphasis on shared intention for *doing* means that, going forward, I am leaving open questions about what uncertainty about intentions means for accounts that focus on shared intention for other reasons, for example learning or emotional connection. These are important questions to explore in the future.

their part are more likely to arise in situations involving imperfect information, and this is far more plausibly the case for planned future interaction than it is for forms of ongoing interactivity in which agents can observe and monitor partner behaviour. This is not to say that agents cannot be uncertain about their partner's intention to contribute in the latter; indeed, It's possible to extend my research in this direction by introducing factors which constrain information individuals have of one another. A better starting point, however, is to limit my scope to contexts requiring shared planning for future shared activities, a view which also informs the selection of authors whose work I analyse; that is, those like Michael Bratman's who place the coordinative, dynamic role of intentions at the centre.

The approach of focusing on future-directed shared intention has implications for whether my findings generalise to alternative accounts with different methodologies for building an understanding of shared intention—as in, it's not obvious that they do. This problem has its philosophical roots in existing difficulties with trying to reconcile the various ways the word 'intention' is used. Setiya (2018), for example, says that

> "philosophical perplexity about intention begins with its appearance in three guises: intention for the future, as when I intend to complete this entry by the end of the month; the intention with which someone acts, as I am typing with the further intention of writing an introductory sentence; and intentional action, as in the fact that I am typing these words intentionally … The principal task of the philosophy of intention is to uncover and describe the unity of these three forms" (pg. 1).

How to do this is surprisingly tortuous given what seems to most like an intuitively obvious concept. A full exposition of the debate is beyond the scope of my project, but it remains true that my centering of future-directed intention will have a bearing on conclusions I come to. With this in mind, I have tried to broaden parts of my analysis to focus on accounts that are quite different to Bratman's in spirit and approach, including one rooted in Elizabeth Anscombe's work, and to show that the problems I identify apply there too. Still, it is likely that more work will need to be done to properly explore if and how shared intention and uncertainty about intentions are compatible with a greater variety of views.

What do I mean by uncertainty about intentions?

How am I defining uncertainty in this thesis? First, I want to be careful in suggesting that any level of uncertainty in my background beliefs is incompatible with an intention reliant on

them. For example, it's not incorrect to confidently assert that my slice of cheesecake is in the fridge (having checked earlier) and rely on that in my practical reasoning (I don't need to check again), even though there's a sliver of uncertainty about whether my recent visitor ate it while I fetched what they came to collect. It would be trivial to assume that my inability to perfectly observe the cheesecake would rule out my claim to knowing the cheesecake is there and using what I know to base my plans to eat it later.

What I am suggesting is that often we form intentions based on quite substantial uncertainty that's of a kind that stems from having good reason to doubt that the cheesecake is still there. This is to say, that if I had no doubts about this, then I wouldn't be uncertain in this way. In the kinds of cases I am talking about, it's not that there's some kind of improbable, unexpected, unusual cheesecake-eating event that I have in mind. Rather, it's something that happens frequently enough, as in a case where I had a roommate notorious for eating my leftovers.

This arguably only pushes the question back a step: how frequently must my roommate help themselves to my food before I am no longer justified in assuming the cheesecake is there? What degree of confidence must I have that makes it safe to presume my cheesecake hasn't been eaten? These are ultimately larger questions about belief and certainty, of interest but beyond the scope of this thesis. Still, I can do more to clarify what I mean, and one useful proposal comes from Berislav Marušić (2015). Though he's talking about *individual* agency and commitment against the evidence, I have adapted one of his analytic principles to suit my scenarios, holding that the following is true:

> *The Temptation-Evidence Principle* (TE–P): If I have evidence that You will be tempted not to φ, I have evidence that there is a significant (non-negligible) chance that You won't φ.

In essence, this says it's sufficient, for my purpose, to establish that an agent's awareness of her partner having attractive alternatives makes her feel uncertain about his motivation to make his contribution. The rationale for this is as follows:

> "[A]ny temptation worthy of its name is such that there is a significant chance that we will give in, despite promising or resolving to resist it. After all, if there is no significant chance that we would give in to a temptation, then that temptation

exercises no pull on us; hence we are not really tempted. But if there is a pull, there is a chance of weakness on our part" (Marušić, 2015: pg. 25).

The TE–P helps articulate what I mean by uncertainty about partner intentions, but is not essential—or at least not in this particular formulation—to appreciate the problems faced in trying to explain how Mya and Iva can share an intention to go to the beach together despite Mya being uncertain about Iva's intention to join him. And so I'll accept this principle as it is for the rest of this thesis[3]. To proceed, then, if Mya has evidence that Iva is tempted to go to the pub to watch football instead of joining him at the beach, and the TE–P holds, then I say that Mya is uncertain about whether Iva will join him at the beach. Now, there remains an open question about if and how *this* type of uncertainty—what I'll call *substantial uncertainty*—is compatible with intention. This is a crucial question for the thesis to address.

Something else to clarify is the category of uncertainty (for lack of better terminology) this thesis focuses on. Michael and Pacherie (2015), for instance, suggest three different types of uncertainty in social situations that can put coordinated action at risk: motivational, instrumental and common ground uncertainty. It should be apparent from the use of the TE–P that I am primarily concerned with the first type, but it's useful to see how this differs from the others. *Motivational* uncertainty concerns the extent to which our interests are aligned or stable enough to further a shared goal together. Imagine you and I had agreed to fill up a large tank. I'm at the pump and you're managing the switch to keep the water pressure high enough for me to pump the water in. Despite agreeing to do so, let's say I am uncertain you will raise the water pressure once I start pumping as you might not be sufficiently willing or motivated and choose not to, perhaps because you are feeling lazy or because it would require too much effort.

This is different from other reasons I might have to be uncertain whether you will raise the water pressure. For example, I might speculate that the mechanism for keeping the

[3] This possibly aligns with a natural language approach to characterising uncertainty. For example, drawing on Wittgenstein's *On Certainty* (1969), we might think that in order to say that Mya is uncertain about Iva's intention, we must have some conception of what it means for Mya to be *certain* what Iva intends. So even if we don't know how to characterise uncertainty, if we know what it means for him to be certain, then we can say that Mya is uncertain if he is *not* certain in the required way. Wittgenstein, though, found it "difficult to provide an uncontentious analysis of certainty…[as there are] different kinds of uncertainty, which are easy to conflate…., that the full value of uncertainty is surprisingly hard to capture…, [and] that there are two dimensions to certainty: a belief can be certain at a moment or over some length of time" (Reed, 2022). Still, the fact that we use terms like 'certain' and 'uncertain' and the fact that others share our interpretation of what these mean suggests that there can be something like certainty and uncertainty. I won't explain this further here; it suffices for my project that we can accept that a reason which causes Mya to be uncertain about Iva's intentions must be worth its name as commonly understood, otherwise it's not an uncertainty-generating reason at all.

water pressure high is more complicated or requires more strength than previously expected and that you lack the capacity to ensure the pressure is kept high. Perhaps the machinery is prone to breakdown, or you are prone to falling asleep while waiting to work. In these cases, even if I knew you were willing to raise the water pressure, I am uncertain about your capacity to do so. *Instrumental* uncertainty therefore concerns how, even if they are shared, our goals are to be achieved; what roles we should play, when and where we should act and whether we'll be able to perform our part.

Of course, I may trust that you are fully ready and capable of raising the pressure but be uncertain that you will see me start pumping from a good distance away, where the pressure switch is located. Or I might be uncertain that you will hear me shout that I've begun pumping over the noise of the machines we are using. *Common ground* uncertainty concerns the fact that though we might be similarly motivated to pursue the same goal and settle on similar plans and roles for doing so, it might still be the case that it's not transparent to us, that either of us doesn't know this, and so the joint action won't go ahead.

To reiterate, then, my focus in this thesis is on motivational uncertainty; the type of uncertainty due to lowered expectations that one's partner is motivated to play their part.[4]

Does this approach exclude specific accounts of shared intention from the thesis scope?

The focus on motivational uncertainty has implications for which authors I choose to analyse. There are, for example, several accounts of shared intention whose infrastructure looks fundamentally incompatible with the possibility of shared intention under motivational uncertainty. These are those explicitly committed to the idea that the people involved feel cooperative or are motivated to act together. Personal intentions in favour of the joint activity are, as is expected, part of most accounts of shared intention, yet some accounts make it especially difficult to entertain the idea that there can be shared intention while there's also

[4] The issue of scalability in joint actions and its implications for increasing common ground and instrumental uncertainty is a fascinating subject. Specifically, how do theories of shared intention make sense of large groups executing seemingly well-coordinated and aligned actions, even when individual participants may not be fully aware of who their partners are and what each of their roles is in the collective performance? Kutz (2000) proposes a modified knowledge requirement, based on differences in publicity, of mutual belief along with overlapping participatory intentions. However, List and Pettit's (2011) study on group agency is more comprehensive, tackling the key challenges in characterising agency at a group or organisational level. Recently, Michael Bratman (2022) has also turned his attention to developing his intentions-as-plans theory of agency at an institutional level. While questions about willingness to participate and free-riding will undoubtedly arise as groups grow larger, this thesis will concentrate on exploring motivational uncertainty in small-scale shared activities.

uncertainty about intentions. Consider the following passage, part of a critique of this feature from Natalie Gold and Robert Sugden (2007):

> "The literature on collective intentions is exemplified by the work of Raimo Tuomela and Kaarlo Miller (1988), John Searle (1990), and Michael Bratman (1992). A general problem for these accounts is how to differentiate collective intentions from the mutually-consistent individual intentions that lie behind Nash equilibrium behavior in games … It is clear that not all Nash equilibria are joint actions. However, the core analyses provided by Tuomela and Miller, Searle, and Bratman seem to imply that all Nash equilibrium situations are instances of collective intentions. Cases in which Nash equilibria are not joint actions are excluded only by stipulation or by the addition of further conditions which are just as problematic as the original concept of collective intention. Tuomela and Miller stipulate that the definition of collective intention includes the condition that the action is joint, Searle that collective intentions involve cooperation in pursuit of collective goals. This amounts to saying that the special feature of collective intentions that distinguishes them from the intentions behind Nash equilibrium behavior is that they are associated with cooperative activity, but this is something that we already knew prior to the analysis. Bratman adds conditions which require each agent to be responsive to the behavior of the other as the joint action proceeds and if unexpected problems occur, but these conditions are stated only informally, and rely on a pre-analytic understanding of the nature of cooperative activity" (pg. 110).

Two things stand out. First, it's unlikely that accounts of shared intention in the spirit of both Tuomela and Miller and Searle's will prove fruitful for exploring my question about shared intention in contexts of motivational uncertainty. The nature of my question effectively rules these out. Second, there is a more general problem for accounts of shared intention that bake cooperativity in up front: we come with a preconceived idea of what shared intention involves—namely, cooperativity of sorts, which enters as a pre-condition—but take for granted where this sense of cooperativity comes from in the first place. If, for example, a possible explanation of this itself involves shared intention, then the account looks circular; and if not, the question of what grounds it remains open.

Questions about what ground cooperativity and interpersonal commitment (in the face of tempting alternatives) are, in fact, central to my project. I want to thus avoid pre-supposing

any forms of cooperativity essential to shared intention. The main account I focus on is Michael Bratman's (who does actually address Gold and Sugden's main concerns). As I will show, in his later work Bratman pays particular attention to explaining how norms of mutual cooperativeness *emerge* from basic norms of rationality, rather than relying up front on conditions that codify cooperative forms of behaviour[5].

1.4 First- versus third-personal perspectives of shared intention

There's another important clarification to make early on about how I have presented the challenge. It is *not* part of BEACH that Iva no longer intends that her and Mya go to the beach together and that she intends to go to the pub instead. If it was the opposite, that Iva no longer continues to intend in favour of the joint activity, then the question of uncertainty looks like a non-starter. There really isn't a question of how to reconcile Mya's uncertainty about Iva's intentions with the two of them sharing intentions, as it's reasonable to assume that they can't possibly share an intention if their intentions are either completely different or one or more of their intentions doesn't feature the role of the other. For even a mightily pared back set of necessary or sufficient conditions for shared intention must surely involve some minimal alignment of intentions. I accept, then, that Mya and Iva can't share an intention if the latter now intends to go to the pub.

---

[5] That said, I will argue in later chapters that the spirit of Gold and Sugden's challenge remains. Though Bratman may deny presupposing cooperativity, he faces the following dilemma: either preselecting contexts that implicitly promote cooperativity—i.e., where agents' interests are closely aligned—so norms of mutual cooperativeness are assured, but meaning his account lacks generalisability; or allowing that his account covers contexts where cooperativity isn't assured—i.e., where agents' interests are imperfectly aligned—and so is generalisable, but then fails to provide a credible reason why norms of mutual cooperativeness emerge.

Tuomela, for his part, has also since developed his account, but still doesn't adequately address the challenge from Gold and Sugden. For example, in recent work he says, for our group, g,

> "… if we, you and I, jointly intend to perform X together, this required that you and I, qua members of g, both intend to participate in our joint performance of X together, this requires that you and I, qua members of g, both intend to participate in our joint performance of X for us (qua members of g) while being collectively committed to performing X jointly and, by implication, collectively committed to the process leading to X. You and I mutually know (or correctly believe) all this. That the joint performance of X is an intentional action *presupposes in this case that there is a group-based reason consisting of the fact that our group has, typically through our collective acceptance formed the intention to perform X*, and thus each participant contributed to X because of the group reason that we (viz. our group) collectively intended to perform X" (Tuomela, 2017; pg. 30, my emphasis).

The collective acceptance referred to is a condition specified in his account of we-intention: that is, each member of the group "collectively accept the truth (correctness) of "We will do X together as a group"" (Tuomela, 2017; pg. 31, (ii)). By including in his definition of we-intention this condition of collective acceptance, Tuomela is arguably still presupposing a form of jointness in the activity (the point made by Gold and Sugden). If this collective acceptance is not itself some form of joint activity, Tuomela doesn't tell us how.

This hints at taking a stance towards shared intention that adopts an 'externalist' view of social phenomena: they exist or are present only if all parties have the required attitudes by which they are constituted. In Richard Moran's (2018) *Exchange of Words*, he explores a version of externalism in what he sees as the distinctively interpersonal nature of testimony—and other social acts more broadly: both speaker and hearer are mutually dependent in the sense that the former counts as telling the latter something only if the hearer recognises this is what the speaker is doing. Success consists in this recognition; without it, there is no 'telling', and the speaker is simply mistaken if she thinks this is the case

An externalist approach to shared intention might take it that if two people are engaged in a shared intentional activity and one party suddenly stops intending the joint activity, then the shared intention falls away—that is, they simply no longer share an intention—even if the other actor, who still intends to act with her partner, is not aware of the latter's change of mind. As before, if she thinks they do share an intention, she is simply mistaken. This may be rare in practice. Given the highly coordinated nature of many joint activities and the ongoing monitoring usually characteristic of them (see Vesper et al., 2010), significant changes in intention likely lead to their rapid breakdown. Nonetheless, taking a broader view of the kinds of activities that can be collectively intentional, and as the BEACH example aims to show, it's not always possible for agents to (perfectly) monitor their partner's actions and must continue to rely on them in spite of this, especially when the joint activity is expected to take place in the future or when it evolves dynamically.

To clarify, then, I have formulated the problem such that Iva does, in fact, continue to intend to meet Mya at the beach, it's just that Mya is unsure of this given the attractive alternatives he's become aware of. So they still might share an intention on an externalist view, but it's probably correct to accept that people can't share an intention when one of them no longer intends their shared activity towards which they were previously directed.

*

BEACH does, however, raise some points about externalist views of shared intention that help frame this project. First, there are important implications that come from adopting a first- versus third-personal perspective when assessing whether a specific case of collective activity involves shared intention. What agents *believe* is the case versus what *is* the case can come apart. And it's plausible that necessary or sufficient conditions for shared intention will differ depending on the adopted standpoint. This is something overlooked by many accounts

of shared intention. A common knowledge requirement, for example, usually justifies equating the two perspectives: agents are correct in their beliefs about what they and their partners intend and believe. What BEACH does is put pressure on this perspective equivalence. By introducing good reasons for Mya to be uncertain about Iva's intentions, but leaving in place their intentions in favour of the shared activity, it forces us to take seriously the fact that Mya cannot come to know Iva's intentions in the way he can his own. This is not to say Mya cannot know what Iva intends, but that either he may not be certain what this is or we must clarify whether the process by which he comes to know has its own set of conditions that must be met. Taking a first-personal perspective pushes us to clarify what these are.

We can be more careful about what this means by borrowing again from Richard Moran, this time from *Authority and Estrangement* (2001), where he addresses the topic of self-knowledge and, in particular, explores what it is that makes knowledge of the self different from knowledge of others. His work is concerned with the mode of awareness an individual has of their own mental states and attitudes, which others do not have access to; what it means to have this privileged access, how this knowledge is both non-inferential and immediate, and why, despite this fact (or perhaps because of it), we grant individuals a certain kind of authority over what they know about themselves: "a person can know of his belief or feeling without observing his behavior, or indeed without appealing to evidence of any kind at all… [and] judgments made in this way seem to enjoy a particular epistemic privilege not accorded corresponding third-person judgements that do base themselves on evidence". (Moran, 2001: pg. 10). A full description of Moran's view is nuanced and beyond the scope here, but the main thrust is that individuals have a self-constitutive role in what they come to know about themselves. Very briefly, three (among several) of Moran's claims are relevant for the issue of perspective. First is the claim that knowledge obtained through introspection shouldn't be thought of as a genuine detection of some independent psychological fact about oneself (Moran, 2001: pg. 13). Second, that it's an agent's own conception of their own state or attitude that is partially constitutive of what that state is; the agent's first-person interpretation of, for example, an emotional state is expected to play a role in constituting the 'identity' of that state, an awareness that is not shared by possible interpretations from external, third parties (Moran, 2001: pg. 35). Finally, it is through a process of deliberation that this interpretation is formulated and self-knowledge is obtained. These describe the self-constitutive role an agent has in what she comes to know about herself, something obtained only through deliberation from a first-personal perspective which implies that this is

not the kind of access she may have to the mind of another person (i.e., from a third-personal perspective). Moran has a lot more to say and defend on this view, but this is enough for here.

With this in mind, if we take Moran's view seriously then we must think that, from an agent's perspective in a shared intentional activity, she cannot simply treat her partners' intentions as if they are her own; that is, she can't claim to know them in the same way she knows her own. To do so would violate the idea of privileged self-knowledge, removing that which distinguishes what we can know about ourselves versus what we can know about others[6]. This means that what ultimately matters for whether shared intention plays its functional role—of leading to shared activity—is what agents believe from their own perspective. An agent is motivated by what she has in mind is the case, not what actually is the case. To understand whether agents are motivated to participate, we are therefore encouraged to analyse collective activities from a first- and not a third-personal standpoint.

Moreover, as I pointed out earlier, what I am concerned with has to do with questions of motivation, in which uncertainty is not simply a practical matter to resolve. There may be material or psychological costs to doing so, which an agent may not want to bear, and so the uncertainty persists and must be part of the account. When we introduce motivational uncertainty we therefore bring out the distinction Moran draws between the first- and third-personal perspectives on social interaction.

*

A second insight we get from BEACH is that substantial uncertainty about intentions forces us to think about metaphysical commitments we might not want to make when developing an account of shared intention. Consider two modifications to BEACH. In the first, imagine Iva did in fact see the football while on the way to the beach and, true to form, decided to go to the pub to watch football instead (as Mya waited at the clock tower anxiously, as he too knows about the game). However, while en route to the pub Iva *again* changed her mind and decided her original plan was best, catching a taxi to the beach just in time to meet Mya as originally planned. In a second modification, imagine the same scenario but where Mya has no idea about the football match and no inkling that Iva might abandon him, but where Iva's decisions and actions are the same. An externalist perspective tells us that in both scenarios

[6] Applying these ideas to intentions rather than self-knowledge leads us to something very close to Elizabeth Anscombe's treatment of intention, which I discuss in Chapter 4. Moran's work is on self-knowledge more generally, but he agrees that his ideas have a lot in common with Anscombe's.

the shared intention first exists when the agreement is made, then disappears, then emerges again, as Iva changes tack—she might even be a very indecisive person, prone to switching her choices often, even only briefly. While I don't take the externalist view to be advocating for something metaphysically new or unique in shared intention—for example, a group-mind-like, irreducible, agent that 'comes into existence' when a shared intention is established, is the bearer of that intention, and disappears when the shared intention falls away—it does push us to ask whether the shared intention should be thought to flicker in and out of existence each time Iva changed her mind? What, exactly, emerges and disappears each time? These aren't questions I'll pursue here, but they are worth bearing in mind, especially for accounts that posit new metaphysical entities in shared agency.

1.5 Why focus on cases involving uncertainty about intentions?

In line with the previous section, what's of interest in introducing Mya's uncertainty about Iva's intentions is analysing shared intention from the perspective of those involved. This includes, ultimately, how they can credibly rely on one another to have the intentions they take them to have and to continue to intend in favour of the joint activity. As pointed out earlier, and as I'll describe in detail in the upcoming chapters, there are assumptions which usually mean this is not a genuine problem. But if it's possible that we sometimes share intentions despite substantial uncertainty about what our partners intend, then some of these assumptions no longer hold. And we might have to explain how, if at all, existing accounts of shared intention can accommodate this possibility—and if not then why not.

This goes to the heart of the purpose of this project. Why do we care about cases involving uncertainty about partner cooperativeness and willingness to participate? I have three main responses to this question.

First, because I take it that uncertainty about a partner's likelihood to make their contribution are common occurrences in social interaction. I do not claim that the cases I have in mind are paradigmatic cases of shared intention, but only that they are familiar enough in our everyday experiences of acting with others to warrant investigation. Moreover, I don't believe the addition of reasons to be uncertain about a partner's intentions is particularly controversial: being unsure what others think, intend, believe, want, hope or long for is part of acknowledging that while we may do things together we are separate people who must figure out, bargain and negotiate how this takes place.

Second, most accounts of shared intention take as a starting point an eager willingness of those involved to work together. While it is certainly true that many joint activities do involve these attitudes, in line with the point just made it's important that theoretical accounts of shared intention which claim to be generalisable should also consider instances like this without assuming cooperative attitudes or cooperativity-promoting contexts up front. These assumptions would face their own explanatory challenges which make taking them for granted risky. Furthermore, while authors might claim to be capturing the most sophisticated cases in their analyses, it's not clear that focusing on situations involving highly aligned interests or several ultra-cooperative agents *are* the most complicated cases to consider. In fact, as we will see, this starting point likely requires *fewer* background assumptions about what agents know about each others' minds and what motivates them to act together.

Third, I believe focusing on contexts where individuals face tempting alternatives and which evoke uncertainty about their continued intention to act together illuminates the central features of joint action which make them genuinely shared in the way most writers seem to want them to be. Analysing uncertainty need not be limited to these contexts. Perhaps I suspect you've simply forgotten to meet me for coffee this morning, or that you are unsure how to get to the cinema. In both cases I might be uncertain about whether you will join me. But uncertainty due to attractive alternatives provides a unique opportunity to study what it is that in some groups binds its members together. Put differently, while the other types of uncertainty have solutions which are perhaps technical and pragmatic, in the cases we are interested in the mechanisms for solving issues of uncertainty must lie in the nature of the social connection itself. Focusing on contexts involving imperfectly aligned interests gives us insight into why, when it's not in people's interests to do so, they nonetheless manage to cooperate and coordinate. What we will see, I think, is that assumptions related to positively cooperative attitudes that many accounts of shared intention make—either explicitly or implicitly—end up doing a lot of the work to explain why individuals are motivated to act in concert. Introducing contexts where this assumption is false shows the impact of taking this for granted—namely, that these accounts lack credible explanations for why individuals continue to participate in joint activities as we observe they do. It also points to where the cracks are and shows us what's required for an account of shared intention to be credible in these circumstances.

1.6 Thesis structure

This thesis is structured as follows. Though Mya and Iva's case is specific, their cause is more general, having to do with the (in)compatibility of traditional accounts of shared intention with contexts involving motivational uncertainty. After the introductory chapter here, in the next three chapters I therefore address three challenges to the possibility that Mya and Iva can share an intention that they go to the beach despite Mya being uncertain of Iva's intention to join him at the beach. I analyse and propose solutions to concerns related to the impact of a now-missing common knowledge requirement, in Chapter 2, a potentially weakened belief requirement on intention, in Chapter 3, and an undermined joint settling requirement in Chapter 4.

In **Chapter 2**, I note that what immediately goes missing when we introduce motivational uncertainty is common knowledge, for if there is common knowledge of intentions and beliefs, then by definition there can be no substantial uncertainty about intentions. I reflect on the role of common knowledge and argue that this requirement, an almost universal feature of accounts of shared intention, typically provides the theoretical route by which the intentions of all parties involved settle matters about what they, individually and together, will do. However, though common knowledge may be sufficient for collective settling, it may not be necessary. I draw on Olle Blomberg's work to support the point that it's plausible that agents can act intentionally together even in the absence of common knowledge. However, I argue that while his view has merit, by using the lens of intentions-settling-matters we can see how the two examples he uses differ—and that in only one is it plausible that, despite there not being common knowledge, individuals' intentions can settle matters. In conclusion, the analysis suggests that the lack of common knowledge per se needn't immediately preclude there being shared intention in BEACH, but that alternative routes for individuals to settle what they will do together must be found in contexts with uncertainty about intentions.

In **Chapter 3** I raise a second concern, about the prospects of a weakened belief requirement on intention that looks likely if we open up the possibility of shared intention in BEACH. If Mya has substantial uncertainty about Iva's intention to join him at the beach, then he cannot both intend that they go to the beach and believe that they will do so. I turn to the literature on individual intentional action, in which questions like this have been asked before, and explore if and how intention and uncertainty are reconciled there. My leading

question asks why we might allow that my intention to *A* can differ from my prediction that I will *A*. I draw primarily on Michael Bratman's planning theory of intention, notably the idea that my intention involves a commitment to action which merely expecting or predicting how I will act does not. I unpack his multi-dimensional characterisation of commitment and, to assess its validity, I discuss research from economics and psychology first on relevant biases and then on recent work on computational rationality, an approach to building cognitive models of planning which I see as having close conceptual parallels to Bratman's work. I then discuss different views on the need for a strong belief requirement on intention, before scrutinising whether Bratman's Asymmetry Thesis—a feature of his account of individual agency, which argues for intention compatibility with doubt but not disbelief—can be mapped across to solve the problem of uncertainty about intentions in the joint case. Though the norm itself is valid, and though it initially looks to provide a solution, I conclude that this falls short of what's required for shared intention. While the AT may be useful for addressing uncertainty both about partner intentions and other facts about the world, truly shared intention—and the kind of settling and commitment it entails—requires treating these two aspects differently.

In **Chapter 4**, I pick up on the intuition from the end of the last chapter that what still looks absent, in contexts involving uncertainty about intentions, is the important sense in which intention settles matters for intender about what the group will do. Without this, it's not plausible that there is shared intention in cases like BEACH. The aim of this chapter is to clarify what exactly is still missing, specifically by exploring how certain authors have suggested that this essential settling characteristic might look in the case of shared rather than individual intentional activity. I present two different theoretical accounts of shared intention in Michael Bratman's and Johannes Roessler's. I focus on how each author proposes overcoming the same problem that because intention settles matters only for the intender, in shared intention it's hard to see how each individual can settle matters about what the group will do, in which the contributions of others are beyond their own control or ability to settle. I show that, despite taking very different methodological approaches, both authors rely on surprisingly similar background assumptions about cooperativity and ordinary predictability to propose how a joint settling requirement might be met. For each author, I argue that these assumptions are plausibly not met in cases where there's motivational uncertainty, in which we cannot take for granted that we can simply rely on others to make their contributions—at least not without additional background assumptions. This suggests that the intuition with

which this chapter opened was correct, and that, though we found possible solutions for the issues identified in Chapters 2 and 3, concerns about joint settling give us an independent reason for potentially excluding cases where there's uncertainty about intentions from the umbrella of shared intention. In cases like BEACH, we don't yet have a reason, given the tools and resources from the accounts in focus, why Mya would be in a position to settle matters about what he and Iva will do.

The aim of **Chapter 5** is to explore a possible way to overcome this problem in two steps: first identifying and then filling in the gap of what's required for Mya and Iva to genuinely share intentions. The first part of the chapter picks up on the idea that there's an important difference between Mya intending versus predicting Iva's intentions and actions. I use Bratman's account of a kind of interpersonal commitment in shared intention to help understand why prediction alone is an insufficient basis for shared intention. Bratman extends his original treatment of commitment in individual agency to the joint case, proposing it as part of a kind of social rationality that emerges from a set of norms of practical and intention rationality already present in intentional action more generally. These place rational demands on parties to a shared intention to act and be disposed to support and help their partners should they need it. The second part of the chapter analysed whether there's a role for this interpersonal commitment to act as a foundation for the kind of joint settling required in shared intention. Given that social commitments are, in the literature of joint action, a popular tool for thinking about reducing motivational uncertainty, it seems reasonable that they might. Indeed, Bratman's interpersonal commitment looks like it provides a plausible explanation for why agents can rely on their partners and depend on them to make their contribution to the joint activity, despite having other reasons to be uncertain whether they will. If agents are committed to one another, and this is common knowledge, then each may be in a position to reliably predict how their partner will act and so, as per the previous chapter, they are in a position to jointly settle matters despite the sense of uncertainty. I conclude by discussing an important, and perhaps controversial, implication of thinking about commitments as partly settling matters, which is that for it to be plausible, it requires that we think about commitments as being distinct from intentions, even if they are usually and characteristically connected.

Chapter 6 begins by analysing a potential issue of credibility with Chapter 5's application of interpersonal commitment. For commitments to work, the receiver must be able to rely on or trust that the commitment provides normative guidance to the maker such

that the maker is more likely to adhere to what they have committed to do than prior to committing. I argue that a big issue in Bratman's account is that his view of interpersonal commitment is not credible, in that his account of shared intention doesn't seem to explain why people are motivated to meet their commitments to their partners. It is his ideas about cognitive and informational constraints which rationalise his original notion of commitment—and which underpin the norm of stability—but in extending this to the shared case he gives us no reason why agents would be motivated to meet their commitments rather than reconsider previously formed intentions when they face attractive alternatives, as in our case at hand. This means that his proposal is not well suited to helping overcome the hurdles of explaining the possibility of shared intention in BEACH, at least not without the risk of either proposing something circular or straying away from his core continuity thesis. In addition to not providing a suitable tool for our particular case, I outline risks this finding poses for Bratman's account more generally. Bratman uses interpersonal commitment—and the supportive behaviours and dispositions which flow out—as evidence of what I call non-tokenistic and non-instrumental social interaction; that is, to defend a strong sense of jointness for his account of shared intention. But if these behaviours do not actually uniquely identify non-strategic behaviour, as Bratman argues they do, then this isn't available to him to use to justify why his account excludes the forms of strategic interaction he says it does, something several critics challenge him on. This also undermines Bratman's proposed 'division of philosophical labour' between describing a minimal account of basic rationality supporting shared intention versus relying on obligations, moral or otherwise, to provide the necessary support. I proceed to analyse an account that emphasises the latter, Margaret Gilbert's theory of shared intention and the joint commitment and mutual obligations she sees as essential. I discuss several positive features of her account, but ultimately argue that it too lacks a credible view of why individuals are motivated to meet their commitments. I also look into Gilbert's sources of inspiration for her claim-rights perspective, unpacking whether she's correct in her proposal that mutual obligations in shared intention should be understood as non-moral in nature. Finally, I revert back to 'reductive'-style accounts, and explore Berislav Marušić's work on individual agency and committing to actions under uncertainty. Looking at an extension he provides to cases of joint activity, in which he presents trust as a mediating factor, I argue that we end up with the same issues as Bratman. I conclude that all of these particular authors' approaches struggle to explain how there can be shared intention in contexts with motivational uncertainty. What's missing is a robust basis for commitment/trust that both makes them credible while remaining within the constraints of minimalism.

Chapter 7 presents an account of interpersonal commitment which provides both an answer to certain issues faced by the accounts of commitment just analysed and a solution to the problem of joint settling in contexts of uncertainty. The chapter is framed in terms of how people experience social commitments in the context of commitment dissolution. This turns out to be a revealing test case for theories of commitment, as it's in these contexts that traditional accounts of commitments have little to say. I set up and derive from them a simple view of commitment and discuss its shortcomings. Using a recent body of work by several authors, I present their framework for a minimal psychological sense of commitment arising in joint activities. I outline the theoretical background—in particular, the basic need to meet reasonable expectations others have of us and the need to maintain good relationships with others—and show how it is this minimal commitment which performs the function of allowing individuals to settle matters in shared activity when there is uncertainty about intentions. Though the proposal is descriptive, and I present recent empirical work in support, I discuss its theoretical implications; notably, the need for greater allowance for both explicit and implicit commitment generation processes, the need to account for both proximal as well as ultimate psychological processes and the need for a graded characterisation of the experience of commitment. I address two concerns about the proposed framework, the first to do with commensurability in weighing up costs and benefits in the reasoning process and the second concerning how close we come to encompassing a kind of instrumentality that many authors tend to preclude for sharing intentions. I respond to the latter by arguing that an account of shared intention must incorporate both instrumental and non-instrumental motivations for participation, and present selected empirical work to show one route towards finding the right balance. I conclude by looking at the implications for using the sense of commitment as a basis for solving the problem of how agents can jointly settle matters under conditions of substantial uncertainty.

CHAPTER 2

Common Knowledge and Collective Settling

The previous chapter introduced the main topic of this thesis by painting a broad picture of the problem we face if we try to explain how there can be shared intention in BEACH. This is that uncertainty about intentions—motivational uncertainty—seems to undermine some key tenets of existing accounts of shared intention. This chapter and the next address in greater detail two issues which stand out; namely, that uncertainty about partner intentions seems incompatible with requirements on shared intention of common knowledge or strict intention-belief coupling, respectively. Given how central these usually are to theoretical accounts, if it's plausible that there is shared intention in BEACH, then we need to find a way to either overcome or manage these concerns.

This chapter picks up on the first issue, as what seems obviously missing from contexts like BEACH is some form of mutual or shared knowledge of intentions and beliefs. This looks like a big problem, given that existing accounts generally involve, as one of several conditions sufficient or necessary for there to be shared intention, an assumption of common knowledge, or something similar. Common knowledge-like assumptions appear in a variety of forms in theories of joint action, but generally perform a similar and important role: common knowledge ensures that when we act together, we're each aware of our intentions to act jointly, we know what roles we must each play, what our parts are and, consequently, what our partners expect of us. The assumption is core to how the structure of shared intention is said to provide a robust mechanism for helping us coordinate, plan and undertake our joint activity. Without it, it's argued, the intentions and beliefs of jointly interacting agents are undermined.

So, what looked like a pretty straightforward observation about social interaction in contexts of uncertainty now generates a surprisingly difficult problem. For if we have common knowledge concerning our intentions that we, for example, go for a picnic while a football match is going on, then there can be no uncertainty about each others' intentions. I can have no uncertainty about yours and you can have no uncertainty about mine. So there's something missing in trying to reconcile accounts of shared intention with the presence of

motivational uncertainty. If there *is* actually a basis for each of us to have the knowledge that I know that you know that I know et cetera, then it's hard to see how there can be both common knowledge and reasonable uncertainty about your intentions. If that I intend to fly to London is common knowledge, then it can't be the case that you're uncertain about my intentions to fly to London. Uncertainty about intentions appears to violate the assumption that there is common knowledge about your (and my) intentions.

The primary concern here, then, is that what goes missing when we introduce substantial uncertainty about partner intentions is the assumption of common knowledge. What's more interesting, though, is not the missing common knowledge per se, but rather what *else* goes missing when the requirement is not met. Indeed, reflecting on a particular role common knowledge plays will help us pinpoint what might be missing in BEACH. It will also give us a good reason—but one not typically discussed—why a common knowledge requirement is important for theoretical accounts of shared intention.

2.1 A proposed role for common knowledge in reductive accounts of shared intention

Several authors have suggested that we should be sceptical of the idea that common knowledge is a necessary feature of minimal accounts of shared intention. There are, they claim, good examples of cases which bear many of the hallmarks of shared intention but which, arguably, lack common knowledge of intentions—and so are traditionally excluded. Debrah Tollefson (2005), for example, writes that young children seem able to engage in joint activity despite apparently lacking the sophisticated theory of mind that common knowledge seems to require (see also Pacherie, 2013). But even in some common cases involving adults, who presumably do possess the necessary psychological capacities, Olle Blomberg (2016) has proposed that this needn't preclude the sharing of intentions. He argues, instead, that any common knowledge-like requirement should not be part of a minimal account of shared intention. Indeed, he says, referencing several leading theoretical accounts in the field, the common knowledge condition "is typically merely assumed rather than argued for" (pg. 317). Blomberg supports his view by drawing on several examples of social interaction (which I'll shortly describe in detail) where common knowledge is not present but which, he argues, still intuitively qualify as cases of shared intention, by meeting some reasonable minimum standards. Blomberg is only partially correct in his assessment of the cases he is considering,

or, rather, he is correct provided some additional assumptions hold in the background. These assumptions have to do with the role common knowledge plays in supporting shared intention, specifically by establishing or providing the conditions for agents to settle matters together about what they will do.

A brief return to Bratman's account will illustrate this. Bratman's account is, to borrow a popular term, 'reductionist' or 'reductive' in the sense that he seeks to reduce the phenomenon of shared intention into component parts that can be explained by referring only to the attitudes and the behaviour of the individuals involved. Following this, his account of shared intentional agency is expressly grounded in his own account of individual intentional agency, and he sees the role intentions play in coordinating and planning action as essentially the same whether individual or joint. As are the specific norms governing practical reasoning that form part of what he calls intention rationality. This is part of Bratman's *continuity thesis*, the idea that there is nothing normatively, conceptually or metaphysically different between individual and shared intentional activity. A consequence of this is that anything needed to explain shared intention, any ideas, tools or theses, should be available from what we know and understand about intention per se.

Accounts like his propose an interpersonal, interlocking complex of individual intentions which play the basic roles characteristic of what we think of as shared intention These accounts thus broadly see shared intention as a network-like structure built up from certain cognitive attitudes individuals have, including attitudes directed towards the joint activity and those directed towards their partners' cognitive attitudes. Taken together, the structure of shared intention is therefore supposed to provide the necessary practical normative force for us to undertake a joint activity as well as a robust mechanism for helping us coordinate and plan in ways that track the goal of our *J*-ing.

One source of this normative force is the settling that comes with intending. We can say that if I settle my *A*-ing, I expect *A* to happen and see *A*-ing as reliant on and brought about by my contribution. It is thus part of my intention to *A* that I see my *A*-ing as up to me. Given the continuity thesis, it must also be an essential feature of shared intention that it settles matters concerning shared activity. A simple mapping of the settling characterisation above to the joint case might look like this:

> If it is part of my intention to *A* that I see my *A*-ing as up to me, then it is part of our shared intention to *J* that we (each of us) see our *J*-ing as up to us (together).

The immediate issue that jumps out is that 'we' refers to each agent's first-personal perspective while the *J*-ing is up to us together, not each of us on our own. How it's possible for an individual to intend a joint activity is something I address in Chapter 4. Cashing out a correct generalised formulation of shared intention which is neither circular nor squeezes out its distinctive form of sociality is a complex task. For now, I want to present a simplified version that embraces this first-personal, singular agent view on the collective action. This will illuminate a role the common knowledge condition typically plays in meeting some form of settling requirement essential to the sharing of intentions.

*

Given the above characterisation, if you and I share an intention then I settle our *J*-ing and so do you. For me to settle matters means I expect *J* to happen, and see *J*-ing as reliant on my and your contributions and brought about by my and your contributions; and the same goes for you. It is thus part of our shared intention to *J* that I see our *J*-ing as up to us and that you see our *J*-ing as up to us. There's still something missing though. While each of us personally sees our *J*-ing as up to us, it's not clear that *we* view it as up to us—it's not clear that you and I regard *us* as seeing it up to us. We have what looks like me settling matters and you settling matters but not *us* settling matters. This is one of the worries just alluded to.

How do accounts of shared intention explain how intentions settle matters in joint activity? We can turn to a simplified version of Bratman's view to answer this. He proposes a version of his *Shared Intention Thesis* (SI) (View 4, 1999c) as follows. With respect to a group consisting of you and me, and concerning joint activity *J*:

We intend to *J* if and only if

(1) (a) I intend that we *J* and (b) You intend that we *J*

(2) I intend that we *J* in accordance with and because of (1)(a), (1)(b) and meshing subplans of (1)(a) and (1)(b);

You intend that we *J* in accordance with and because of (1)(a), (1)(b) and meshing subplans of (1)(a) and (1)(b)

(3) (1) and (2) are common knowledge between us.

Most reductive accounts of shared intention involve a condition like (3), such that participants' intentions are common knowledge. Common knowledge ensures that individuals' relevant cognitive attitudes are public. It is relied upon to ensure that "everything amongst the parties is above board… that we are aware of what is happening, aware of our each being aware of this, and so on" (Pettit and Schweikard, 2006)[7].

A common knowledge requirement may be important for many reasons, but one that is underexplored in the literature is that it justifies individuals being in positions where they can settle matters about what their group will do. We can see this by reflecting on the conditions above. First, from condition (1) both you and I have an intention in favour of the joint activity—that is, that we *J*. Second, from condition (3) I know your intentions and you know mine. This means we can, for example, plan properly in light of what we know about each other and the shared activity in question. I know that you, as a rational actor, will take steps towards making your contribution, and you know this about me. Third, again if we view one another as rational agents, each of us can rely on the other to be mutually responsive in their intentions and actions to what we ourselves intend and do, where "such public mutual responsiveness involves practical thinking on the part of each that is responsive to the other in ways that track the intended end of the joint activity" (Bratman, 2014: pg. 79). If I know that you intend that we *J* (1), and I know that you see *J*-ing as partly up to me as well (2), then I know that there is rational pressure on you to respond to my intentions and actions in ways that support us achieving our joint activity.

These three features—personal intentions in favour of *J*-ing, knowledge of others' intentions in favour of *J* and expectations of mutual responsiveness, all in context of mutual awareness of rationality—means I am in a good position to predict with good confidence how you will act, and vice versa. This supports each of us having the expectation that *J* will occur.

In addition, from (2) I intend that we *J* in part because you intend that we *J*. Given an appropriate intention-action link, I then expect that our *J*-ing is partly to be brought about and reliant on your contribution (as well as my own). I therefore see our *J*-ing as partly up to me and partly up to you. Given the symmetry of these conditions, the same therefore holds for you, and I know this holds for you because I have knowledge of (2), again via the common knowledge condition. This suggests that each of us settles our *J*-ing. My intention supports the expectation that *J* will occur by seeing *J*-ing as reliant on my and your contributions and

[7] These authors' CK condition is: "they each believe in common that the other clauses hold"

brought about by my and your contributions. And your intention likewise. I therefore see our *J*-ing as up to us and you see our *J*-ing as up to us.

Note, though, this conclusion doesn't require there to be common knowledge, only that each of us have knowledge of (1) and (2). Perhaps this is sufficient for a narrower view of settling but there's still what looks like the unanswered question of whether *we* see matters as up to us, rather than each of us seeing matters as up to us. This is a notorious problem for reductive accounts of shared intention. For example, in Bratman's SI thesis above, the account appears circular if we build sharedness into the personal intention in (1), something Bratman himself acknowledged. We need a different way to somehow incorporate the sense of collective agency. Focusing on the particular problem of settling shows how common knowledge can help us with this.

Common knowledge means not only do I see our joint action as up to us and you see our joint action as up to us, and that I regard you as seeing our joint action as up to us, but that I know that you regard me as seeing our joint action as up to us, and I know that you know that I regard you as seeing our joint action up to us. This means that we share an awareness that the collective action we intend is up to us both. Each of us is aware of the same thing—that the collective action is reliant upon and brought about by both of our contributions—and we are aware that we are both aware of this. This has directly to do with settling in the joint case: as a group, we settle matters together. But it is only because we each see our intentions as settling matters as part of a group that the group is formed. The group depends on each of our personal recognitions that the group's performance relies on and is brought about by each of our contributions and our contributions together. This gives us as good a sense of collective settling as we'll get given the constraints of the continuity thesis—the theoretical continuity between individual and shared intentional activity precluding any normative, conceptual and metaphysical differences between them.

The structure of attitudes in the SI thesis is supposed to provide a good basis for successful coordination partly through ensuring that each participant is in a position to settle matters while seeing those matters as up to them both to settle. Knowledge of intentions, essential contributions of both and meshing sub-plans support our expectations that *J* will occur by seeing *J*-ing as reliant on our contributions and brought about by our contributions. This means that I see our *J*-ing as up to us and you see our *J*-ing as up to us—each of us settles our *J*-ing. Common knowledge moves us beyond this by building in the sense of

awareness that we are part of a group, a group which is responsible for a collective performance: we see our *J*-ing as up to us, individually and together.

2.2 Blomberg on common knowledge in shared intention

Common knowledge therefore provides one justification for how individuals sharing intentions settle matters together. This connection is, I think, something that tends to be overlooked, so highlighting it in this chapter helps us appreciate the role this condition plays in joint action. It also opens up a promising avenue for thinking how it's still possible for there to be shared intention without there being common knowledge. For if it's truly possible for each of us to settle matters with more limited knowledge sharing than under common knowledge, agents could still be in a position where each of them settles matters about what the group will do—though it may be arguable whether they settle it together—and that this can plausibly support activities which are in some sense intentionally shared.

I'll use this lens to explore two examples used by Olle Blomberg (2016) in his claim that common knowledge should not form a necessary part of a minimal account of shared intention. What we will see is that while he uses both examples to illustrate the same point—that is, there are cases of plausibly-shared intention without common knowledge—if we keep in mind that all agents' intentions must settle matters we can see how the two cases he presents are crucially different. I argue that in only one of the possible scenarios is it plausible that both individuals can settle matters (I see our *J*-ing as up to us). I also argue that in both scenarios it's not plausible that the sense of collective settling is met (We see our *J*-ing as up to us).

This shows us why a common knowledge requirement is important, and what might go missing when it's dropped, which has an important bearing on my thesis. One reasonable extension of the discussion in the previous section is to think of common knowledge as sufficient but not necessary for collective settling. This opens up the possibility of other routes for thinking how agents can see themselves as part of a group who, together, settle matters about what the group will do. It tells us that in cases where there's substantial uncertainty about intentions—and in which by definition there is no common knowledge—then shared intention might be possible provided we find another feature of the interaction that justifies each individual being in a position to rely on their partners to have the intentions they think they have, and so settle matters.

*

Many authors present a set of conditions either necessary or sufficient for shared intention which typically include there being common knowledge but without making an extensive case for why this is so or explaining how this comes about. For example, Bratman simply says that when we share an intention, each of us "is responsive to the intentions and actions of the other in ways that track the intended end of the joint action—where all this is out in the open" (Bratman, 2014: pg. 79), and where "it seems reasonable to suppose that in shared intention the fact that each [of us] has the relevant attitudes is itself, out in the open, is public" (Bratman, 1992)[8].

Blomberg (2016) questions whether a common knowledge requirement (or similar) should be a necessary feature of a minimal account of shared intention. His challenge is based on proposed cases of social interaction where common knowledge is not present but which, he argues, intuitively qualify as cases of shared intention. In both of Blomberg's examples, one party in a shared activity misattributes a false belief to her action partner about her own intentions. That is, she falsely believes that her partner falsely believes that her intention is *X*, where *X* is not an intention to do one's part in the joint action but some other pattern of behaviour. The presence of these false beliefs, he aims to show, needn't preclude the possibility that the two individuals share an intention, at least in some minimal sense.

There are a few more general issues a reader might raise regarding Blomberg's analysis which could shade what follows and so are useful to address up front. First, there's no doubt that common knowledge of intentions and beliefs would make shared intention—or any general structure of interconnected attitudes—more robust to environmental fluctuations and more likely to result in joint action. If it's truly possible to share intentions against a backdrop of false beliefs, this situation is probably more fragile and prone to breakdown than when there is common knowledge. In fact, as we will see, in Blomberg's scenarios it turns out to be quite a coincidence that individuals' intentions don't conflict and for there to be sufficient overlap to enable some form of joint action. Moreover, it would ordinarily be odd (or symptomatic of a divergence from rationality) if those involved continued to engage without taking some measures to rectify these false beliefs. It would feel, instead, like it's just a matter of time before the joint action collapses. An interesting question is thus whether

[8] See Bratman (2014: 57–59) for a broader, but still brief, discussion of 'out in the open'. He retains the idea that common knowledge is a non-controversial feature of shared intention.

Blomberg's are examples of shared intention or merely a sort of precursor to it. However, he is keen to emphasise that success conditions on shared intention should not be burdensome:

> "It is true that the presence of CK-defeating false higher-order beliefs will typically make coordination of a joint activity somewhat precarious and inflexible. Other things being equal, such false beliefs increase the likelihood of glitches and breakdowns in communication and coordination. However, the importance of efficiency and robustness should not be overstated. Glitches and breakdowns rarely put a complete halt to coordination or communication in joint activity. Typically, they merely result in brief temporary obstacles that participants quickly overcome… At any rate, the aim of reductionist accounts is to explicate the difference between intentional joint action and mere coordinated parallel activity. It is not to specify conditions under which joint action is reliably successful. Hence, reductionists cannot appeal to considerations concerning efficiency and reliability in support of the CK-condition" (Blomberg, 2016: pg. 8).

Of course, though, and this is the second concern, if we accept the above it looks like we're going to be relying on intuition to justify which cases qualify as involving shared intention. This approach is always open to someone saying that they simply don't share the same intuition. To be fair to Blomberg, this issue is symptomatic of the literature on shared intention more broadly, in which philosophers seem, at heart, to be merely exchanging intuitions about different cases. The main point is that it's not obvious that my upcoming discussion of Blomberg's argument is tapping into anything particularly deep—and we might be unsure whether we'll discover anything significant via this method. Still, raising Blomberg's main ideas will shed light on an important question about what is required for there to be shared intention in contexts of uncertainty, so even if Blomberg is wrong, his examples highlight the importance of keeping settling in mind when analysing whether a potential scenario involves shared intention or not. In addition, my project also begins with an intuitive starting point—namely, the possibility that there is shared intention despite individuals having reason to be uncertain about their partner's intentions—so perhaps it's best left up to the reader to decide whether this approach has merit.

Finally, in what comes I have taken a particularly narrow approach to thinking about why shared intention requires common knowledge. My focus is on a particular functional role that I've argued common knowledge plays. There may be other good reasons why authors

care about having common knowledge, such as grounding a unique phenomenology of shared activity, like a sense of shared agency. Seeing whether upcoming solutions to the functional aspects can support other reasons for common knowledge requires further analysis.

*

Let's turn to Blomberg's two case studies, one in which Hector and Celia are busy building a block tower together and another where You and I plan on walking down to the valley together. In TOWER BUILD, Hector and Celia each intends that they build a block tower, and each intends to do their bit of this joint performance. Each also believes that the other both intends to do their bit and sees their bit as part of a joint performance (i.e., each believes the other intends that they build a clock tower). And each intends to do their bit because the other intends to do theirs, and also intend that their plans and sub-plans mesh. Blomberg says:

> "These intentions and beliefs appropriately cause Hector and Celia to build the block tower, that is, the attitudes cause them to take turns putting blocks on top of each other so that a block tower is built. Celia starts by putting down one of her blocks, Hector then puts down a block on hers, and so on until the tower is completed. Note that, as a side effect, the attitudes cause Hector to cover the top face of each of Celia's blocks" (Blomberg, 2016: pg. 318).

So, Blomberg, takes it, several conditions of a general and minimal account of shared intention are met. However, he carries on, "suppose that Hector falsely believes that Celia falsely believes that he intends to cover the top face of each of her blocks rather than to do his bit of their joint performance" (Blomberg, 2016: pg. 318). A common knowledge condition is not satisfied, for even a weak version of it would require that joint action partners lack false beliefs about what their intentions are. Yet, Blomberg claims, "Hector's false belief can be present and persist while Hector and Celia successfully execute their intentions that they build a block tower… the false belief can persist without any failure of rationality on the part of either Hector or Celia" (Blomberg, 2016: pg. 318).

Blomberg's claim is that, though undeniably rare, in these situations agents are not merely performing individual actions in parallel or where they accidentally have a joint effect. Instead, each intends that they, both of them, undertake the whole action, and their intentions are interdependent so that they each settle that they, together, undertake the joint action performance. Furthermore, this doesn't require either party seeing the other as

behaving irrationally, as the actions of both are compatible with either the tower being built or the block faces being covered. A real-life example Blomberg gives is of someone who makes the mistake of thinking their ostensive partner is mistaken in his belief that she'll go ahead with the activity whether or not the other joins her,

> "so that the satisfaction of [her] intention is compatible with [his] involvement but doesn't require it… Such false higher-order beliefs are arguably a common upshot of insecurities and mild forms of paranoia that are often present in human relations. And such false higher order beliefs and doubts can persist throughout joint activities that at least appear to be jointly intentional" (Blomberg, 2016: pg. 319).

(An important and project-relevant reminder is that while one party has a belief about another party's belief which is mistaken, from a third-personal standpoint we know that they actually do have the same intentions. The first- and third-personal perspectives diverge.)

This example undermines one motivation for common knowledge, which is that in its absence the intentions and beliefs reflected in the other necessary or sufficient conditions of shared intention will be undermined. The logic of this "Rational Intending (RI) Argument", as Blomberg calls it, is that if we assume that an agent can intend to *A* only if she believes she will *A*, then if she doesn't believe her partner intends in favour of joint activity *J* like she does, then she cannot believe he intends to perform his part of *J*, and so she cannot intend her own part of *J*—as *J* depends on both their contributions. They therefore cannot share an intention that they *J*. If all their intentions and beliefs were out in the open, this common knowledge would enable these conditions involving intentions and beliefs to be satisfied. Blomberg's second example focuses directly on this argument.

In VALLEY WALK, You and I each intend that we walk down to the valley. Similar conditions hold as in TOWER BUILD, including that each of us intends to do our bit because the other intends to do theirs. Now, says Blomberg,

> "… suppose that I mistakenly think that you believe that I intend, rather, that we walk up to the hilltop. If the Rational Intending Argument is sound, then my false belief about your belief about my intention will undermine my intention to do my bit of our walking down to the valley, as well as my intention that we walk there.
>
> The argument rests on the assumption that an agent can intend to [*J*] only if she believes that she will [*J*]. Given this, I can intend to do my bit of our walking

> down to the valley only if I believe that you will walk down to the valley. After all, if you don't, then I will not be able to do my bit of our walking there. Since I intend to do my bit in part because you intend to do your bit, I must not only believe that you will walk down to the valley; I must also believe that you intend to walk there (as your bit of our walking there). Furthermore, I realize that you cannot intend to do your bit of our walking down to the valley unless you believe that I will walk down to the valley. Now, I falsely believe that you believe that I intend that we walk up to the hilltop rather than down to the valley, so I will believe that you believe that I will not walk down the valley but rather up to the hilltop. Hence, from my mistaken point of view, you cannot rationally intend that we walk down to the valley, or intend to do your bit of this walking. This will in turn undermine my own intention" (Blomberg, 2016: pg. 319–320).

A common knowledge condition sorts this out because, according to the RI argument, participants can only rationally have the previously assumed intentions and beliefs when it is common knowledge that they have these attitudes. However, says Blomberg,

> "[t]he case of Hector and Celia shows that the Rational Intending Argument fails. Hector's intention that he and Celia enact the joint performance of building a block tower is not undermined by his false belief that Celia mistakenly thinks that he merely intends to cover the top face of each of her blocks" (Blomberg, 2016: pg. 320).

Taken together, these two examples are supposed to show how it's possible for there to be social activities closely resembling shared intention when there is no common knowledge of intentions.

2.3 False beliefs and compatibility constraints on intention

There are, however, important differences between the two examples which Blomberg has not spelled out but which are relevant for his conclusions as well as our focus on settling. My settling *J*, recall, usually involves both that I expect *J* to occur and that I see *J* as reliant on and brought about by my and your contributions. And your intention similarly settles matters for you. I therefore see our *J*-ing as up to us and you see our *J*-ing as up to us. A lens of intention-settles-matters will help see why it's incorrect to use the conclusion from TOWER BUILD to argue for the same in VALLEY WALK, as Blomberg has done.

In TOWER BUILD, Celia's actual view is that Hector intends to build a tower with her and she likewise with him. So Celia, from her own perspective, is in a position to settle matters about what she individually will do as well as what they as a pair will do. She intends to build a tower with Hector and thinks that Hector intends to build a tower with her. She therefore takes actions which form a pattern of behaviour which, for all intents and purposes, lead to the tower being built. All's well with Celia! But Hector, who correctly believes that Celia intends that they build a tower together, thinks Celia's view is that he really only intends to cover the tops of blocks (and not that he intends that they build a tower together, which is the truth). Fortunately for the success of the tower being built, this is compatible with what Hector truly intends, namely that he and Celia build the tower together. I say fortunately as it is only because of this compatibility—or, in Bratman's words, meshing of sub-plans—that Hector is in a position to confidently expect that Celia will perform actions that allow an intention to cover the block faces as well as an intention that they build a tower together to be met. We can see this more clearly by outlining several important compatibility constraints on Hector's beliefs that make it possible for there to be shared intention in the way Blomberg envisages:

> H(1) Hector's actual intention (that he intends that they build a tower together) is compatible with Celia's actual intention (that she intends that they build a tower together);
>
> H(2) Celia's actual belief about Hector's intention (that he intends that they build a tower together) is correct;
>
> H(3) Hector's actual intention (that he intends that they build a tower together) is compatible with what he believes about Celia's intention (that she intends that they build a tower together);
>
> H(4) Hector's actual intention (that he intends that they build a tower together) is compatible with his belief about what Celia believes he intends (that he just intends to cover the blocks);
>
> H(5) Hector's belief about what Celia believes he intends (that he just intends to cover the blocks) is compatible with what he believes about Celia's intentions (that she intends that they build a tower together).

(There are parallel constraints on what Celia's intentions must be which are not necessary to go into as Hector's perspective is enough to make the point.)

Taken together, H(1) to H(5) imply that Hector can confidently predict that Celia will continue to make her contribution to their collective activity. But crucially note that this is partly because he believes that *she* can confidently predict how he will act—via H(5)—given that she would, if she had this belief, see her intention that they build a tower together as being compatible with his intention to simply cover the face of the blocks. This makes it plausible that each will, without being irrational, make their contribution and so form a pattern of behaviour that Blomberg claims is, in essence, joint activity.

This gives us one element of Hector's intention settling matters: it supports his expectation that they will build a tower together. What about the sense of control, or 'up to us' that is also supposed to characterise settling when sharing intentions? We can respond to this by answering several relevant sub-questions.

Does Hector see their joint activity as up to them both? Yes, he sees it partly as relying on and being brought about by his contribution (building a tower) and Celia's contribution (building a tower).

Does Hector regard Celia as seeing their joint activity as up to them both? Yes, in his eyes Celia sees it as relying on and being brought about partly by her contribution (building a tower) and partly by his contribution (covering the tops of faces). Even if she thinks he just intends to cover the tops of the blocks, she still sees it as up to him to perform, and it still contributes to their tower building.

Does Hector regard Celia as seeing that he sees it as up to them both? Plausibly yes, if, in Hector's eyes, Celia sees that Hector's sequential covering of the block faces requires that he sees her contribution (e.g., to put the next block down so he can then cover the next block face) as also required and up to her to perform. But plausibly also no, if Hector thinks Celia sees that all he intends is to cover one face and be done with it.

This gives us reason to think that Hector's intention settles matters, for him, about what they will do, and the same goes for Celia. And they can do this without behaving irrationally. This is perhaps enough for there to be a kind of coordinated interaction with aligned and closely interdependent intentions to make Blomberg's claim that there is shared intention here plausible.

What is missing, though, is the sense in which they settle matters *together*. There is no shared awareness that, from their own perspective, *we* settle matters about what we will do. We can see this by noting additional background assumptions that haven't yet been made clear but which highlight this lack of shared understanding. The first is that Hector must believe that Celia will participate even if she thinks Hector and she don't intend the same thing. For her part, Celia seems to be going ahead and making a contribution (in reality because she believes Hector intends the joint action like her) and so Hector must assume, given his beliefs, that she's happy (perhaps benevolently or patronisingly) to intend their tower building while simultaneously being content to let him intend to just cover the block faces. We must also assume that Hector must be happy to tolerate what looks to him like a misunderstanding on her part or even a mild form of deception on his. He believes (though falsely) that Celia is mistaken about his intention so, if Blomberg is to be believed and the shared intention persists despite this false belief, then it must be that Hector is content to let Celia be wrong. Absent these assumptions, it's hard to see how Blomberg's conclusions hold.

2.4 Sub-plans meshing both down and up

Before moving on to VALLEY WALK, let's adapt TOWER BUILD. Imagine, now, that Hector's false belief is instead that Celia falsely believes Hector wants to build a tower to a certain height in order to knock it down. Again, Celia's actual intention (that they build a tower together) and her belief about Hector's intention (that they build a tower together) causes her to make her contribution and expect him to make his. But now Hector thinks that Celia thinks he only wants to build the tower to knock it down, which is incompatible with what he believes about her own intention (i.e., that they build a tower together). In this case, could Celia be in a position to rationally, confidently predict that Hector would continue to perform a pattern of behaviour that would satisfy them building a tower together, if she thought he was going to knock it down at some point? Probably not.

To be sure, there's a subtle point about the extent to which sub-plans may overlap. Say Hector sees Celia as thinking that he wants to build the tower to be 10 blocks high before knocking it down. What he believes Celia's intention to be is now compatible with what he believes she thinks his intention is, so it's possible that each makes their contribution until the tower is 10 blocks tall—and all real and perceived intentions are satisfied. Such a situation would reasonably require additional assumptions, one of which is that Hector must believe that Celia's intention that they build a tower together is being satisfied while they are still

building on the way to reaching 10 blocks high. There wouldn't be the compatibility described above if he thought her intention was instead not just that they build a tower together but, for example, that they build a tower together to be 50 blocks high, or that they build a tower until they are tired but leave it standing for more work tomorrow, with anything short of this not worth the effort. Otherwise, Hector should think that Celia would feel that her intention that they build a tower together will indeed be undermined by what she takes as his intention to destroy the tower or stop building at some point. And because of this he should think that she can't expect to continue to build a tower together, so he should think she won't make her contribution. And because of this, he won't make his.

This suggests we need to add something to the already existing condition that agents' sub-plans must mesh. Comparing these adaptations with the original TOWER BUILD tells us that in order to generalise Blomberg's ideas about shared intention with something like persistent false beliefs, we need an additional assumption about intention compatibility and plans and sub-plans meshing, which could look something like this:

> Agents' plans and sub-plans must mesh (i.e., not be incompatible) both all the way down and all the way *up* to the point where at least one (but it could be both) agent's intention is fully satisfied[9].

Up to the point at which sub-plans stop meshing, Blomberg's argument that shared intention can persist despite the presence of false beliefs is plausible. Beyond it, however, it's reasonable to think that the collective activity will break down and that there can be no instance of shared intention in which both agents are behaving rationally. And it also implicitly covers situations in which an intention has no finite endpoint, such as simply intending to build something together (i.e., the process of simply doing something with someone else), or when both intentions persist but are not yet fully satisfied.

Meshing upwards is nicely illustrated with our visual of tower building. If Hector and Celia both intend that they build a tower 50 blocks tall but Hector believes that Celia thinks that all he wants to do is build the tower to this height to knock it down, provided they are still on the way to 50 blocks then there's never an opportunity for these false beliefs to be corrected, and so it's reasonable to think that these beliefs can persist as their tower building

[9] Or alternatively: The satisfaction of at least one agents' intention (i.e., their entire system of plans and sub-plans) must be fully compatible (i.e., mesh all the way down and *up*) with the partial satisfaction of the other agent's intentions (i.e., their entire system of plans and sub-plans).

is underway. It's also possible that they never reach a point at which their plans and sub-plans diverge, or that there is no finite goal (like desired tower height) being aimed at. In Blomberg's original example, Hector's false belief can persist as his covering of the faces meshes with Celia's sub-plans of putting one block onto another (which incidentally leads to covering the block faces), and each can continue as is as long as their intention remains.

*

Blomberg's second example, VALLEY WALK, resembles the adapted cases above more than the original Hector and Celia scenario that he claims it does. In particular, the earlier H() constraints are not all met. Most importantly, it's not true that My belief about what You believe I intend (that I intend that we walk up the mountain) is compatible with what I believe about Your intentions (that You intend that we walk down the valley); that is, an H(5) equivalent is violated because our intentions don't mesh all the way up. Again, we can see why the meshing upwards assumption is needed. If walking down the valley (whether together or individually) is incompatible with walking up the mountain (whether together or individually), from my perspective why would You think that you can confidently predict that I will perform my part of walking down the valley? You would not, despite me having the actual intention that we walk down the valley. You therefore wouldn't see the matter of us walking down the valley as settled. Finally, if you only intend to walk down the valley if we do so together, you wouldn't even begin to perform your part, if we haven't started yet, or would abandon the performance, if we had.

The problem here is that there is little to no meshing of subplans at all. We could be generous and allow that there is a road that we must walk together a short way which eventually splits to lead either up the mountain or down the valley, so helping us get shared intention 'off the ground', though this would be kicking the proverbial can (as it were). Besides, this further strengthens the point just made, that Blomberg's conclusions only hold if we make additional assumptions about the compatibility of plans and sub-plans with the beliefs (false beliefs about false beliefs) he has in mind. It shows that the assumption that plans mesh all the way up for at least one agent's intention satisfaction is needed. What about if a large proportion of the walk is done together until the path splits? Perhaps they share an intention, as per Blomberg, on the way, but at some point, given the incompatibility between walking both up and down the valley, their plans will diverge.

In VALLEY WALK, the lack of common knowledge due to false beliefs does therefore seem to undermine the intentions and beliefs of the agents involved. If the shared intention is to persist in spite of false beliefs, as Blomberg is after, then what he needs to do is rule out intention and belief incompatibility. This would look something like H(5) earlier, which would depend on assumptions about sub-plans meshing both down and up. This is met in TOWER BUILD but not VALLEY WALK, a distinction Blomberg doesn't make; instead he equates the two by saying that findings from the former can help us in the latter, despite these differences.

2.5 The relevance of a belief requirement on intending

Blomberg does, though, provide a possible defence for his equating the two scenarios. He argues that "we arguably shouldn't accept a strong belief condition on intending" (Blomberg, 2016: pg. 320) according to which to intend to *A* an agent must believe they will *A*. This would undermine my argument that You see your intention that we walk down the valley as incompatible with what You take to be My intention that we walk up the mountain. I might go ahead with my actual intention of intending to walk down the valley, despite what I believe about your beliefs about me, as "[a]fter all, you could intend to do your bit of our walking down to the valley because you hope that I will change my mind and start to intend to do my bit of this joint walk" (Blomberg, 2016: pg. 320). This defence is reasonable and, in fact, the next chapter is dedicated to analysing the implications of a strong belief requirement on the possibility of shared intention in contexts with motivational uncertainty.

With respect to the discussion here, my view is that even accepting this, Blomberg's claim that rational intending is not undermined in the absence of common knowledge (and so shared intention is possible) is much less plausible in VALLEY WALK than in TOWER BUILD. First, there's arguably something very different between my *hoping* you will come to intend like I do in situations where a purported intention of mine (what I falsely believe you mistakenly believe my intention to be) is highly compatible with my actual intention—so our plans and sub-plans closely align and mesh, as in TOWER BUILD—versus when compatibility is low—little to no overlap of plans and sub-plans, as in VALLEY WALK. It seems much more plausible that when agents perform actions in parallel and which support the current intentions of both parties, their intentions might at some point become more closely aligned or even the same. But it's hard to see this occurring when agent's perceive

there's minimal overlap of their intentions, and even harder to see how any sort of collective activity can get off the ground in the first place, barring any failure in rationality.

Second, and following, the alignment of each person's primary goals also matters. In TOWER BUILD, Celia and Hector's purported intentions mesh all the way up to at least the satisfaction of his. His intention is fully satisfied by covering the block faces, which meshes with any sub-plans of Celia's intention that they build a tower together. There is no point in the future at which their sub-plans diverge and their intentions clash, so it's possible that shared intention persists. In VALLEY WALK, in contrast, My and Your intentions don't mesh all the way up to a point where, in that collective activity, at least one of our intentions is fully satisfied while the sub-plans of the other are entailed. In this case, at some point in the future our sub-plans will diverge and our intentions will clash. There is a lot more riding on Your hope that I will come to intend like You do in this case than there is for Celia hoping that Hector will come to intend like her. It's plausible to think that shared intention with false beliefs is more likely when the risk of goals remaining unaligned is lower, or at least that this should be a factor taken into account by a rational planning agent.

Third, part of Blomberg's justification refers to the fact that "Bratman [likewise] rejects a strong belief condition on intending" (Blomberg, 2016: pg. 320, fn 5). But even Bratman requires that intentions must still be based on beliefs reliable enough to provide a screen for action options (as I discuss in the next chapter). Whether 'hope' provides a robust enough basis to rely on and settle matters may depend on the extent to which sub-plans are expected to mesh or the stakes in play. There's far less risk to go ahead and intend when what I (mistakenly) take to be your intention is compatible with my own than when it's not.

Fourth, there are other general normative principles that must still constrain practical reasoning, including being guided by the available evidence for one's beliefs. In TOWER BUILD, it looks like Hector's false belief about Celia's belief about his intention can persist without a breakdown of rationality for either of them. He can maintain his false belief while each continues to perform their part of the intentional action they think they are performing together. In VALLEY WALK, it's arguable that there's greater pressure on Me to change my beliefs (that you believe I intend that we walk up the mountain and not, like you, that we walk down the valley) if we start walking down the valley; the evidence begins to point towards Me being comfortable that we walk down the valley which You, as a rational actor, should take into account if you truly had that belief about me. The more we walk together, the

closer it looks like a violation of rationality to continue holding the original false beliefs. It thus seems more plausible that there's ongoing shared intention with persistent false beliefs in TOWER BUILD than in VALLEY WALK.

This brings me to my final point. One important feature of Blomberg's proposal is that false beliefs *persist* while individuals share intentions. To repeat: "And such false higher order beliefs and doubts can persist throughout joint activities that at least appear to be jointly intentional" (Blomberg, 2016: pg. 319). A reason for this, I take it, is that if false beliefs don't persist, but are quickly corrected (e.g., by communicating), then we're back to having shared intention *with* common knowledge (presumably generated by the communication). In this case, it's hard to defend the initial interaction, in which false beliefs were present, as an instance of shared intention, rather than another type of interaction acting as a precursor to it. It may not be unusual that we are sometimes mistaken in what we believe others intend, but we sort this out early on, with shared intention as the outcome. Because of this, Blomberg's argument is weakened without persistent false beliefs, which is something that's plausibly part of TOWER BUILD but less so in VALLEY WALK, as per the previous paragraph.

2.6 Conclusion

The aim of this chapter was to address the immediate issue that jumps out from the proposal that individuals can share intentions though one or more of them is uncertain what the other intends. This is that such uncertainty seems incompatible with there being common knowledge of intentions. Given that common knowledge is an almost-universal requirement on most accounts of shared intention, this makes it hard to see how there can be both shared intention and motivational uncertainty.

To respond to this it's important to understand why common knowledge is thought to be important for shared intention. In the first part of this chapter, I make the case that one role common knowledge plays is to enable individuals sharing intentions to settle matters about what they intend and will do. Common knowledge justifies us saying that agents are in a position to settle matters both individually and together. This shows us what goes missing when we introduce substantial uncertainty about partner intentions; namely, that if we have this uncertainty then we no longer have common knowledge of intentions, which means we can no longer rely on common knowledge to provide the route by which intentions settle

matters. If it's possible that individuals can still settle matters in these contexts, we need to find another reason why they can be said to do so.

The next part of the chapter highlighted the value of keeping in mind this connection between common knowledge and settling, by focusing on Olle Blomberg's argument against the need for a common knowledge condition in minimal accounts of shared intention. Comparing the two examples he employs, I argue that though he takes them to be equivalent they do, in fact, differ in meaningful ways. Comparing them shows us that, for Blomberg's account to work, several assumptions must be made about compatibility constraints on individuals' sub-lans related to how they mesh both down and up. Still, Blomberg's proposal gives us insight into how agents can plausibly engage in shared activity even when there are potential unknowns about those involved, importantly without violations of rationality. One positive development is, to reiterate my point from before, that he pushes us to take a first- rather than third-personal perspective on shared intention. Rather than focusing on what's minimally required to actually be in place for there to be shared intention, he discusses what's minimally required from each agents' perspective. This forces us to confront real limitations faced by agents in what they know about their partners; that is, that private information is a reality of social interaction that cannot be assumed away. And because of this, it pushes us to answer a motivational question of why, in the face of these unknowns, agents nonetheless expect and rely on their partners to perform their parts of their collective activity.

Turning back to the focus of this thesis, of how there can be shared intention under motivational uncertainty, there are several useful takeaways from this chapter. First, the analysis opens the possibility of there really being shared intention without common knowledge. While Blomberg focuses on one specific reason for why common knowledge is absent, false beliefs, my thesis focuses on a different reason, namely uncertainty about intentions. Nonetheless, what he has shown is that it's plausible that individuals can have the right intentional attitudes to support interaction that looks, on the face of it, like joint activity involving shared intention despite the absence of common knowledge.

Second, there's an important connection between common knowledge and intentions settling matters. Reflecting on possible reasons for a common knowledge condition in shared intention, I claimed that one is that it provides the grounds for individual and collective settling, and a lens of collective settling shows us that we mustn't lose sight of the important role common knowledge plays in this process—and to keep in mind that arguing for dropping

the common knowledge requirement is not to argue for dropping the joint settling requirement. It highlights the fact that if it's true there can be shared intention in the absence of common knowledge, then there must be some other feature of the situation which is doing the work to allow individuals to settle matters both individually and jointly.

Finally, if we accept this, then it's also plausible that common knowledge is sufficient but not necessary for joint settling: it provides one but not the only route, if we think people can rely on something other than knowledge of their partner's intentions (and their partner's knowledge of their own intentions, etc.) to settle matters together about what they will do. If the purpose of a common knowledge condition is to justify settling, and if common knowledge is sufficient but not necessary for this, we could develop a more correct characterisation of the conditions necessary for shared intention by removing the common knowledge requirement and emphasising the settling requirements instead (already required, though often implicit and hidden in the background). Of course, there may well be other important reasons for requiring common knowledge, so while this is an interesting avenue of exploration, it's beyond the scope of my project.

In summary, I agree that there can't be both common knowledge and motivational uncertainty. I have, though, argued that this needn't undermine the possibility of shared intention in contexts where there's substantial uncertainty about intentions (a) if it's plausible that there can be shared intention without common knowledge, provided that (b) there's justification, given by several additional background assumptions, for why agents' intentions can still see the joint action as settled—an important feature of shared intention otherwise supported by common knowledge. I will return to the relevance and importance of collective settling in more detail in Chapter 4, but hopefully this assuages any concerns that the possibility of there being shared intention in BEACH is simply a non-starter, and that thoughtful reflection on the role of common knowledge in shared intention can explain why.

CHAPTER 3

A Belief Requirement on Shared Intention

The previous chapter raised questions about how there can be shared intention when there isn't common knowledge of intentions, a typical requirement on most authors' accounts. In response, I discussed Olle Blomberg's claim that agents can plausibly share intentions and engage in joint activities even in the absence of common knowledge, while arguing that a lens of intentions-settling-matters suggests that additional background assumptions about the alignment of agents' plans and sub-plans are necessary. If we accept this, then it seems we've addressed the first problem, raised at the beginning of this thesis, that motivational uncertainty seemed to pose for existing accounts of shared intention. It's plausible, in this case, that shared intention can persist even without common knowledge, provided certain quite narrow constraints which support settling are met.

This brings us to the second reason why theories of shared intention may face a problem when we introduce motivational uncertainty. Suppose we drop the requirement of common knowledge, it's still the case that Mya must have an intention about something to do with him and Iva going to the beach later today. But if Mya doesn't believe that they're going to do that, because he thinks it's more likely than not that Iva is watching the football, then it's not clear that he can have an intention at all. More generally, we might think that where there's a requirement on ordinary individual intention that intention entails belief, we would expect a version of that requirement on shared intention. The problem, now, is that this requirement may very well not be met in BEACH, given what Mya believes.

To explore this potential issue of a belief requirement undermining Mya and Iva sharing intentions, one approach is to look for insights from research on individual intentional action, to see whether the ways authors have proposed for reconciling intention and uncertainty there can provide some help for cases involving a collective. For example, we might think that if "I intend that I *A*" is compatible with uncertainty about whether I will *A*, then "I intend that we *J*" is compatible with uncertainty about whether we will *J*. If correct, then we could show that existing accounts of shared intention can accommodate the kind of motivational uncertainty present in BEACH.

3.1 Intending to *A* versus predicting I will *A*

Our guiding question is this: does an intention to *A* necessarily involve the belief that one will A? What we'll see is that there's a robust debate—and not much consensus—in the literature about the necessity of a belief requirement in intention. However, there hasn't been much discussion on if and how this belief requirement applies to *shared* intention as well. I engage extensively with Michael Bratman's work throughout this thesis, and now is a good time to both introduce part of his account and use it to explore the above question. His writing on individual as well as shared agency provides a comprehensive, carefully thought through body of work that hints at several different solutions to the problem of uncertainty under the microscope. Moreover, his generosity in engaging with competing viewpoints means that insights gained from analysing his work can potentially generalise; indeed, we will see that questions about uncertainty highlight more overlap between his and other very different accounts than we might at first think.

Bratman's early work on individual agency, *Intentions, Plans and Practical Reason* (1987), presents his view of intentions as plans. The ethos of Bratman's work is that human capacities to plan for future activity is something special, and that a theory of agency must both encompass and find the balance between

> "the centrality of planning in the constitution and support of fundamental forms of organization, and our important capacities for conceptual openness, spontaneity, and flexibility. And here it will be natural to … appeal to relevant practical virtues that are involved in well-functioning planning agency" (Bratman, 2014: pg. 24).

Bratman's broad goal is to identify what these 'practical virtues' are, which he does by locating them primarily in what he takes to be a core feature of practical agency; namely, intention, or, more precisely, intentional attitudes. He says:

> "The planning theory is a theory about the nature of intentions understood as central elements in this fundamental form of human, temporally extended agency. Such intentions bring with them a complex nexus of roles and norms that is characteristic of planning agency. And these structures go well beyond simple, temporally local desire-belief purposive agency. So it seems reasonable to see intentions, so understood, as distinctive elements of the psychic economy of planning agency. This is the distinctiveness of intention" (Bratman, 2014: pg. 24).

Bratman's approach is to detail the roles intention plays in agency, propose a set of intention-related norms, principles and mental processes and describe how these guide practical reasoning. Understanding what these are in their original form is crucial as, given his continuity thesis Bratman sees intention playing the same role in both individual and shared agency and draws exclusively on resources available almost exclusively in the former to explain the latter. His account of shared intention thus develops later, primarily in his collection *I Intend That We J* (1999b) and then down the line in *Shared Agency: A Planning Theory of Acting Together* (2014). At the heart of his account lies his conceptualisation of a "certain kind of public, interlocking web of [individual] intentions" (Bratman, 1999b: pg. 143) which plays roles distinctive of agents' shared intention to J, "roles such that it would be plausible to identify shared intention with what plays those roles" (Bratman, 1999b: pg. 142). This includes supporting the coordination of agents' intentional activities in pursuit of J, the coordination of their planning and structuring the relevant bargaining between them regarding their roles to play, actions to take, et cetera.

*

Going back to Bratman's early work on intention is necessary to understand where the settling and control requirements on intention mentioned in the last two chapters come from. Why is it necessary that my intention settle matters about what I will do?

Bratman's ideas are grounded in the way he says we talk about and typically observe intentional action, and his starting point is a focus on future- or forward-directed intentions, which lends itself to the idea that intention is functional in nature. His view is that, in reflecting on this, we see that our "commonsense framework [of the psychology of intentional action] sees intention as a distinctive attitude, not to be conflated or reduced to ordinary desires and beliefs" (Bratman, 1987: pg. 20). Intentional attitudes, he says, involve characteristic dispositions which support its specific functional role—namely, to facilitate planning over time—but which attitudes of desire and belief do not. Intention is thus a different sort of attitude and not reducible to belief, desire, or a combination of both. Intentional attitudes thus feature in the reasoning of ordinary planning agents who are rational and who make decisions dynamically in light of competing preferences. But that's not all. We must take seriously, he says, the fact that we are cognitively constrained, and that

> "so as not to use deliberative resources inefficiently, we frequently depend on general, nondeliberative habits and strategies about when to reconsider. And given a somewhat

> reliable environment, habits and strategies that to some extent favor non reconsideration will be likely, in the long run, to be conducive to the overall effectiveness of our temporally extended agency" (Bratman, 2014: pg. 22)

There are therefore two features of shared agency his account must address. First, scepticism about intention as a distinctive attitude—about there being such a thing as an 'intention' at all. Second, how agents make plans given their limited resources for attention, deliberation, calculation, and judgement. His neat idea is to interweave these two requirements: he incorporates the latter into an account of practical reasoning by seeing them as norms and principles by which intention, but not belief-desire per se, operate. In other words, an agent who has an intention reasons in the way characteristic of one who is reasoning within constraints described. And so we get both a plausible account of planning and one which distinguishes intention from other states of mind.

The norms and principles Bratman appeals to can be separated into what he calls norms of practical- and intention-rationality. The former involves traditional ideas about an agent's desires and beliefs at a certain time providing her with reasons for acting in various ways at that time—a desire-belief model of practical reasoning. It also specifies an 'intention-action principle' requiring that the present intention to *A* and the resulting action of intentionally *A*-ing must be tightly connected. In turn, various norms of intention rationality provide the mechanism by which we can link present reasons for action with future activity, and they include requirements about intention agglomeration, intention-belief consistency and assumptions about means-end reasoning.

The kicker, though, is Bratman's central idea that norms of intention rationality also involve a *commitment* to acting appropriately to the intention, as visible in this pithy description: "Intentions are conduct-controlling pro-attitudes, ones which we are disposed to retain without reconsideration, and which play a significant role as inputs into reasoning to yet further intentions" (Bratman, 1987: pg. 32). Unpacking this statement will illuminate what Bratman sees as two distinct dimensions of the kind of commitment characteristic of future-directed intention (see from Bratman, 1987: pg. 25). The first is a *volitional* dimension of commitment, which concerns the relation between intention and action. It derives from the idea that intentions are attitudes in favour of a particular course of action and are conduct controlling: if my intention persists until the time of action it will control my action then; it

will move me to act in ways appropriate to the intention[10]. The second is a *reasoning-centred* dimension of commitment, which concerns the role intentions play in the period between their initial formation and their eventual execution. This is constituted by 'characteristic dispositions concerning reasoning' over time, including dispositions to avoid reconsidering an intention and to use a retained intention as inputs into, and providing constraints on, reasoning about further intentions. These are grounded in our "capacity to act purposively… and… the capacity to form and execute plans" (Bratman, 1987: pg. 11), commonly taken to mean treating intentions as plans which guide and constrain our actions *as plans*, and not only because of the previous, original reasons for making them. Bratman supports this claim by observing that often we reason from prior intentions to further intentions: from intended ends to intended means towards that end; from more general to more specific intentions; or when prior intentions constrain other intentions formed later, such as to maintain consistency.

But why would it be rational for my previous intention to go to the beach to override my current desires to meet up for coffee with a friend—especially given everything else I feel and know now but which I didn't when first making plans? Bratman defends this by appealing to agents' bounded rationality, better overcome if we're disposed to retain our intentions without reconsideration. Furthermore, when we talk of intentions we talk of being settled on a certain course of action. If I form the intention to go to the beach this afternoon, this usually means I won't continue to deliberate about whether to go: "I will normally see (or, anyway, be disposed to see) the question of whether to go as settled and continue so to intend until the time of action. My intention resists reconsideration: it has a characteristic *stability or inertia*" (Bratman, 1987: pg. 28, author's emphasis). In the way they guide future

[10] Bratman tackles various challenges to his notion that intention is conduct-controlling in a way that 'ordinary desires' are not. How, for example, is intention's commitment to action different from a basic commitment to meeting one's desires, already included in a standard conception of practical rationality? These are different, Bratman says, because while both intentions and desires motivate us to act (in light of our beliefs), intentions are conduct-controlling whereas ordinary desires are merely potential influencers of action. Desiring an ice-cream for lunch, I would not be guilty of irrationality, he says, if, when lunchtime arrives, I decide against having one after weighing up this desire against a conflicting desire to lose weight. So, while my desire for an ice-cream might influence where I go for lunch, I might not even try to eat an ice-cream. Conversely, if I formed the intention to have an ice-cream and at lunchtime my intention remains, this intention will normally guide my actions, and I'll go ahead and order the dessert.

Intention thus controls, and not only influences, my conduct. But this could also be true if we think that we're treating intentions as attitudes that simply weigh more, qua intentional attitudes, than desires as inputs into practical reasoning. The problem with this view is that there's no particular reason why we should think this is true just by virtue of their being an intention and not a desire. The weather becoming unexpectedly hot might make an ice-cream suddenly very desirable—it's now up to our intuition if this desire trumps a previous intention to avoid an unhealthy lunch. Bratman's solution seems to be that a special kind of normative cognition is involved with intention: in the normal course of events, intentions provide immediate grounds for action without being weighed up against new desires. Though interesting, I won't take this discussion further as my focus in this thesis is on Bratman's second dimension of commitment.

action, intentions are not simply abandoned or revoked in the face of fluctuating desires. It's not that they cannot, but rather that it is characteristic of their ordinary working that they are generally not, or tend not to be, revocable. So, he says, we tacitly accept this particular norm of stability, so that in the normal course of things, lacking new reasons for reconsideration an intention is normally simply retained up to the time of action.[11] Settling thus grounds the commitment to action associated with intention; if I do not settle that I will *A*, I might still be disposed to deliberate about what to do in the interim and might not be moved to *A* when the time comes… and so I do not intend that I *A*. Without the matter being settled, I am not committed in the right way. In fact, an attitude that doesn't involve settling is, on Bratman's account, not an intention at all.

*

All of this tells us that my intention to *A* is not the same as my prediction that I will *A*. Most importantly, my intention to *A* involves my settling that I will *A*, which, in turn, means I am committed to *A*. My prediction that I will *B* does not mean that I settle my *B*-ing, and does not entail this characteristic commitment to *B*. Looking at the weather prediction for today, I can predict that later I will get thirsty and pour myself a glass of water without currently forming the intention and associated commitment to do so. Returning to the question of whether one can intend to *A* despite being uncertain whether one will *A*, this suggests that it might not be as simple as thinking of this as intending in light of a lower-confidence prediction that I will *A*. Whether it's still possible to commit oneself to an action one doubts one will succeed in performing is the subject of the next section.

However, first we should briefly reflect on whether the idea that intention and belief differ in terms of a settling requirement extends beyond Bratman. An alternative idea to focusing on cognitive constraints is to refer again to the processes by which each is formed. Recall from Chapter 1 the difficulty in explaining how practical reasoning leads to intention modification: "agents who have resolved the question of what they ought to do still have a question to settle, about what they are going to do" (Wallace, 2020). Elizabeth Anscombe's treatment of this is perhaps one of the clearest. She argues for "a difference of form between reasoning leading to action and reasoning for the truth of the conclusion" (Anscombe, 1963:

[11] What counts as a good reason or what triggers reconsideration is something which, unfortunately, Bratman himself gives us little idea about, as I discuss in Chapter 6.

pg. 58). It's important, she says, to see the processes involved in practical and theoretical reasoning as non-identical. If we do not, she states,

> "the disadvantage, so far as its being practical is concerned, [is] that though the conclusion is necessitated, nothing seems to follow about doing anything… It is obvious that I can decide, on general grounds about colouring and so on, that a certain dress in a shop window would suit me well, without its following that I can be accused of some kind of inconsistency with what I have decided if I do not thereupon go in and buy it; even if there are no impediments, such as a shortage of cash at all" (Anscombe, 1963: pg. 57).

Anscombe's point is that we need an explanation of the additional step from ought to action.[12] If we adopt something like Broome's proposed *Enkrasia* requirement earlier, then one idea is that, having judged that *X* is best for me to, in *deciding* to *X* I form the intention to *X* and settle on a course of action. And in deciding to *X* I settle that I will *X* in the future. Settling is bound up here in deciding, and is essential to intention formation. And it is part of the authority I have to decide that I do something that I see it as up to myself to do. Making a prediction (i.e., forming a belief about something occurring) involves no such *Enkrasia*-like normative requirement; there is no sense in which the conclusion of my theoretical reasoning is something that pushes me to decide to act. This is just to say again that the idea that my intention to *X* is not only my prediction or expectation that I will *X* or am *X*-ing is widespread on even quite different authors' views.

3.2 Empirical support for Bratman's commitment-related norm of stability

As I'll be reflecting on Bratman's concept of commitment throughout this thesis, before turning back to deal directly with uncertainty about partner intentions I should assess his

[12] Anscombe continues with an acerbic take on any attempt to look for solutions which rely on logical reasoning:
> "The syllogism in the imperative form avoids this disadvantage; someone professing to accept the premises will be inconsistent if, when nothing intervenes to prevent him, he fails to act on the particular order with which the argument ends. But this syllogism suffers from the disadvantage that the first, universal, premise, ("Do everything conducive to not having a car crash") is an insane one, which no one could accept for a moment if he thought out what it meant. For there are usually a hundred different and incompatible things conducive to not having a car crash; such as, perhaps, driving into a private gateway immediately on your left and abandoning your car there.
> (...)
> Thus, though general considerations, like 'Vitamin C is good for people' may easily occur to someone who is considering what he is going to eat, considerations of the form 'Doing such-and-such quite specific things in such-and-such circumstances is always suitable' are never, if taken strictly, possible at all for a sane person, outside special arts" (Anscombe, 1963: pg. 58–59).

theory's validity. One approach is to look for empirical support, which would actually align with Bratman's own, given that he starts with a descriptive and moves to a theoretical explication of agency:

> "For present purposes, however, we can rest content with a pair of ideas: First, the planning theory involves both a descriptive account of the underlying, accepted norms, and an account of the normative force or significance of those norms. Second, we can understand this normative significance both by appeal to the importance of the general forms of functioning the acceptance of these norms supports, and by appeal to the distinctive, non-instrumental significance of the satisfaction of these norms in the particular case" (Bratman, 2014: pg. 17).

Recall that Bratman's commitment has two dimensions. The first is volitional, referring to an intention's conduct-controlling nature. There are interesting empirical questions about whether agents experience intention like this (see for example, Pacherie's, 2007, discussion on the sense of control and sense of agency). However, whether in BEACH Mya and Iva share an intention has precisely to do with perceptions of whether each continues to intend as they had before, given changes in available options. This is less to do with the experience of agency and more to do with the second, reasoning-centred dimension of commitment, concerning an intention's disposition to be retained and not reconsidered—his norm of stability. I'll thus focus on empirical support for this.

Though Bratman's multi-dimensional account of commitment hasn't been the direct subject of experimental testing, his view is based on a common sense view of agency and there's research in the social sciences which tests related ideas. Most obvious would be studies on public pre-commitment, with Schelling's (1960) treatment in contexts of strategic bargaining a notable early work. More recently, there've been multiple studies in psychology and economics on pre-commitment to a goal and its use in effecting behaviour change, such as for overcoming procrastination and improving savings rates (for a brief overview, see Sunstein, 2014). One view is that pre-commitments work by overcoming issues related to future discounting (Kurth-Nelson & Redish, 2012), but another reason, and one which better aligns with Bratman's view, is due to the well-documented phenomenon of *status quo bias* (Samuelson & Zeckhauser, 1988). This refers to individuals' preferences for maintaining their current or previous state of affairs, or avoiding taking actions to change this state. It's possible that several 'non-rational' cognitive processes drive this, including loss aversion, the

endowment effect, sunk cost thinking, regret avoidance or a need to feel in control. A different school of thought, though, is that status quo bias can be rational in the face of informational and cognitive constraints: automatically sticking with what is known or has worked in the past can be easier and safer than trying something new and unknown.

There are various explanations which adopt this evolutionary psychology perspective in the context of judgements under uncertainty, in which decision outcomes—and hence the utility they bring—are often uncertain. Early on, Herbert Simon (1956) made the point that we needn't presuppose a utility function or make any demands on psychology (e.g., to calculate marginal rate of substitution) to make sense of an organism's survival in a typical environment. As long as previous decisions are 'good enough', sticking to them can yield a high probability of survival. Simon's later conceptualisation of bounded rationality (Simon, 1982), which emphasised deviations from traditional views of rationality due to limitations in our thinking capacity, available information and time, helped provide a framework for investigations into, among other things, status quo bias just mentioned. For example, where there are informational limitations, Haselton and Nettle (2006) predict that a history of asymmetric costs of false positive relative to false negative errors should favour a bias towards making the least costly error, which may favour what is known to have worked in the past. Research focused on the cognitive cost of choice has found that people are more likely to postpone decisions or avoid change when more alternatives are added to a choice set, because of the increasing complexity of making a new decision.

Moreover, there's evidence that the increased mental effort of attending to status quo alternatives can lead to a superior choice's benefit being outweighed by decision-making costs (Dean et al., 2017). And status quo bias can even strengthen over time if, for example, its mere existence and longevity is taken as a prima facie case for goodness (Eidelman & Crandall, 2012). Nebel (2015) argues that, in situations with high uncertainty and high deliberation costs, status quo bias cannot be criticised as irrational on subjective theories of rationality. The author presents an objective theory holding that a conservative bias towards existing things we value is rational. Finally, there's interesting evidence from neuroscience which supports a "regret-induced status quo bias", in which people experience a greater feeling of regret when making errors after rejecting rather than accepting a status quo option (Nicolle et al., 2011), with the effect stronger for difficult relative to easy decisions (Fleming et al., 2010).

*

The above provides limited evidence in support of Bratman's view that it can be rational to stick with an intention given cognitive limitations and task complexity. But there's still an important step between what the above studies are testing and what Bratman is after. These studies test a willingness to maintain the status quo in light of attractive alternatives an agent *is aware of and faces choosing between*. Generally, the agent is deciding between an option whose value is known and an option whose value is not known, while the cognitive costs of switching are manipulated. Bratman's proposal is slightly different: his norm of stability provides resistance to the reconsideration process itself—rather than actual reconsideration across options with a bias towards the status quo. This suggests a more accurate parallel with Bratman's work could be found.

One promising area of research is in the field of computational rationality, which has a good deal of conceptual overlap with Bratman's view of intentions-as-plans. Computational rationality is a sub-theory part of a broader literature of computational approaches to cognition, developing models of agent planning built on base processes for perceiving, predicting, learning and reasoning under uncertainty. A lens of computational cognition

> "[involves the] development of computational representations and procedures for performing large-scale probabilistic inference; methods for identifying best actions, given inferred probabilities; and machinery for enabling reflection and decision-making about tradeoffs in effort, precision, and timeliness of computations under bounded resources" (Gershman et al., 2015; pg. 273).

The guiding principle is that agents maximise utility while taking into consideration the costs of computation. There are, of course, various ways of building models of cognition, of which probabilistic models provide perhaps the clearest overlap with Bratman's methodological attempt to characterise norms guiding intentional action. Both are, in terms of Marr's three levels of analysis, situated at the computational level, characterising the problem faced by the mind and how it can be solved in functional terms. They identify *ideal* solutions to cognitive obstacles, presenting them as a set of norms and principles required to solve challenges faced by a decision-making agent. These are a starting point for model development:

> "Probabilistic models of cognition pursue a top-down or 'function-first' strategy, beginning with abstract principles that allow agents to solve problems posed by the

> world—the functions the mind performs—and then attempting to reduce these principles to psychological and neural processes. Understanding the lower levels does not eliminate the need for higher-level models, because the lower levels implement the functions specified at higher levels" (Griffiths et al., 2010).

However, before exploring research on computational rationality, a first step is to ask whether even broader, more basic computational models of planning have any validity. There's a large literature on this, so two studies will help illustrate the general applicability of the model. Jara-Ettinger et al. (2016) present a commonsense psychology account of planning based on a naïve utility calculus, reviewing multiple studies for evidence in support. They argue that the simple model captures much of the rich social reasoning humans engage in, even from infancy (see also Perfors et al., 2011; Gopnik & Bonawitz, 2015), suggesting that

> "human social cognition is structured around a basic understanding of ourselves and others as intuitive utility maximizers: from a young age, humans implicitly assume that agents choose goals and actions to maximize the rewards they expect to obtain relative to the costs they expect to incur" (Jara-Ettinger et al., 2016; pg. 589).

A similar approach to the study of adults is also supported. Baker at al. (2006, 2009) present intuitive models of action understanding, proposing

> "a computational framework based on Bayesian inverse planning for modeling human action understanding. The framework represents an intuitive theory of intentional agents' behavior based on the principle of rationality: the expectation that agents will plan approximately rationally to achieve their goals, given their beliefs about the world… Our model captured basic qualitative inferences that even preverbal infants have been shown to perform, as well as more subtle quantitative inferences that adult observers made in a novel experiment" (Baker et al., 2009: pg. 329).

So even simple computational approaches involving basic models based on desires and beliefs help explain planning behaviour, supporting the assumption that agents are guided by maximising their net reward (see Vlaev et al., 2011, for an evidence-based review of this and models with other value calculation strategies). It also suggests that agents infer from others' actions that they plan in the same way.

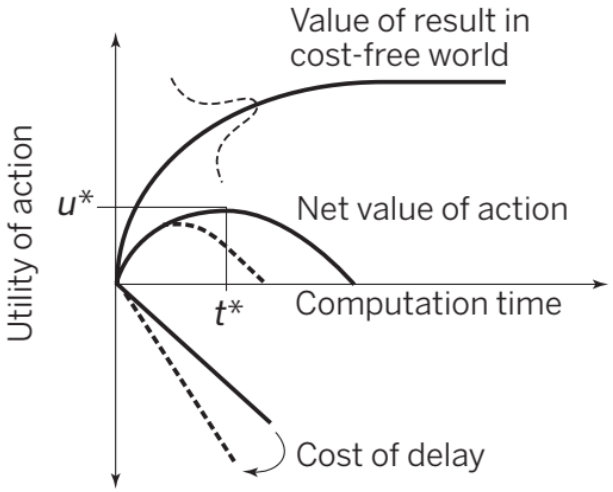
However, while the simple reward-based model is not overly controversial, it does face an important theoretical and computational challenge: namely, how to shift from the

traditional view of rational utility-maximising agents, deciding how to act in an environment that generates an enormous (and exponentially increasing) set of options, towards one of resource-bounded agents operating in real time and under significant computational constraints. Bratman's motivation for the distinctiveness of intentional attitudes echoes this. Computational rationality approaches address this concern by augmenting the existing probabilistic models described with the goal of "identifying decisions with highest expected utility, while taking into consideration the costs of computation in complex real-world problems in which most relevant calculations can only be approximated" (Gershman et al., 2015; pg. 273). The key idea is that the choice of how best to approximate—that is, the choice of which probabilistic model to use—is itself a decision subject to expected utility calculus. Because deliberation is costly, there is value in a psychological mechanism or process that optimally allocates time and costly cognitive resources. Models of computational rationality thus try to build this intelligence in by including principles that guide inferences and decisions at the 'metalevel' to regulate those at the base-level. These models thus incorporate a form of meta-reasoning—or, reasoning about reasoning—which applies rational decision-processes to the decision process itself (Griffiths et al., 2019). For instance, mechanisms that adaptively select between model-free and model-based systems to balance computational tradeoffs in response to changes in the environment (Daw et al., 2011).

An example of the economics of a computational rationality approach is shown in the image below. The figure shows how the decision-making agent must consider the expected value and cost of computation to find the ideal balance between, on one hand, the cost of effort or delay of additional computation—for example, attending more closely to the problem or taking longer to make a decision—and, on the other hand, the expected value or quality of the action chosen. In an ideal world, free of time pressures and little decision-making effort, the decision-maker would keep computing—for example, reasoning through the problem, weighing up competing options—until the utility curve of taking an action flattens out and there's little additional benefit to be had, and then making a decision on how to act. But introducing costs to the process (here assumed to be constant, a straight line) means that at some point the improved value of additional time spent deciding is outweighed by these costs, where sticking with the decision problem incurs a net cost (where the net value curve intersects the computation time axis). Furthermore, flexible computation processes enable us to identify ideal stopping times (seen at t*) to optimise this net value of taking action (u*). In short, this example shows how "the time at which further refinement of

the inference should stop and action should be taken in the world are guided by computation of the expected value of computation" (Gershman et al., 2015; pg. 275).

Figure. Economics of thinking in computational rationality



Source: Fig. 2—Gershman et al. (2015; pg. 275)

A relatively new and growing body of evidence supports the validity of these models. Lieder et al. (2014) find that 'rational metareasoning' holds promise as a framework both for inferring how people choose among cognitive strategies and for incorporating findings into improved solutions for the algorithm selection problem—that is, for finding the optimal stopping point described earlier. And Gershman et al. (2015) provide a nice summary of evidence for the application of models of computational rationality to various domains of cognition. Recent attention has been given to a proposed framework of 'resource-rational analysis' of agent planning, which allows for "[deriving] rational models of people's cognitive strategies from the assumption that people make rational use of limited cognitive resources" (Callaway et al., 2018; pg. 178). Lieder and Griffiths (2020) provide a comprehensive treatment of the validity of these models given currently available evidence, as well as commentary on their application across a range of domains. They suggest that the

> "integration of rational principles with realistic cognitive constraints makes resource-rational analysis a promising framework for reverse-engineering cognitive mechanisms and representations … that resource-rational models can reconcile the mind's most impressive cognitive skills with people's ostensive irrationality" (pg. 1).

*

While the evidence presented in this section is not meant to be a robust defence of computational models of cognition more generally, and computational rationality approaches in particular, it does provide support for Bratman's view of the type of commitment in planning, which takes as a starting point an assumption similar to one that "people's cognitive strategies are jointly shaped by function and computational constraints" (Callaway et al., 2018; pg. 178). To better see this, the table below outlines some broad conceptual features of Bratman's characterisation of the norms of intention rationality as well as a general resource-rational framework for understanding agent planning. We can see, I think, a promising overlap that could provide a fruitful future avenue for exploring how these accounts can inform one another. For example, of relevance to my project, resource-rational analysis already hints at an answer to the issue, identified earlier, that Bratman says nothing about when and under what conditions agents should reconsider their intentions. These models show that algorithms to select the best approximation strategy will be rationally adapted to the agent's needs and environment, evolutionarily or online using metareasoning (Gershman et al., 2019). Perhaps it's the case that the more likely one's partner is to face tempting attractive options the more likely an agent will reconsider their intentions, all else equal. For now, this is likely sufficient to conclude that we can move ahead with Bratman's version of commitment and intention.

Table. Conceptual comparison between Bratman's proposal and computational rationality approaches

	Bratman's norms of intention rationality	**Resource-rational framework for planning**
Agents are planners	Understand agency in terms of planning, intentional action guided by beliefs, desires and intention	Understand agency in terms of rationalistic probabilistic planning and maximising expected utility
Top-down approach	Start with general principles (norms) guiding practical reason; processes not discussed	Start with computational principles facing cognitive challenges; processes not discussed

Cognitive limitations guide model	Pragmatic rationale: planning is constrained by limited capacity to deliberate	Cognitive strategies depend on rational use of limited cognitive resources
Cognitive solutions	Commitment characteristic of intentional attitudes entails persistence and resistance to reconsideration of intentions	Planning can be effectively approximated or simplified by algorithms that adjudicate between more or less complex models
"On the one hand, we need to coordinate … and we need to do this in ways compatible with our limited capacities to deliberate and process information… This argues for being planning creatures. On the other hand, the world changes in ways we are not in a position to anticipate […]" (Bratman, 1987; pg. 43) "[… so we must] explain how human decision-making … can be so accurate and so fast yet still flexible to replan when circumstances change—the essence of acting intelligently in an uncertain world" (Gershman et al., 2015; pg. 278).		

3.3 Intention, uncertainty and Bratman's Asymmetry Thesis

The first section in this chapter argued that intention is not the same as mere prediction. Even a high confidence in success does not guarantee the kind of commitment to action that intention requires. However, the fact that intention settles matters does usually support me taking the means to enable me to *A* and seeing me as bringing about my *A*-ing when the time comes. This means that if I have settled on my *A*-ing, then it's hard to see how I can be uncertain whether I will *A*, barring unexpected difficulties which prevent me from doing so. And if I intend that we *J*, I should take it as settled that we *J*, so how can I also be uncertain about our *J*-ing? While predicting is therefore not intending, the latter seems to require at least believing that I will *A* or that we will *J*. This provides a potential obstacle to thinking that there can be shared intention in BEACH, given Mya's uncertainty about whether Iva intends to join him at the beach, and so his uncertainty that they will go to the beach together.

Like in the last chapter, one idea is to turn to the field of *individual* intentional action to look for possible ways to overcome this. In fact, whether an intention to *A* entails a belief

that one will *A* is, as mentioned earlier, subject to robust debate in this literature. Though a full exposition is beyond the scope of this thesis, we can nonetheless look for insights by getting a sense of the main arguments on each side.

Answering 'no' to whether intending to *A* requires believing one will *A* raises several concerns. First, more abstract, is that it's strange for me to say "I am going to dive to the bottom of the dam this weekend, but I probably won't", a strangeness which can partly be explained by us generally treating intention as a type of belief, or at least something closely tracking it (see e.g., Grice, 1971). Bratman, likewise, says that an intention to *A* usually supports the belief that one will *A*, as …

> "... the combination of … two dimensions of commitment help explain how intentions play their characteristic role in supporting coordination, both intrapersonal and social. Both the inertia of intention and the fact that it is a conduct-controlling pro attitude provide support for the expectation that when the time for action comes, an agent will at least try to do what she intends to do. Further, the dispositions to figure out how to do what one intends, and to settle on needed preliminary steps, provide support for the expectation that an agent will both be in a position to do what she intends and succeed in doing it" (Bratman, 1987: pg. 29).

This can also be interpersonal, such that my intention supports your expectation about how I will act, and vice versa.

A second, more specific, concern is that intention without requiring belief in action success might make it rational for me to intend an action which I think there's only a very small chance I'll perform. It could accommodate, for example, my intention to dive, later this year, in a bathysphere to the bottom of the Challenger Deep. It's hard to define a set of rational principles that would differentiate, when it comes to making plans, between this intention of mine and another intention I have to learn to SCUBA dive in the same time period. It would not be irrational to take means to either of these ends, for my intention and beliefs would still be coherent and consistent.

Conversely, the worry in answering 'yes' is that it seems to ignore the observation that we routinely deliberate and settle matters about what we will do, form intentions and make plans about means to chosen ends despite not being fully certain we will carry them out. I intend now to join a boat trip next week to visit a seal colony (even getting excited

about possibly observing seal pups), though there's a good chance the excursion will be called off due to stormy weather. I nonetheless plan for the trip, drawing cash to pay the captain and making allowances for arriving home late. Though I lack full confidence the trip will proceed, we would be hard pressed to say that I don't intend to go to see the seals next week, given how all of my planning revolves around this. Manchester United's goalkeeper intends to beat the record for number of clean sheets (games with no goal conceded)—he trains extra hard and is wholly fixated on this task—despite being unsure if he will be able to stop the opposition from scoring in the upcoming matches.

However, intentions do normally guide planning partly by providing support for expectations that the intended action will be successfully carried out. We also take it that intentions and beliefs should be consistent. These together suggest that an intention should normally *not* support the expectation that one *won't* act as one intends, but not that the two statements do not actually entail this. Bratman shows this in examples involving an intender who is agnostic whether she will even try when the time to act comes:

> "I might intend now to stop at the bookstore on the way home while knowing of my tendency toward absentmindedness… If I were to reflect on the matter I would be agnostic about my stopping there, for I know I may well forget. It is not that I believe I will not stop; I just do not believe I will" (Bratman, 1987: pg. 52).

Or she is agnostic about whether she will succeed even if they try:

> "Perhaps I intend to carry out a rescue operation, one that requires a series of difficult steps. I am confident that at each stage I will try my best. But if I were to reflect on the matter, I would have my doubts about success. I do not have other plans or beliefs which are inconsistent with such success; I do not actually believe I will fail. But neither do I believe I will succeed" (Bratman, 1987: pg. 52).

In addition, this need not be limited to thoughts about future-directed intention, but could apply to present-directed intention, or intentional action, as well. Donald Davidson's observation about a copywriter—trying hard to make ten carbon copies as he writes, doing so intentionally despite being unsure that he's succeeding—is that we might sometimes perform an action with an intended outcome we are unsure we are achieving.

In what follows in this thesis, I choose not to rule out the possibility of shared intention when there's substantial uncertainty about shared activity success. I have two

reasons. First because, as per the discussion above, though it's common that when I intend to *A* I expect to *A*, I find it plausible that this need not always be the case. Of course, I might typically try to avoid doing things I am not sure I'll be able to do, seeing this as a waste of time and effort. But there are times when embarking on action despite no guarantee of success might be useful, as when practising or learning something new, or desirable, if the stakes of not trying or doing nothing are very high. Second, because of taking seriously the need to explore a first-personal perspective of shared intention, as I proposed at the outset. It is a (if not the) defining feature of shared intentional activity that we need to recognise and explain it in terms of some organisation of the attitudes of several, different individuals who, from a first-personal perspective, don't come to know their partners' attitudes in the same way as their own (at least those accounts which avoid metaphysical additions like group or plural agents). This requires acknowledging that there is *some* process for learning what others' attitudes are which must admit some degree of freedom in what's known for sure. To rule out from the get go any uncertainty in what we know about others would therefore avoid answering a crucial question about how there can be shared intention at all. Going forward, then, I embrace the view that not every intention of mine to *A* involves my belief that I will *A*. We now need to reconcile how I can settle on matters being done while being uncertain about this being the case.

*

The positive and negative answers to the belief requirement on intention suggest that it's going to be a challenge to reconcile two general needs. On the one hand, we must allow for intentional action where there's reasonable doubt about success—which we might call weak belief, low or weak credence, a low degree of confidence, low subjective probability, being uncertain—while, on the other, ensuring the bar for what seems reasonable is set high enough to rule out the kinds of strange, close-to-irrational behaviours described earlier.

One intuitive solution is to consider a threshold level of belief certainty, above which it's rationally acceptable to incorporate this belief into planning. This may be accompanied by additional rational pressures, like a nudge to make contingency plans as when Mya believes there's a 70% chance of rain he takes his surfboard to the beach in addition to his beach bats (he enjoys surfing in the rain). Though individual preferences might well influence the threshold risk likelihood for making contingency plans, this could hold in general. We are, however, in deceptively tricky waters when we start to consider belief in intention in this

way. To see this, consider the challenge Bratman faces when taking a likewise pragmatic view of intention under uncertainty. What his account seems to require is that an agent's planning takes place against background beliefs which are, in his words, "all-or-nothing" or "flat-out" beliefs, and not merely degrees of confidence or subjective probabilities (between 0 and 1) (Bratman, 1987: pg. 51). Flat-out belief must come as part of the package if intentional attitudes are truly distinguishable from belief and desire in the way they settle matters. For this is partly what explains the way intention provides a screen for filtering options inconsistent with one's existing intention or current beliefs, and so promotes one taking specific means to ends on which one has settled. This function is incompatible with beliefs which merely involve high probabilities (<1): "What distinguishes flat out belief from the assignment of some probability is, partly, the role it plays in further planning—notably, the role in "providing a screen of admissibility for my options."" (Bratman, 1987: pg. 51).

The problem is that there's no reason why a belief, even one with a very low level of confidence, should screen out any option (this is similar to the worry in answering 'yes' earlier). There is, for example, no issue of inconsistency if we treat belief like a subjective probability. If (to paraphrase Bratman) I have only one surfboard, then a plan to both leave it at home while taking it to the beach wouldn't be inconsistent if I simply assigned a high probability to the proposition that I have only one surfboard, rather than flat-out believing I have only one surfboard. Likewise, it wouldn't be inconsistent if I planned to both take a surfboard to the beach while also leaving one at home for my friend to pick up on her way, if I simply assigned even a tiny probability to the proposition that I have more than one board (perhaps I faintly recall my brother suggesting he'd once left one). It's the fact that I have a flat-out belief that I have only one surfboard that makes it inadmissible for me to plan to leave a surfboard at home, for my friend to pick up, once I've decided to take it to the beach. It doesn't make it irrational to plan in light of beliefs with only the smallest hint of being correct. Without flat-out belief we don't have grounds for norms—like the norm of consistency—which play a core role in Bratman's idea of intentions as plans.

*

It's useful to summarise several of the claims just made:

(1) My intention usually supports an expectation that I will act appropriately to that intention.

(2) Agnosticism about whether I will actually try or even succeed if I do, and cases involving learning or when stakes are high, means we should plausibly allow for cases where there is uncertainty in my expectations in (1).

(3) Intention must provide a (normative) screen for admissible options (i.e., they involve norms of consistency and stability), a function which requires flat-out belief and which is incompatible with beliefs thought of as subjective probabilities (as to the truth of their propositions).

(4) It's irrational for me to intend in ways inconsistent with my beliefs.

Accepting these means we are still left with needing to explain how uncertainty in intention is reconciled. Bratman does this by providing a neat solution via the introduction of what he calls an Asymmetry Thesis (AT), which says:

(5) My intention to *A* is compatible with doubts I have that I will *A*, but which is incompatible with my belief that I will *not A*.

In other words, intending without believing in action success—'intention-belief incompleteness'—is fine, but intending while believing in action failure—'intention-belief inconsistency'—is not, the latter being "closer to criticizable irrationality" (Bratman, 1987: pg. 53). As always, Bratman's rationale is rooted in how these separate attitudes contribute to planning: "One good reason for accepting the asymmetry thesis is that intention-belief inconsistency more directly undermines coherent planning than does intention-belief incompleteness" (Bratman, 1987: pg. 53). A planning agent who merely doubts success can, for example, make contingency plans in case of failure, this being a practical necessity as they're not in a position to settle and plan on being either successful or not (in fact, not making contingency plans might be a criticisable form of irrationality itself). In contrast, to believe I will fail means it would be strange to plan as if failure were not an option. I am in a position to settle and plan on the belief that I will fail, so it would seem odd and inefficient to plan for the possibility I will succeed.

3.4 Mapping the Asymmetry Thesis to the shared case

The AT looks like it can be used to address the problem that Mya's uncertainty about Iva's intention precludes them sharing an intention because of relevant belief requirements. The simple application of the AT to the joint case would support the idea that Mya and Iva can

share an intention if Mya is merely uncertain whether Iva intends to meet him at the beach—even if the level of uncertainty is substantial—but not if Mya believes Iva has diverted to the pub to watch football and does not intend to meet him at the beach.

There are, arguably, other routes we could take instead of using the AT. We might reject outright that one can intend while uncertain about action success—that is, stick to a strong belief requirement on intention—in which case it's just impossible for Mya and Iva to share intentions. Or we might accept a weakened belief requirement on intention but reject Bratman's AT, and include the possibility that, in certain circumstances, one can rationally intend to *A* even while believing, rationally, that one will not *A*—and that this combination is rational and not inconsistent. Whatever the motivation here for allowing that one can intend to *A* and yet not believe that one will *A*, it's hard to see how this motivation would differ in the individual versus the shared case, and so hard to see how uncertainty about partner intentions would be a special problem at all.

If, however, we take it that Bratman's AT is a valid norm of planning in the case of individual intentional action, it's plausible that it's likewise valid in the shared case. This gives us the potential solution above, in which Mya and Iva's shared intention that they go to the beach is consistent with either or both of them harbouring doubts about whether they actually will go to the beach together. Though Mya might be uncertain about how things will pan out, as long as there isn't an obvious conflict with his beliefs, then there isn't a problem of inconsistency. In Bratman's view, it looks like everything is fine in the case of shared intention; it will be fine for Mya to go ahead and see the matter as settled. There isn't, then, a special problem of uncertainty about partner intentions in joint action that isn't already addressed when we take into account uncertainty generally present in planning for the future.

*

Before concluding, I want to temper enthusiasm about the readiness of applying the AT to joint action contexts. It's possible that in directly mapping it (and other tools) from individual to joint action I have forced it to fit. To see why this might be inappropriate, or at least why we might consider other questions raised by its application, compare these three statements.

A) I am uncertain whether my intention to *A* will translate into my *A*-ing.
B) I am uncertain whether our intentions that we *J* will translate into our *J*-ing.
C) I am uncertain whether you intend that we *J* (as I do), so am uncertain that we will *J*.

What's noticeable is that there's not such straightforward mirroring from uncertainty about action success—the AT's original target—to uncertainty about partner intentions. This chapter focused on what is actually best described as bringing along the AT in a shift from (A) to (B). However, the motivational uncertainty we have in mind is best captured by (C), which requires an additional step to get to the point where we can apply the AT; namely, from being uncertain about partner intentions to being uncertain about joint action success. The step is plausible, as it's hard to see how I could be certain about action success despite being uncertain about your intention to play your part, so it's not to say the AT's application here is invalid. Nonetheless, it highlights the subtle difference between uncertainty about our joint action success despite our willingness to perform it versus uncertainty about our willingness to perform it in the first place.

Another way of putting this concern is to consider two, of potentially many, different sources of uncertainty Mya may have which are relevant to BEACH. Mya is uncertain about whether Iva intends that they go to the beach together, and he's also uncertain whether it'll rain once they're there. Let's generalise and call these examples of uncertainty about intentions versus uncertainty about other facts about the world. I've used the AT to explain how there can be intention in the face of both sources of uncertainty. But note that to do this we must be treating Mya's beliefs about Iva's intentions in the same way as Mya's beliefs about other facts about the world, at least in the ways they are relevant for his intending. If applying the AT is valid in case of both sources of uncertainty, it means we are assuming that when sharing intentions we treat our partner's intentions just like any other fact about the world. We are, in a way, treating their intentions like other facts about them; they have their intentions, they are also a certain shape, size and mass … a mass which, for example, makes it useful that they will be able to prop open a door (but not stop an oncoming bulldozer). On this view, their intentions look like they are features of them as objects. Moreover, it looks like we're saying there is no interesting difference between beliefs about partner intentions and beliefs about other facts about the world.

Why might we hesitate to accept this? This is awkward and difficult to answer at the moment, given only what I've introduced in the thesis thus far, but the broad idea might be that this would violate something important about the nature of the social interconnection we typically think of as essential when sharing intentions. More specifically, my treating your intentions like other facts about the world may not give us an account in which our intentions are 'genuinely' or 'properly' shared, and that we could end up allowing for 'strategic' rather

than intentionally joint action. Though I raise this now, going beyond this intuition and pinpointing what is wrong with this approach takes some work and is an important focus of the upcoming chapters.

3.5 Conclusion

This chapter aimed to show that Bratman's Asymmetry Thesis removes one problem about motivational uncertainty undermining the possibility of shared intention. The problem has to do with the relationship between intention and belief. For example, some authors think that in order to walk to the village, I have to believe that I will walk to the village. Generalising this requirement means that if I intend that we will walk to the village, then I believe that we will do so. This, of course, is immediately blocked by the presence of motivational uncertainty, meaning it is impossible for me, if I am uncertain about your intention to walk to the village with me, to intend that we will. And so, irrespective of other concerns about common knowledge or something similar, motivational uncertainty blocks shared intention. If, however, we drop a strict belief requirement on intention, then we can draw on the AT to help us resolve the issue. The AT tells us that my intention to walk to the village is compatible with doubts I will successfully do so but is incompatible with believing that I will not walk there. If we generalise the AT to cases of shared activity, as I've done in this chapter, then we can see how my being uncertain that we will walk to the village is compatible with my intention that we do so, as long as I don't believe that we will not.

Having a strong view about intention and belief therefore means there's going to be a problem in explaining shared intention in contexts with motivational uncertainty—all kinds of uncertainties, in fact, as this is a more general issue. But if it's reasonable to bracket this, because the AT or something like it seems well supported, then there's a reasonable case to be made that shared intention is compatible with even substantial uncertainty about partner intentions. Moreover, adopting the AT also gives us a good way of thinking about exactly when motivational uncertainty is or isn't a barrier to shared intention; namely, that motivational uncertainty, per se, is not the obstacle, but rather the extent to which it undermines beliefs about action success. This also means that shared intention in contexts of uncertainty needn't involve the idea that agents must have a special attitude or belief.

At this point, we appear to have an answer to how it is that agents can share intentions yet still be uncertain about what one another intends. Our job is, in essence, complete. We

must however, be careful, as just applying the AT doesn't seem to capture the distinction between dealing with facts about the world and facts about interaction partners' intentions (and even other attitudes). If we stand by the intuition that there is something special about the latter when individuals share intentions, such as a special sense of sharedness or jointness between those involved, then the AT by itself looks insufficient. There's now a separate problem that just appealing to the AT doesn't answer. However, in analysing the applicability of the AT, we do learn something crucial: namely, that there can only be a problem about uncertainty in shared intention if it's specifically linked to an agent's perspective which treats beliefs about a partner's intentions differently to beliefs about facts about the world. This provides a clue as to what's now missing, which has something to do with the idea that, in having a shared intention, I not only think of my intention as settling matters but I think of your and my intentions as settling matters together, and that I do not settle matters concerning other facts about the world in the same way.

CHAPTER 4

Joint Settling in Theoretical Accounts of Shared Intention

There are many events whose occurrences we have no say over but which are nonetheless relevant to our planning and which we must therefore form expectations about. In BEACH, this may include the afternoon's weather forecast, if trains are running to schedule, the presence of lifeguards at the beach and the disposition of the ice-cream truck driver to set up shop nearby. There is a sense in which we take these features of the world at arm's-length when thinking about the future. Conversely, the success of certain events depends on our contribution (and so also their failure) or whose course we can influence by our actions and, crucially, which we *intend*. Whatever the weather, I intend to surf; whichever trains are running, I intend to get to the beach. My surfing and going to the beach I do not see as events taking place in the world around me. Rather, in intending them, I see matters as characteristically settled, something we saw in the last chapter as setting intention apart from my prediction of how I will act in the future. This distinction helps us begin to clarify what was missing at the end of the last chapter with the straightforward application of Bratman's Asymmetry Thesis (AT).

In applying the AT to address uncertainty both about partner intentions and about other, arms-length facts about the world, it looks like we are treating agents' attitudes towards both as mere predictions of what the agent's expect, or predict they will be. Using something like the AT to reconcile Mya's uncertainty with his intention that he and Iva go to the beach, suggests that we see him as merely treating her intention as a prediction of how she will act and intending their joint activity based only on this prediction. To be sure, this is not to say that the AT wouldn't generally apply when forming intentions in light of uncertainty about what others intend. As a basic principle of practical rationality, Mya shouldn't plan on Iva having an intention she's very unlikely to have. He shouldn't plan to meet her at the zoo when he knows she is mortally terrified of zebras. But Mya could well make plans based on only his expectation of Iva's intentions. Perhaps Mya overhears from a mutual friend that Iva,

whom he hasn't seen in ages, will be at the beach this afternoon and he hopes to surprise her there. Indeed, the AT seems like a reasonable norm to guide planning.

So what's wrong with seeing Mya as treating Iva's intention only like he would a prediction of how she will act? The issue comes in if we think that, when sharing intentions, those involved intend not only their own actions and the shared activity but also intend, in some sense, their partner's actions too. It's a problem, for example, if, when you and I share intentions that we *J*, I settle matters about what I will do *and* I settle matters about what *you* will do, and vice versa, when, crucially, my settling what you will do requires *more* than me merely predicting how you will act. In the upcoming Chapter 5, I will expand on exactly what more is required by focusing on how a conceptual difference between intending versus predicting a partner's intentions can help clarify what it means for intentions to be shared in any meaningful sense.

In the current chapter, however, I want to give expand on and analyse the concern we had that, even after dealing with the issues of the absence of common knowledge and weakened belief requirements on intention in the last two chapters, it looks like we still have a problem explaining how there can be shared intention in contexts of uncertainty when we reflect on the familiar role of intention settling matters, but this time in the context of shared rather than individual intentional activity. The purpose of this chapter is therefore, first, to formalise how intentions are thought to settle matters in the context of joint action—that is, agents *jointly* settling what they are or will be doing—and, second, to argue that uncertainty about partner intentions undermines certain background assumptions which are either explicitly or implicitly relied on to enable joint settling to take place. I aim to show that when there is motivational uncertainty, we don't yet have a good reason why parties can rely on one another in ways that allow them to settle what they will do. Furthermore, we will see that the issue is not limited to Bratman's AT but is, rather, reflective of a broader risk in trying to begin with insights from individual intentional action and adapting them to the joint case.

To explore the importance of settling in the context of shared intention, how it should be conceptualised and the challenges faced in doing so, I analyse two authors' accounts of shared intention: those given by Michael Bratman (1992, 1999b, 2014) and Johannes Roessler (2020). In each, I locate where settling enters and explain its framing in terms of joint rather than individual agency. What I'll show is that while the two accounts are quite different in their aetiology of shared intention, both authors explicitly regard settling as an

essential characteristic of shared intention and both identify the same problem faced; namely, to provide a theoretical explanation for how exactly individuals sharing intentions are in positions where they can *jointly* settle matters about what they will do.

Not only do Bratman and Roessler face a common challenge but, as I will argue, their solutions, though methodologically and conceptually quite different, rely on similar background assumptions about cooperativity and the ordinary predictability of typical occurrences of shared intentional activity. These assumptions are, I go on to suggest, rather taken for granted by both authors. Introducing reasons individuals have to be uncertain about partner intentions shows why; they are undermined when we squeeze these background assumptions. This means that if it's true that there can be shared intention in contexts which generate substantial uncertainty about intentions, we need to either do more investigation given the resources available or introduce additional assumptions to establish why individuals would still be in a position to jointly settle matters.

4.1 Michael Bratman's account of shared agency

Across the last three chapters I described Bratman's intentions-as-plans theory of individual agency and why it's necessary that intention settles matters about future activity. Intentional attitudes involve a commitment to acting as one has decided and settled on doing. I also discussed how his theory of shared agency springboards off this theory, in particular drawing on norms of practical and intention rationality to also explain the normativity present in shared activities too. Remember that Bratman is not simply drawing a conceptual parallel here. His view is of a strict continuity between the nature of planning in the case of solo activity and when two or more individuals act together. Accepting this implies that intention's essential functional role in practical agency must also be continuous.

The logical conclusion of this is that the settling requirement plus the continuity thesis gives us a joint settling requirement in joint action. Shared intention must settle matters about what the group will do. However, how this is supposed to take place is not straightforward to establish. Recall from the SI thesis in Chapter 2 that Bratman sees shared intention as a network of interdependent personal intentions in favour of the joint activity, each of which looks like this: "I intend that we *J*". Because he sees shared intention as reducible to component parts purely available from individual agency, Bratman sees nothing incorrect in stating a personal intention in this form. Specifically, there is neither a conceptual nor

metaphysical mistake in the subject of the intention being the individual while the content being the group activity, that is, that "we *J*". But this formulation seems problematic, for how can *I* intend *our* activity? In the last chapter, I described how intention is bound up with a sense of volition and control; we usually take it that what I intend are my own actions over which I have control. As Annette Baier puts it: "any intender assumes discretionary powers, powers to settle the moment-by-moment details of how the intention gets implemented, or how the intentional activity gets sustained" (Baier, 1997: pg. 26). But, she says,

> "if what each favours is joint intentional activity, then we still have unreduced we-intentions embedded in the interlocking individual intentions. It seems reasonable to maintain that I cannot intend what I believe to be beyond my power or control, so I conceptually cannot have the prior intention that we do anything at all, unless I have executive power to give the orders to the rest of 'us'" (Baier, 1997: pg. 25).

David Velleman, writing at the same time, raises a similar concern. His critique rests on intentions' planning function as a means of settling issues. In what he calls a settling condition, which should be part of a planning account of intention, he says "your intentions… are the attitudes that resolve deliberative questions, thereby settling issues that are up to you" (Velleman, 1997: pg. 32). They do so, he says, both factually, by causing the issue to turn out a certain way, and notionally, through representing the issue as turning out a certain way.

In *I Intend That We J* (1999b) Bratman provides a general response to both authors' concerns; to what he calls the control condition—that I may intend only those actions I think I control—and the settle condition—that I may intend only those actions I think my intending settles (both notionally and actually). This is his statement of the problem:

> "The problem arises in those standard cases in which "[w]hat we are going to do is supposed to be determined by you and me jointly". In such a case how can I intend that we *J*, consistent with the S condition? For me to intend that we *J* I must—according to the S condition—see my intention as settling whether we *J*. But that seems incompatible with seeing you as also intending that we *J* and so as also having an intention that settles whether we *J*" (pg. 149).

The problem with my intention in the form "I intend that we *J*" is that it therefore seems incompatible with you having the same intention, where both of us each settle our *J*-ing. How

can we each see *J* as up to ourselves to settle when *J* requires both of our contributions, and the contributions of the other are expressly not up to ourselves to control?

*

The gist of Bratman's response is that I *can* intend an action which is not strictly within my control or ability to settle, provided two things hold: first, I am in a position to see my control mediated by your intention; second, that such mediation is in response to you recognising my intention. He states that "plausible [control] or [settling] conditions should allow for such, as I will say, other-agent conditional mediation" (Bratman, 1999b: pg. 152).

He supports his view using a range of examples which progressively build on a version of Anscombe's story of Abe the pumper. In the original, Abe is a person who intends to pump water into the house and who intentionally moves a pump handle, thereby pumping water into the house, believing both "that his intention settles whether he will pump water into the house, and that he is in control of whether he will pump water into the house" (pg. 150). Over the course of a long and tiring day, Bratman has Abe interacting with a series of partners on whom he relies to raise the system's pressure (using a mechanism they control) so that he can succeed. Abe's partners differ in their dispositions towards him: Barbara's job is to keep the pressure high at all times; a machine partner raises the pressure when it detects Abe pumping; Bill monitors Abe and raises the the pressure when he sees Abe pumping; Charlie raises the pressure when he sees Abe only intending to pump, for example through communication from Abe; (Charlie and Bill do raise, but could also, with a logical shift, be shown to intend to do so); and, finally, Dianne, who, though she does not yet intend to raise the pressure, "is a kind soul and has access to the pressure valve" (pg. 154) which she turns when when she notices Abe intending to pump. Bratman (1999b) concludes that

> "[t]hese examples suggest that plausible [control] or [settle] conditions on intention should allow that control can be mediated by another agent and that this mediation can itself be conditional on that very intention. I may intend X while believing that my control over X would proceed by way of a process that involves other agents responding to my intention. I need only see my intention as settling whether or not X given what will happen, and what others will do, if I do so intend" (pg. 152).
>
> (…)

> "The answer is, first, that I can "frame" the intention that we J in part on the assumption that you will, as a result, come also so to intend. While I confidently predict you will come so to intend, I also recognize that you remain a free agent and this decision is really up to you, just as I can recognize that your decision to tell me the time, in response to my query, is up to you but fully predictable. Second, even after I have formed the intention that we J, in part because I predict you will concur, I can recognize that you still need to concur: It is just that I am fully confident that you will. Third, and finally, once we arrive at a structure of intentions … we can each see the matter as partly up to each of us" (pg. 157).

It's here as well that Bratman introduces the idea of *persistence interdependence* between the intentions of the agents, such that each will continue to intend if and only if the others do as well (all else equal). Persistence interdependence is required for me to see my intention as controlling your intentions by way of supporting the persistence of your intentions and, conversely, to see my own intention's control mediated by your support of mine. This allows each of us to experience the sense of control over our *J*-ing, as each of our intentions goes partly by way of the others', under conditions of common knowledge.

4.2 Bratman's view and issues with motivational uncertainty

The reference to common knowledge shows more clearly the argument I made in Chapter 2, that one purpose of a common knowledge condition in minimal accounts of shared intention is to explain why intentions settle matters in joint action. In Bratman's account, common knowledge ensures that agents' intentions are persistent interdependent. Of course, the introduction of motivational uncertainty rules out there being common knowledge. But this particular feature doesn't rule out Bratman's proposal, which is more subtle, as what enables Abe to settle matters about what him and his partner will do together is not that she does, in fact, recognise his intention and come intend the same as he does, but only that Abe can rely upon her to do so. To cash this out—that is, what it means for Abe to be in a position to 'confidently predict' what she'll come to intend—we can look at various descriptions of this Bratman (1999b, my emphasis added) provides:

> "Abe also *believes that Bill will turn his valve when the time comes*, for Bill monitors Abe's pumping and responds accordingly. Abe believes that if he intends to pump water into the house he will. But he knows that this is in part because of Bill's

actions… So long as Bill's contribution is *known by Abe to be reliable*, Abe can form an intention whose success requires Bill's contribution" (pg. 151).

(…)

Suppose that Diane does not yet intend to raise the pressure once Abe intends to pump. But Diane is a kind soul and has access to the pressure valve. Recognizing this, Abe might be *justifiably confident* that if Diane knew that Abe intended to pump water Diane would decide to turn the pressure valve. And *he might be confident that if he intended to pump Diane would know it*" (pg. 154).

(…)

"Suppose now that the issue of whether we paint together is one that is obviously salient to both of us. I know you are not yet settled on this course of action because you are not yet confident of my attitude. But *I know that you would settle* on this course of action if only you were confident about my appropriate attitude. I infer that if you knew that I intended that we paint, then you would intend that we paint, and we would then go on to paint together. *Given this prediction*, I form the intention that we paint and make it known to you; and then, *as I predicted and as a result,* you too form the intention that we paint" (pg. 155).

(…)

"Abe settles the matter of pumping the water into the house, even though he knows his success depends on Diane's recognition of his intention and her supporting action. Granted, *he settles this matter only given his predictions about Diane;* but *he is in a position reliably to make those predictions,* so he is in a position to settle the matter in a sense plausibly required for intention. But if Abe settles the matter in such a case, it seems to me that I can settle the matter of our painting *so long as I am in a position reliably to make the appropriate predictions*. I do not settle the matter in a sense that precludes that the route to success involves further voluntary activity on the part of another agent" (pg. 156).

Despite all of this, Bratman doesn't spell out how agents come to be in the positions he describes or what justifies us in saying they are! He doesn't, in other words, give us much insight into what grounds the partner-reliability condition that supports his proposal for

control-mediated intention and joint settling. What justifies Abe's expectation that his partner's will come to intend as he does (i.e., in favour of the joint activity)? Diane being a "kind soul" is hardly a universal feature of social interaction. More generally, Bratman's approach seems to be that we can safely assume that expectations like Abe's are not unusual but involve, rather, "just the ordinary predictability of ordinary agents" (Bratman, 1999b: pg. 155) on whom we can rely to intend that we *J* when our intentions in favour of *J* are made manifest. Furthermore, the process for establishing these expectations is something Bratman likewise takes to be non-controversial. He asks, for instance,

> "[h]ow might this happen? Well, I might just report: "I intend that we paint on the assumption that you will thereby be led as well to an intention that we paint." Or I might simply start painting, given that I expect that you will see this [or it will become salient to both of us] and thereby, knowing me fairly well, recognize my intention that we paint and so arrive as well at such an intention and just jump in" (Bratman, 1999b: pg. 155).

Various forms of explicit and implicit communication therefore provide a basis for partner reliance. Note also the reference to a background of familiarity which provides powerful cues as to what each of us is likely to do next, perhaps because we've done this many times before. Presumably, Bratman feels a complete account of how agents come to be in a position of confidence about what their partners will intend is outside his scope—maybe there are too many ways to be able to provide a general account. We might also say that Abe can reliably predict how his ostensive interaction partners will come to intend because he has no reason to doubt otherwise. And because of this, he can settle matters about what they will do[13].

As I raised at the outset, however, it's plausible that there are cases in which you and I might be in the process of or anticipate doing something together where I'm aware of tempting alternatives you have available and so have good reason to be fairly uncertain about your intentions. In such cases, there are good reasons why I may not be in a position to predict—with the kind of confidence that Bratman's proposal seems to require—that you will come to intend like I do. It's plausible to think, therefore, that the partner-reliability condition

[13] It's important to reiterate that it's not just an agent's ordinary predictability about her partner's intentions that needs justification, but ordinary predictability about what her partner will intend *specifically in response to* her own intentions. It's not just Diane's predictability per se, as if she was a fisherman who sailed out every morning and whom Abe watched each day from his balcony. It's predictability that's partly explained by Abe's intentions. This hints at a possible solution to the problem we face, in that if we can find some essential feature of their social interaction to boost predictability through responsiveness, then we have a solution within the conditions of the current account.

will not be met in these circumstances. If I have reasons to doubt that you will come to or continue to intend the joint activity, then it could very well be that I am not 'in a position reliably to make the appropriate predictions' about you coming to intend the joint activity and may not be in a position where your 'contribution is known by [me] to be reliable'. And if the partner-reliability condition is not met, then the conditions for joint settling—as per Bratman's response to the Velleman-Baier challenge—are not met either.

Yet, as observed in Chapter 1, in BEACH many of the conditions generally required for shared intention are met. So if it's plausible that there is shared intention in this case, we have yet to figure out why Mya can settle matters about what they will do. It's true that he may well see his control over the outcome mediated by Iva's intention, despite having reasons to be uncertain what she intends. He may initially be unsure whether Iva will meet him at the beach, but he might nonetheless decide to rely on her to join up with him, make plans and settle on their playing beach tennis together, catch the bus and arrive at the beach to find Iva who, it turns out, had always intended to meet up with him. The problem, though, is that it seems too big of a leap, in cases like this, to rely on the 'ordinary predictability' of others to confidently predict and rely on how they will act. The Temptation-Evidence principle (from Chapter 1) says any temptation worth its name must make Mya uncertain and create a problem for her to resolve. It is, simply, not just a matter of ordinary predictability.

On our reading of Bratman's account thus far, we don't yet have any additional reasons to think that Mya should more likely expect Iva to join him than to skip the beach for the football match. To be fair, the problem we're dealing with is not a focus of Bratman's. Still, as we'll see in the next chapter, digging deeper into his account does seem to offer a possible solution to the problem of settling matters when there's substantial uncertainty, though it takes some work to get there and the results may not help us as much as we'd like.

*

One final point, before moving on. To avoid the problem motivational uncertainty poses, we could take Bratman at face value when he claims to limit his account to cases of small-scale, modestly-social shared activities, in which the kind of shared intention he proposes is merely one of several different species, and not meant to cover situations involving this kind of uncertainty. We are, we would be saying, excluding situations where agents have any reasons to doubt their partner's intentions.

There are good reasons to doubt this strategy. First, as I've said thus far, we might think that situations involving uncertainty about another's intentions are commonplace and that we face them regularly in our everyday lives. These should therefore fall within the scope of any adequate account of shared intention which aims to be general, as Bratman claims his account to be. Second, and following, many of the examples Bratman uses to describe instances of shared intention can be tweaked in small ways to introduce the kind of uncertainty I have in mind here. This means we are not necessarily talking about cases of shared intention 'at the margin' but, rather, that again it's possible these are a regular feature of social interaction. Third, Bratman hopes that his account of shared intention in the case of small-scale shared activity can provide insight into a theory of larger-scale social institutions. One very likely consequence of scaling up is the impact on what agents know about their partners—including the extent of their knowledge about what each of their (potentially many) interaction partners intend. For Bratman to succeed in this aim, it would seem odd that such an account is, at base, completely incompatible with the presence of at least some form of uncertainty about others' attitudes. Finally, it's core to Bratman's view that shared intention provides the necessary supporting infrastructure for agents to negotiate and bargain over time about what each will do as part of their joint endeavour. These bargaining processes are an identifying feature of the essentially *social* nature of the interaction. But such bargaining processes must take seriously the fact that each agent cannot know their partners' minds in the same way they know their own. True bargaining implies the presence of private knowledge such that we can't treat multiple agents as a single unitary agent making decisions on how to act—*at least without making some additional assumptions* about how each member perceives and participates in the joint activity. Together, these suggest that a proper, general account of shared intention needs to explain, in part, how agents deal with the fact that sometimes they have reason to be uncertain what their partners intend. The problem is that in Bratman's account we don't yet have an answer as to how agents settle matters despite having reasons to be uncertain about their partner's participation in their joint activity.

4.3 Johannes Roessler's relational account of shared intention

Are these problems of explaining joint settling under uncertainty about intentions limited to Bratman's account? Though not an exhaustive approach, one idea is to take an account very different to his and assess whether the problem persists. Johannes Roessler's recent work on plural practical knowledge (2020) provides an opportunity for this. His view is broadly

‘anti-reductive’, seeing shared intention as irreducible to individual intentionality alone: we intend as a group (“We intend to J”), rather than you and I each intending that the group does something (“I intend that we J”). At the same time, Roessler avoids explaining shared intention by appealing to a kind of supra-personal or plural agent, thus providing a good counterpoint to Bratman’s methodological individualism, while avoiding a notable challenge facing many anti-individualist accounts.

Roessler’s proposal is also attractive as it draws on Elizabeth Anscombe’s treatment of intention, which is markedly different from Bratman’s in several ways. Most notably, Bratman sees intention as a distinct type of attitude, separate from, for example, the attitudes of belief and desire. Anscombe does not. Bratman also views intention as having a functional role as an input to and constraint on practical reasoning about further intentions and actions. Anscombe, conversely, views intention as something known without inference, not represented in deliberation as such. Finally, Bratman’s main focus is on future-directed intention—current intention for future action—while Anscombe’s is an analysis of present intentional action. Bratman and Anscombe may, of course, be using the same term but talking about different things (this is the ‘unity of intention’ problem alluded to in Chapter 1), or they may not be mutually exclusive—Bratman argues that his account can accommodate present-directed intention too. Either way, while it’s hard to establish criteria to directly compare their accounts, they are often taken as providing quite different, though both foundational, approaches to the subject of intention and intentional action. Exploring an account of shared intention that uses Anscombe’s work as a springboard could therefore provide useful insights for my project.

*

Certain authors have asked if it’s possible to extend Anscombe’s treatment of individual intentional action to the collective case. Perhaps the most important contribution she makes is her description of non-observational knowledge—and specifically a kind of practical knowledge associated with intention:

> “Anscombe’s account of practical knowledge develops from observations about the way we ordinarily make, and engage with, claims to knowledge as to what we are or will be doing. One observation is that such claims often simultaneously purport to express knowledge and intentions. Another is that there is a distinctive pattern of appropriate (and inappropriate) responses” (Roessler, 2020: pg. 2).

To illustrate practical knowledge—or knowledge in intention, or knowledge without observation, terms often used interchangeably—consider the example of Mory, whose car has broken down on a winding country lane one late afternoon and who has opened the car bonnet to inspect the car's engine. Zoe, a passer-by who happens to be out for an evening stroll, notices the bonnet squeaking as Mory moves it up and down, in his attempt to lock it open, and stops to chat to find out what is going on. Anscombe's idea is that intentional actions are characterised as those "to which a certain sense of the question "Why?" is given application", the sense in which "the answer, if positive, gives a reason for acting" (Anscombe, 1963: pg. 9). Zoe asks: "Why are you lifting the bonnet?", to which Mory responds: "To inspect the engine". However, Zoe notes that Mory's actions also lead to that rhythmic squeaking noise. Not only that, but his lifting the bonnet is also casting a shadow onto the road (in which Mory's small dog has taken comfort). These are two unintentional actions, as evidenced by the reason Mory gave in response to Zoe's question. To see this, Zoe might have asked: "Why are you lifting the bonnet? Is it to…:

- … make a rhythmic squeaking noise?"
- … cast a shadow for your dog to lie in?"
- … inspect the engine?"

Answering 'yes' only to the third shows which of Mory's actions is intentional. Given the variety of things he might be doing, the clearest indication of which is intentional is given by the positive, reason-giving answer to the open-ended question "Why?".

Zoe's three questions illuminate another important feature of Anscombe's account, which is that only Mory can have the kind of basic knowledge of his intention that Anscombe calls *practical knowledge*. She contrasts this with *speculative knowledge*, which is arrived at differently through evidence and inference, like observation, inductive reasoning, or testimony. Practical knowledge is characteristically *not* explained by citing evidence for what we are doing. An intentional agent knows her intention without needing to refer to any way of finding out what she is doing, as she would if she was a third-party observer of the action. Zoe can learn Mory's intention by asking him, but she cannot know, without posing the question, which of her three questions would receive a positive response. Mory, on the other hand, 'simply knows' what he intends: while opening the bonnet intentionally, he knows that

he is opening the bonnet. Knowledge in intention is thus characteristically non-observational, unlike knowledge one might have about another person's intentions.

It is, furthermore, a hallmark of what Anscombe calls *an expression of intention* that it would strike us as out of place were Zoe to request an explanation from Mory of how he knows what he intends by lifting the car bonnet: "How do you know what it is you intend?" or "How do you know that you intend to inspect the engine?" would be odd questions to ask[14]. Conversely, it would not sound inappropriate to ask Zoe how she knows what Mory intends. This would, instead, seem to invite a valid response, for example, "He told me that he is inspecting the engine". Accepting this looks like it also means acknowledging that Mory is *entitled* to answer the factual question of what he is or will be doing just by expressing his intention. In fact, he treats it, rather, as a practical question to which he can make a claim to knowledge just by expressing a practical choice, an action: "I am inspecting the engine"[15]. This leads Anscombe to say that Mory's expression of intention, then, is simultaneously an *expression of knowledge* of what his intention is. And only Mory is thus entitled to express knowledge in this way (i.e., without providing grounds for how he comes to know it).

*

Consider a new scenario where this time Zoe comes across two people, Mory and Rory, positioned on either side of a car and pushing it along the road. Zoe asks Rory, who happens to be pushing the side closest to her: "Why are you pushing the car?" Rory responds: "To get it to the garage to be fixed." Mory and Rory's pushing the car seems an uncontroversial example of a shared intentional activity. The focus earlier in this chapter was on Bratman's explanation of shared intention as a 'bottom-up' construction of the individual psychological attitudes of those involved, whose interconnection supports the goal of getting the car to the garage. The 'top-down' view we have now, takes it that if there is a collective activity which is genuinely intentional, then it cannot be reduced to a combination of individual attitudes or activities, no matter how complex.

[14] "Odd, or "'off-key' means more than 'brusque' or 'tactless' or 'conversationally inappropriate'. 'How do you know you and these other people are pushing the car to the gas station?' may be any of the latter, yet…this might be glossed as: 'leaving us at a loss as to what would count as a good answer', or 'erroneously presupposing that the addressee knows about the fact in question by exploiting some way of finding out'" (Roessler, 2020: pg. 10).

[15] Of course, there are a range of cascading reasons for this, which at some point must stop. It's explained by the next practical reason along: e.g., I am lifting the bonnet to inspect the engine (the intentional action is the lifting of the bonnet); not I am lifting the bonnet to lift the bonnet; or I am moving my fingers this way to lift the bonnet.

In this direction, several authors have taken Anscombe's ideas above and asked whether participants in a shared intentional activity might have the same kind of knowledge of what they're doing jointly as when acting alone intentionally. That is, is there a first-person plural corollary to first-person singular practical knowledge? Their idea is that Anscombe's methodology—of relevant questions and responses, as just narrated—can explain how a collective action can be intentional partly by appealing to how "participants in a collective activity can have the sort of knowledge possession of which, as Anscombe taught us, is part of what it means to be acting intentionally, viz. 'practical knowledge' of what they are doing" (Roessler, 2020: pg. 5). Note that this plural practical knowledge approach to characterising shared intention doesn't attempt to characterise shared intention directly, but says only that individuals who can be said to share intentions have a particular sort of knowledge. On this view, then, if Mory and Rory have a plural form of the distinctive practical knowledge Anscombe saw as characteristic of intentional action, then they have practical knowledge of their car pushing, and so must share an intention to do so.

To see whether two or more agents have plural practical knowledge, Roessler (2020), building on previous work (see Stoutland, 2008; Laurence, 2011), explores whether Anscombe's observations about ordinary claims to knowledge of what we are doing have "plausible analogues in the case of collective activities" (Roessler, 2020: pg. 2). For example, whether Zoe's question to Rory elicits the same kind of responses we saw in the case earlier when Mory was pushing alone. Zoe asks, "Why are you pushing the car?", and Rory responds, "(We're pushing the car) to get it to the garage."

Why should we understand this as Rory expressing their (his and Mory's) shared intention—and so a valid example of plural practical knowledge? First, the question is most naturally heard as being addressed to the collective, where the subject of the question is second-person (plural), and not asking what either Mory or Rory are doing pushing the car on their own. So it's naturally heard not as a request for either of their reasons for acting (or for both of their individual reasons for acting), but the reason they are pushing together. Second, Zoe's question appears to meet Anscombe's special sense of the question 'Why?' (…are you pushing the car). For this type of question, Rory's positive reason is appropriate while the response "To get some exercise" is not. Though both of them may be getting some exercise, it is not the reason for which they are acting together. Finally, Rory's response is naturally interpreted by Zoe as expressing knowledge of what they intend. Recall the mark of this is that, in response, a request for a reason for their actions ("Why are you pushing it to the

garage?") would be appropriate and invite a response ("To get it fixed"), but a request from her for evidence of how they know what they intend would not. "How did you find out / How do you know that you intend to get it to the garage to get fixed" would seem odd. Overall, if these three points are correct, we could say that Mory and Rory have non-observational knowledge, the question has application and they are entitled to express their intention.

*

But there's a clear problem with this conclusion. The question "How do you know what it is you intend?" does not actually seem strange. It would not in the slightest be inappropriate for Zoe to ask Rory how he knows what they intend. Hans Bernhard Schmid notes this issue in his work on collective intentionality:

> "If you tell me what you intend to do, individually, it does not make much sense for me to ask how do you know what it is you intend. You just know—that's it. But if you tell me, that is what you intend to do together with your partner, no such reply seems to be possible. You don't 'just know'. You'll have to quote some evidence, and you are likely to reply with: 'That's what we have agreed to do, and here we are', or some such. In that sense, too, joint intentional activity seems to be deeply different from individual intentional activity. Whatever knowledge of what it is we are doing together cannot be basic, but implies individual self-knowledge and observation. Neither of us has immediate awareness or introspective access to our intention to go for a walk together, but only to whatever individual contributive intentionality we have, individually" (Schmid, 2016: pg. 12).

That it's not strange for Zoe to ask Rory how he knows what they intend can also be seen in what she expects as a response. Rory can't get away with saying "I just know" while still respecting that the question is appropriate. He could describe the conversation that led to them agreeing to push the car to the garage. This needn't even be explicit: Mory's sigh, a quick nod to Rory ("How many times have we been here before?"), and both climb out to push. Whatever the case, a response would rely on evidence that would seem to render Rory's knowledge of their shared action speculative—no different to what a third-party observer to their discussion might have.

Anscombe's thesis was that an intentional agent has an entitlement to knowledge of what she intends that others do not—and is entitled to express this knowledge by expressing

her intention, rather than providing proof. This unique authority appears incompatible with the idea that multiple agents can share the same intention. Rory seems to have speculative and not practical knowledge of their joint activity and so isn't in a position that grants him the authority to settle matters about what they, Mory and Rory, intend to do. This is the *issue of authority* to which Roessler refers, saying that because we are

> "committed to separating two roles that, in Ancombe's discussion, are invariably co-occupied … how should we understand the authority of the addressee's account of the reasons for which the collective activity is being undertaken?" (Roessler, 2020: pg. 7).

As Schmid puts it: "What I take 'our intention' to be does not settle the question of what it is we're doing together in the same way it does in the case of my own intentions" (Schmid, 2016: pg. 61).

It's worth revisiting why Anscombe requires that an intentional actor has practical knowledge. Not simply for the sake of it, but, rather, because this type of knowledge is tied up in a very distinct aspect of agency: namely, it is up to the intentional agent to settle matters about what he is or will be doing. He settles matters not through prediction or speculation about this, but, instead, by deciding what to do—and in deciding, expressing an intention to do so (which need not be explicit). Agents with only speculative knowledge are not in a position to decide to get something done and settle matters in this way. In Roessler's words, what's at stake, then, is that

> "so long as our questions as to what the group are doing, and why they are doing it, are addressed to an individual participant in the collective activity, the authority of our interlocutor's response will be theoretical, not practical. He cannot settle the question of what the group are doing by deciding what to do, or by expressing his individual intention. That is why his answer, if knowledgeable, will be an example of speculative, not practical knowledge.
>
> (…)
>
> If he is expressing the group's intention and knowledge, the question of how to understand an individual's knowledge of what the group are doing remains wide open" (Roessler, 2020: pg. 10).

The authority relation is delicate. Agreements do often, if implicitly, grant their parties an authority to speak on behalf of their group—and Mory is unlikely to object to Rory doing so. But it's arguably a similar authority an eyewitness to their discussion might be granted. An exhausted Mory, out of breath and unable to speak, might wave towards Etel, a construction worker digging at the roadside near the breakdown, to explain on their behalf. Authority, then, is the authority to settle matters, an authority granted only to those with practical knowledge of their intention, something Etel, whose speculative knowledge came from overhearing Rory and Mory's initial conversation, does not have. It seems purely up to Mory and Rory whether they will push the car to the garage; whereas Etel can only make a prediction about this. In other words, practical knowledge involves the kind of settling I've discussed in this thesis so far (deciding what I will do), while speculative knowledge does not (predicting what you will do).

*

To make plural practical agency work, the challenge is therefore to characterise knowledge of a joint action in a way that overcomes the problem that settling matters on behalf of the group is not possible given only speculative knowledge. Roessler does this by (1) accepting that intentional action involves practical knowledge and (2) accepting that practical knowledge is characteristic of the authority to settle matters, but (3) rejecting the pre-supposition that an agent's knowledge of their group's activity (including their partner's participation in it) is necessarily speculative. He starts by introducing a different expression of shared intention (that is, an appropriate response to the question "Why?" aimed at the group), one which, he says, better reflects the interpersonal nature of joint activity. According to his relational view of plural practical knowledge,

> "if we have plural practical knowledge of being engaged in some activity, then we both must have practical knowledge of acting with each other; more precisely: practical knowledge we could articulate by the use of the first- and second-person pronoun, 'I'm doing x with you" (Roessler, 2020: pg. 2).

To make this clearer using our current examples, consider two different replies—expressions of shared intention—Rory might make to Zoe:

> Ex (1): I am pushing the car to the garage with Mory.
>
> Ex (2): We are pushing the car to the garage.

While both responses would not be out of place, Roessler's proposal is that the relational form Ex (1) is the correct formulation of an expression of shared intention, one indicative of plural practical knowledge. Even if Rory chooses to say Ex (2), Ex (1) is the right way to understand his expression of shared intention.

I'll take at face value Roessler's claim as to why Ex (1) should be regarded as a valid expression of intention, and focus instead on the issue of authority that seems to remain. It's still the case that, in response to Ex (1), Zoe might reasonably (though rudely) ask Rory: "How do you know Mory wants to do this?" She's asking for an account of his grounds for this: if he didn't know Mory wanted this, Rory wouldn't be entitled to make such a claim—he doesn't have the authority, in Anscombe's terms, to make it. But again, such a claim to knowledge seems obviously speculative. So, getting the car to the garage is not something Rory on his own is in a position to settle, at least not in the way he settles his own actions.

To progress, Ex (1) must therefore be reconciled with the kind of knowledge that allows Rory to settle matters on both of their behalf. Roessler's solution is to reject the assumption that the only kind of knowledge one person can have of another, including their attitudes, is speculative. He argues, instead, for a particular kind of knowledge agents in a joint activity have—namely, knowledge each has of jointly intentionally acting with others—which is basic and non-inferential, and thus to be understood as practical rather than speculative knowledge. Moreover, he says, this requires no new and unique processes to obtain. Simply by communicating together, agents can gain the basic knowledge required to jointly settle matters. He explains:

> "There is a sense in which the relational form of plural practical knowledge is basic. More specifically, what is basic is a second-person version of the relational form. Shared practical knowledge that 'we' (you and I) are doing x depends on our being in communication with each other, enabling us to articulate our practical knowledge by saying 'I'm doing x with you.' … Of course, there will often be no point in making our activity explicit in this way. What matters is that we are both in a position to do so insofar as we are communicating with, and able to address, each other. It is this that makes it possible for us to 'settle' *together* what to do and to acquire an intention that is 'the object of shared recognition'" (Roessler, 2020: pg. 14, author's emphasis)

It's still true that Rory can't on his own settle whether he and Mory push the car to the garage—and so he cannot have practical knowledge, by himself, that they are. But this is not

what Rory is claiming by uttering Ex (1) "I am pushing the car to the garage with Mory". Rather, Rory is claiming to be engaged in a jointly intentional activity *with* Mory. And the turn Roessler makes is in suggesting that Rory's knowledge of *this*—that is, Rory's knowledge that he and Mory are together engaged in a joint intentional activity—is practical in the sense required (i.e., basic, non-inferential, non-observational). Rory's practical knowledge of his personal intentions thus settles matters for himself and his practical knowledge of his joint interaction with Mory allows him to settle matters about what Mory is or will be doing, and settles that they are acting together.

4.4 Roessler's view and issues with motivational uncertainty

In the most recent quote above, the language of 'being in a position' to settle matters sounds familiar. But what does this mean and why is Rory entitled to his knowledge claim? As Roessler puts it, what is the "salient prerequisite of [his] entitlement to that claim, viz. whether and how [Rory] knows [Mory] is cooperating[?]" (Roessler, 2020: pg. 14). The term cooperation, Roessler continues, is used in the sense of Mory having the right attitudes to make Ex (1) true. And his answer is that, provided agents are able to communicate, to address each other, then they are in a position to have this knowledge. This prerequisite (i.e., that Rory knows Mory is acting cooperatively with him) is therefore minimally satisfied only if there is communication among the participants (Roessler, 2020: pg. 16). Communication allows agents to articulate their practical knowledge (of each of their engagement in the joint intentional activity) to one another, and this provides a basis for their entitlement to the knowledge that their partner is truly 'cooperating' with them (i.e., truly is engaging in a joint activity with them), and so jointly settle matters about what they will do.

The characterisation of the knowledge agents have of their acting with one another as basic has the advantage of overcoming a common challenge to accounts of shared intention, which is that the explanation of joint settling is circular, given the process of joint settling itself seems to presuppose some shared intention to do so. As Roessler notes,

> "jointly *settling* what to do does not have to take the form of a joint activity that's intentional under a 'we are settling what to do. A sensibly pluralist account will recognize the enormous variety of ways in which people start shared enterprises.
>
> (…)

> Again, ‘a gesture may suffice’ to attract someone’s attention and thus to start a conversation. The sense in which agents jointly settle what they are doing, in such a case, may come to something like this: both agents perform their respective parts in a shared activity of which they are mutually aware” (Roessler, 2020: pg. 17, author’s emphasis).

Roessler is thus using the term communication quite broadly, meant to encompass the myriad ways in which an agreement to act jointly might be reached. It seems he regards this as relatively non-controversial, in that he doesn’t so much emphasise how a decision to act together is reached, only that at some point it was. All that matters is that, at a minimum, communication reflects an interpersonal process which instantiates a mutual basic awareness of being jointly engaged.

This approach to grounding shared intention on basic psychological processes has other recent advocates. Hans Bernhard Schmid (2016, 2018), though not advocating for a practical knowledge view of shared intention per se, proposes something similar. Like Roessler, Schmid’s account avoids a role for any sort of plural agent as the subject of the shared intention. Unlike Roessler, though, Schmid (2016) argues the correct formulation of an expression of shared intention is given by the first person plural form—as in Ex (2) (“We are pushing the car to the garage”). Schmid’s major challenge is therefore to explain how this “we” is constituted without running in a circle, like via a prior shared intention to constitute the collective. Like Roessler, Schmid’s solution is to argue that there’s no issue of circularity if we see the ‘upstream’ process establishing shared intention as involving the transfer of a basic kind of knowledge. His proposal is that participants in a shared intention gain a first personal “plural awareness” of those with whom they are jointly engaged. This awareness is plural both in the sense of an agent in a joint intentional action being aware she is acting together with others and in the sense that it is each agent which has this awareness. Crucially, this plural awareness is said to be basic in that it is not to be explained by a shared intention to generate it.

*

The idea that shared intention is grounded on minimal psychological processes, which track an agent and her partners’ joint participation in a collective activity, is appealing. It both explains how shared activity ‘gets off the ground’ and provides a route for joint settling that isn’t circular—there is no recursive requirement to explain the origins of shared intention via

another shared intention, and so on. Recall that it's minimal communication with Mory, accompanied by a basic form of relational awareness with him, which gives Rory the entitlement to his claim to know that they are engaged in a joint activity.

More than that, though, this entitles Rory to rely on Mory's being cooperative with him, an important additional factor. More generally, it's important for Roessler's account that all agents are entitled to rely on their partner's cooperativeness—to rely on them having the right sort of attitudes which make Ex (1) true. His view seems to be that this entitlement is non-controversial, as is evident in these passages:

> "The mutual dependence between agents who share an intention, then, has both a practical and a psychological dimension. We depend on others' cooperation for getting things done, in cases where we are unable to do so by ourselves. But we may also depend on others' cooperation for engaging in the activities we think we are engaging in, and even for having attitudes of the sort we take ourselves to be expressing when we make claims such as [Ex (1)].
>
> (...)
>
> Second-person thinking, it has been argued … essentially depends on the addressee's disposition to recognize being addressed; in Moran's words, it involves 'a content and act that is the object of shared recognition'. On such a view, only if you have the required dispositions will I be able to have a second-person intention to tell you that p" (Roessler, 2020: pgs. 13-14).

The proposal is that in the kinds of joint actions that we tend to analyse, agents are rationally permitted to assume that their partners are disposed to be cooperative, and that shared intention is the result of this mutual recognition. This is to say that Rory can claim to know that Mory is interacting with him in part because of an entitlement he has, in light of a general pattern of cooperation, to depend or rely on those with whom he is interacting to have the cooperative attitudes he expects them to have (this is further echoed in Roessler's extract, not shown, from Dorothy Frede's ideas about how joint activities begin). This rationale is closely related to Bratman's grounding on the 'ordinary predictability of agents' from before. Rory can rely on Mory to be cooperative because such dispositions (and expectations) are such an everyday feature of our social interaction as to be rendered basic.

*

But the step from Rory and Mory communicating to Rory having the authority to express to Zoe a claim about what they are doing together is perhaps not something to take for granted. This additional step might sometimes require additional clarification and support. For instance, how does Rory know Mory is disposed to be cooperative? Mory told him this, he might say. But simply assuming that agents intend to communicate cooperatively—for example, according to Grice's (1975, 1989) cooperative principle, or Brown and Levinsohn's (1987) politeness theory—only pushes the question back further. What grounds do we have for assuming that these communicators are being truthful or communicating with cooperative intent? Simply stipulating mere communication might sometimes not be enough to establish Rory's entitlement to know that he and Mory are jointly engaged. This is just to say that the assumption that we adopt a cooperative stance in our communication is important: each of us must adopt a minimally cooperative stance towards our partner (in the sense of each of us having the attitudes our partner expects us to have) and expect that our partner does likewise. Otherwise we can't necessarily say that agents can rationally rely on their partners having the right sort of attitudes which make Ex (1) true. In other words, the assumption of a minimal cooperative stance can sometimes be doing much of the work to overcome the issue of authority, to enable individuals to 'know without observation' that they are in a joint action with their partners, and so jointly settle matters.

As with Bratman's account earlier, introducing the presence of non-aligned interests into the picture, which generates substantial uncertainty about partner intentions, violates this assumption of cooperativity. In these contexts, we don't now have a reason why an agent is entitled to know and rely on her partner's cooperation. More specifically, there's no guarantee that communication per se is sufficient for her to claim to know what her partner's attitudes are—and so the conditions for non-observational knowledge are not met. If I have reasons to doubt that you have "attitudes of the sort we take ourselves to be expressing when we make claims such as [Ex (1)]", then I'm not in a position to depend on your disposition to be cooperative. The question arising, then, is why in these cases we can assume the basic knowledge acquisition is guaranteed through communication.

This is perhaps beyond the scope of Roessler's project, if examples like BEACH don't strike him as possibly involving shared intention, or do, but involve a different kind of basic knowledge of joint participation. However, while we want to be careful not to simply trade intuitions, it's good to push this analysis as far as we can, especially given the overarching goal of searching for a minimal, generalisable account of shared intention.

Moreover, having reasons to doubt a speaker's veracity is not an unfamiliar concern in collective activity, and why individuals might have a reason to mask or shade their intentions in their interactions with others is worth exploring. For example, people sometimes insinuate their intent and don't speak plainly—"would you like to come up and see my etchings" and other sexual come-ons, veiled threats, polite requests and concealed bribes. This makes communication potentially inefficient, prone to misunderstanding and seemingly unnecessary (as people generally understand what's being meant, and it's within the grasp of the speaker to be clearer), which poses a challenge to theories of language like Grice's cooperative speaker mentioned earlier. To explain this, Pinker et al. (2008, 2010) develop a theory of the strategic speaker, borrowing ideas from signalling in evolutionary biology and evolutionary game theory to highlight potential advantages of indirect speech, under their assumption that most social relationships involve a mixture of cooperation and conflict. The authors argue that a strategic speaker might seek *plausible deniability* when she is uncertain whether the hearer is cooperative or antagonistic (e.g., paradigm case of bribing a policeman who has a chance of being corrupt). Indirect requests allow for this, as cooperative listeners can accept the request while uncooperative listeners cannot react adversely. Furthermore, they propose that communication serves the dual purpose of conveying information and negotiating the type of relationship holding between the speaker and hearer (one of dominance, communality or reciprocity), the emotional costs of a mismatch in the latter helping select for indirectness.

Another view that focuses on scepticism about others' testimony takes as a starting point that because we depend on communication with others we are open to exploitation—for example, being accidentally or intentionally misinformed (for an overview see Michaelson, 2018). This provides us with a reason to doubt an interaction partners' truthfulness, which has led over time to humans developing a "suite of cognitive mechanisms for *epistemic vigilance*" (Sperber et al., 2010), various mechanisms that track the quality of testimony and so discern between truth-tellers, the uninformed and liars.

In short, if it's reasonable to accept both that many of our interactions involve a mixture of motivations and that we depend on communication to do things with others, then there are good reasons why we cannot always take for granted that others are predisposed to be cooperative in what they tell and do with us.

*

Mechanisms of plausible deniability and epistemic vigilance are two factors that suggest that more needs to be said in general about the minimal cooperative stance assumed earlier. In cases like BEACH, this is even more pressing: it's clear that there are good reasons for Mya to doubt that Iva will join him at the beach. Furthermore, it's not enough for Mya to assume that simply picking up the phone will sort the matter out between them; Iva may ignore or miss the call, finding it too emotionally costly to tell the truth. Consequently, we don't yet have an answer in Roessler's account for how there can be shared intention in contexts with motivational uncertainty. It's actually surprising that we don't have much in the way of why we *should* depend on others given how he says we 'depend on others' cooperation for getting things done' and 'depend on others' cooperation for engaging in the activities we think we are engaging in'. This language suggests it is dependence and interpersonal reliance that really matter. Communication establishes the experience of 'with' but its normative force arguably comes from an appreciation of one's partner as a fellow cooperator; it's not a normative force intrinsic to communication itself.

That this dependence is underexplained is more strange when we reflect on a key driver of Roessler's view; namely, to distinguish between 'with' versus 'we'. The nature of the second-personal relational formulation of an expression of shared intention is core to Roessler's view of plural practical knowledge—and to his account of shared intention more broadly. A primary motivation for accepting this formulation is his hesitancy to accept a straightforward mapping of Anscombe's treatment of intentional action to cases of collective activity, as authors before him had done. Attempting to find strict parallels, he says,

> "fails to give sufficient weight to the fact that collective intentional activities are not merely a matter of groups exercising their supra-personal powers of agency: it also, essentially, involves *interpersonal* agency among the individuals making up the group" (Roessler, 2020: pg. 11, author's emphasis).

Certain issues arising in the search for a plural form of Anscombe's practical knowledge are, he thinks, a direct result of this failure. In particular, previous concerns about how joint settling takes place are unanswered precisely because too little attention has been paid to the nature of the knowledge relation that is characteristic of the interpersonal relation when acting with others. It is precisely because previous authors, who come close to invoking forms of plural agency, have not analysed how and what jointly interacting agents know about each other that they leave open questions about settling matters on behalf of a group.

Taking this view seriously must surely mean also taking seriously the possibility that others may *not* have the attitudes we think they have. Or at least that this requires some work to establish. Another way of putting this is that we perhaps lose sight of what it means for others to be cooperative if we don't think that they can be otherwise. If there is no consideration that one's partner won't be cooperative, then we might lose part of what Roessler is aiming at with his second-personal relational formulation. These observations are perhaps not a fatal flaw for Roessler's account, but they do suggest that more needs to be said about which factors support background assumptions of cooperativity in his account, and why these factors should also form part of the basic knowledge agents have of acting with each other.

4.5 Conclusion

Chapters 2 and 3 identified two issues—an absence of common knowledge, and a violation of a strict belief requirement on intention—that arise when trying to use traditional accounts to explain how there can be shared intention when agents have substantial uncertainty about what a partner intends. Though I argued that plausible solutions are available, in working towards a resolution in those chapters we realised that there was something separate not yet resolved; namely, that the important way intention settles matters for the intender still seemed absent. The aim of this chapter was to clarify what exactly is missing. First, by exploring how different authors—Michael Bratman and Johannes Roessler—have suggested that this characteristic settling might look in the case of shared rather than individual intentional activity. And second, by arguing that the background assumptions both authors make to justify agents being in positions to jointly settle matters—assumptions which are surprisingly similar despite very different methodological approaches—are plausibly not met in cases where there's motivational uncertainty. This suggests our intuition, at the end of Chapter 3, when relying on Bratman's AT to solve the problem of motivational uncertainty was correct; concerns about joint settling give us an independent reason for potentially excluding cases with motivational uncertainty from the set of those involving shared intention.

Despite this conclusion, in getting here we have learned something valuable about how we might plausibly resolve the problem of joint settling in contexts with motivational uncertainty. If we can't simply rely on background assumptions of ordinary attitudes of cooperativity and predictability, it's still possible there is some other reason why agents might justifiably be in a position to rely on their partners to act cooperatively and why this might form part of a common or basic knowledge of acting together. Another valuable lesson is

that, in breaking down both Bratman and Roessler's accounts, we've shown that the problem of explaining shared intention in the face of motivational uncertainty is common to both. Though their commitment to reducing shared intention only to individual attitudes is not shared, both authors adopt a form of methodological individualism to explain how agents *jointly* settle matters: Bratman springboards off his individual theory of intentions as plans, and Roessler begins with Anscombe's observations of practical knowledge when speaking about individual, not collective, intentional activity. The risk of this approach is that drawing on deep, well-trodden solutions to problems identified in contexts of individual agency can fail to provide good solutions in contexts of shared agency if we fail to fully specify and ensure are met all background assumptions, some of which may have originally been taken for granted. For example, when Bratman formulates his AT, though he doesn't say so explicitly, it's plausible he assumes that the uncertain agent is still the one in control of any future activity. Though perfectly reasonable in Bratman's original formulation, when mapped across to joint action this hidden assumption became visible because it's precisely the apparent lack of individual control over a joint action that makes explaining joint settling so tricky. Of course, it may at times turn out that background assumptions aren't violated in the shift from individual to shared, but as an artefact of the methodology taken it is not something we should take for granted. In our case, we *do* need a better explanation of how a joint settling requirement can be met in cases where there's motivational uncertainty. The next chapter continues in this vein.

CHAPTER 5

Social Commitments and Joint Settling under Motivational Uncertainty

The previous chapter discussed how a settling condition can be accommodated in the context of shared intention. It presented two diverging authors' views of how joint settling might be formally incorporated into theoretical accounts. Despite taking quite different methodological approaches, both authors rely on similar assumptions—including relying on the ordinary predictability or cooperativity of social agents—to ground joint settling. But in turning back to BEACH, we saw how introducing motivational uncertainty poses a challenge for these accounts. They both find it difficult, with the resources immediately available, to explain how it is that individuals can jointly settle matters when they have reasons to be sceptical about their partner's motivation to perform their part. This is because these reasons clash with the aforementioned background assumptions of cooperativity. Unlike in Chapters 2 and 3, in which we found solutions in the literature of individual intentional action to the problems identified, we haven't yet managed to reconcile the possibility of joint settling with substantial uncertainty about partner intentions. This chapter aims to do just that. It argues that a form of interpersonal commitment, a separate feature of some theories of shared intention, can ground the kind of reliance that would enable Mya to settle whether he and Iva go to the beach together, despite him being uncertain that she intends to join him. This approach holds promise, but, crucially, it requires that we reconsider the connection between intention and commitment established in Chapter 3.

5.1 Introducing social commitments in collective action

In the previous chapter, I made the initial case that substantial uncertainty about partner intentions threatens settling requirements on shared intention. We can lay out the argument to make this clearer:

(1) *Settling requirement*: It is characteristic of intentional attitudes that they settle matters for the intender about what she will do.

(2) *Continuity thesis*: An individual's intentional attitudes are characteristically the same across individual and shared intention.

(3) *Joint settling requirement*: From (1) and (2), shared intention requires that each individual settles matters about what the group will do.

(4) Individuals can rely on ordinary predictability and cooperativity of their social interaction partners for certain behaviours.

(5) Because of (4), individuals are in a position to confidently predict how their partners will come to intend (Bratman) or to take as given their partners have the intentions they think they have (Roessler), and so rely on them.

(6) Joint settling: (4) and (5) mean (3) is met.

That is, individuals acting together jointly settle matters because they are in a position to confidently predict and rely on their partner's intention to contribute. In addition:

(7) It's plausible that individuals sometimes intend a joint activity though they have good reason to be uncertain about their partners' intentions (e.g., in BEACH).

(8) If (7) is true, then (4) is false.

(9) If (4) is false, then (5) and (6) are false.

(10) (9) contradicts (7), and we have shared intention without joint settling.

If individuals aren't in a position to know or confidently predict their partner's intentions, then they cannot rely on them to settle matters. Thus, if we are prevented from seeing matters as settled because of good reasons to be uncertain about what our partners intend, then shared intention cannot function in its characteristic way. If it's plausible that we can, at times, have shared intention in circumstances with motivational uncertainty, then it appears that the argument is either missing a premise or one of its premises is false.

In Chapters 2 and 3 the approach was to use tools from our understanding of individual intentional action to help solve the problems faced. This approach does not, however, appear to bear fruit for the problems of explaining joint settling as per the argument above. However, we may not yet be convinced we have reached this violation, as there's more to say about the relation between (4) and (5). If ordinary predictability and

cooperativity is necessary for relying on partners (i.e., relying on them to be motivated to make their contribution), then the conclusion is true. However, if (4) is merely sufficient for (5), there may be other routes to the satisfaction of the latter. In this case, there could be both reasons to be uncertain about a partner's intention to play their part while concluding that they will, in fact, do so—indeed, even expecting and relying on them for this.

This helps us frame the question of what makes (5) true if (4) is false, a question of motivation—whether and to what extent agents expect their partners to be motivated to play their part in the joint action. It's possible, then, that an agent might perceive her partner as having reasons that motivate him both for and against making his contribution. This, in turn, provides her with reasons to be uncertain her partner will make his contribution and reasons to expect that he will. The question, of course, is what these alternative reasons are and, relevant to this project, how they might work to enable intentions to settle matters. An alternative approach, then, is to first look for possible solutions directly in the *social* nature of the interaction itself. We should not, of course, add anything new if we are to abide by the continuity thesis, but it's possible that there are existing features of the situation which are candidates to help reduce the motivational uncertainty that's present in BEACH.

One area to look for an answer is in the part of the literature that sees interpersonal commitments as typically involved in joint activity. A popular suggestion is to think of interpersonal commitments as one important way of reducing uncertainty about partner intentions. There are many different types of commitments we can make, which Michael and Pacherie (2015) (with help from Clark, 2004) break into a helpful typology. First, commitments differ by who their authors and recipients are, depending if they are the same person (self-commitment) or different people (other-commitment). Second, both types can be private, known only to author and recipient, or public, having an audience. Third, these can be unilateral, made by one party to another, or interdependent, in which each party makes a commitment to the other. Finally, these can be bilateral or joint, the difference in the latter being the inclusion of a shared goal. It is this final group with which I am concerned, joint commitments, in which two or more parties make commitments to each other, of which they're all aware, to perform some part of a joint goal. I'll refer to these as mutual, social or interpersonal commitments.

In the context of temporally-extended joint activity, social commitments typically reduce motivational uncertainty by shielding long-term benefits from cooperation against

short-term gains from defection. Under the right circumstances, a commitment does this by encouraging an agent to meet expectations her partner has of her, and so provide greater assurance about her future action performance on which he can rely (Michael & Pacherie, 2015). If both agents are committed to one another, then this assurance goes both ways. Social commitments thus make behaviour more predictable by providing grounds for those on the receiving end to rely on those making them, crucially more than in the absence of any commitments. One important function of commitments is thus to reduce uncertainty about a partner's willingness to contribute to the joint action, thus reducing uncertainty about intentions and, in turn, reducing uncertainty and settling matters about prospective joint action success (Michael & Pacherie, 2015). This is especially relevant when tempting alternative options are available to one or more parties, as in cases like BEACH with which we are concerned, where motivations to participate are more likely to be called into question.

Why do interpersonal commitments provide assurance, that is, why do people believe others will be motivated to meet their commitments? Much of the remainder of this thesis is dedicated to responding to this in contexts where there's motivational uncertainty. As a starting point, though, we can look for answers in existing theoretical accounts for something more general. In the literature on shared intention there are two broad views of what makes social commitments credible. What they have in common is the idea that for commitments to be credible they must impose costs on action non-performance if performance is expected. These can be economic, material or psychological, and they can be real, potential, or even opportunity costs. Costly repercussions ensure that commitments guide our actions in ways that 'cheap talk' wouldn't, and notably when perceived, attractive alternative options are available and which the committed person might be otherwise tempted to take.

Where the two views differ, though, is in terms of what commitment violations and associated costs are. The first sees a failure to meet one's commitment as a violation of certain basic norms and principles of agency. As already discussed, for example, Bratman sees intention as involving a kind of commitment to action that's part of a set of guiding norms of intention rationality. If individuals should plan rationally, then not being and acting committed to an action one intends is symptomatic of a breakdown of rationality. And given that these principles of reasoning usually help us in achieving our goals, this breakdown implies costs in higher inefficiency or action failure, even spilling over to other plans beyond the immediate action itself. Agents should therefore be motivated to abide by these norms to achieve their goals, making their commitments credible. The second view is that

commitments work because they generate obligations on the part of the person making the commitment to meet them. These obligations are directed towards a particular person, the receiver of the commitment, and meeting one's commitments means honouring said obligations and failing to meet them implies a failure to do so. Moreover, there are psychological benefits and costs attached to meeting these obligations, including emotions related to the social context and relation. An agent might therefore experience pride or satisfaction just by fulfilling their obligations, and their failure to do so might lead to feelings of guilt, shame or sadness. Mechanisms like these are usually taken to motivate agents to meet the person-directed obligations they have in joint activity and which thus make their commitments credible.

In summary, as commitments are a useful tool for reducing motivational uncertainty in joint action, they provide answer as to how agents who are mutually committed are in positions to expect and rely on one another to have certain intentions—so (5) above is met—and so jointly settle matters when there is substantial uncertainty about partner intentions—so (6) is met. On this view, if I count on you being committed to both your and my role in the shared intention, I can rely on you to make your contribution though I have reasons, arising from changes in our environment, to believe that you may not intend to do so. I can therefore rationally settle matters about our joint action, and so shared intention can perform its characteristic functional role in shared agency.

5.2 Bratman's interpersonal commitment in shared intention

I focus in this chapter on Michael Bratman's view of interpersonal commitment in shared intention for several reasons. First, to continue in and build on the vein of the thorough analysis of his account in the thesis thus far. Second, because although an account like Margaret Gilbert's is often held up as paradigmatic treatment of the subject, placing joint commitment at the heart of joint activity, sticking closely to a reductionist approach to shared intention rules it out (we will see why in my analysis of Gilbert's view in part of the next chapter). Lastly, because the role of interpersonal commitment in Bratman's account of shared agency tends to be underappreciated. It is, as we will see, fundamental to his explanation of why, in certain situations, intentions should be thought of as genuinely shared, and what it is that makes shared intentional agency different to other kinds of strategic interaction. Without a robust account of interpersonal commitment, Bratman's account of shared agency is surprisingly vulnerable to challenge.

To begin outlining Bratman's view of the commitment typical of shared intention, one approach, that tracks the insights gained in this thesis thus far, is to see it as a mirrored response to a parallel question about intention in individual agency:

1) How is my intention to *A* different to my prediction that I will *A*?

2) How is our shared intention to *J* different to our predictions that we will *J*?

Bratman's thoughts about commitment are already familiar to us from Chapter 3. It's core to his view that my intention to *A* involves a commitment to action that my prediction that I will *A* does not, the main response to (1). A version of interpersonal commitment also features heavily in his theory of shared agency, and because of his continuity thesis these are tightly linked. It's Bratman's view that any principles of practical reasoning which feature in a minimal account of shared agency emerge from the combination of his original guiding norms (of practical and intention rationality) and the groups' network of intentions. The notion of interpersonal commitment, therefore, is also constructed from individualistic components. As intention is bound up with commitment, the same applies to shared intention, such that shared activity absent this commitment cannot involve shared intention. This prods us to ask whether the answer to (2) is similar to (1). More specifically, is it sufficient for an action *J* to be intentionally joint (i.e., that the joint settling requirement is met) that I intend to make my contribution to *J* in light only of my expectation that you will make your contribution to *J*? If the answer is negative and, moreover, that what takes it beyond mere prediction is a kind of commitment in shared intention (mirroring the answer to the first question), then we have before us a notion of social commitment, essential to shared intention, which might prove a useful tool for reducing motivational uncertainty.

*

Whether prediction alone is sufficient grounds for joint action cuts to the core of certain challenges raised against Bratman's account, specifically, and reductive accounts of shared intention, more generally. These concern perceived difficulties they experience in explaining what it is that "makes joint action intentionally joint" (Pacherie, 2013: pg. 1818). This becomes clearer when we acknowledge there are plenty of cases of social interaction which meet several criteria for shared agency—such as having multiple agents involved, a common goal, intentional behaviour on the part of each, a mutual awareness of agency and the need for coordination (Butterfill, 2012)—but which lack a stronger notion of action that is

intentionally shared. Under a 'strong' sense of shared intentional action, "the individuals who engage in this activity must think of its goal not just as bringing about outcome *O*, but as bringing about outcome *O together*" (Pacherie, 2013: pg. 1821; author's emphasis). Explaining this is, of course, a primary task for any account of shared intention, but accounts like Bratman's which advocate a strong form of methodological individualism are often specifically criticised for providing too weak an explanation of this distinctive 'togetherness' or 'sharedness', or having explanations which appears circular (charges of circularity, to be fair, are also levelled at 'non-reductive' accounts).

There are two ways this general concern about weakness can be cashed out. The first frames the issue as leading to an account of shared intention which too closely resembles disconnected agents operating together in parallel, sharing some knowledge about the world but nothing distinctly collective about their interaction. The action is thus only *tokenistically joint*, in that there's an outward appearance of collectivity which belies any real, uniquely social, interconnection. Hans Bernhard Schmid, one of the more blunt critics, says:

> "Knowledge of intentional joint action, it thus seems, is each participant's self-knowledge of his or her own doing plus mutual ordinary knowledge (e.g., the common knowledge appealed to in received accounts of collective intentionality) of what the respective partners are doing. If this line of argument is sound, it leaves us with what we might call the singularist view—*the view that the only subjects that exist are singular subjects, and that whatever plural attitudes there are have singular subjects*.
>
> (…)
>
> Let us spell the consequences out in practice: Whenever you think it is actually one token tango dance which you and your partner intentionally perform together, you're under an illusion—*all there really is your intentional part and your partner's, perhaps with some structure of mutual knowledge so that it (hopefully) adds up to something that looks like that one token action*. But there is never one collective dance with many participants, but just several suitably combined individual dances" (Schmid, 2018: pg. 237–238; my emphasis).

As might be apparent, a characterisation of shared intention along these lines does not satisfy Schmid, nor would it those authors who regard approaches aimed at reducing shared intention

to individual attitudes as misplaced[16]. Roessler (2020), whose plural practical knowledge account of shared intention we saw earlier, puts the point more generally:

> "One philosophical question in this area is whether collective activities can be genuine cases of intentional agency. Of course, one answer would be that they can, insofar as they amount to nothing other than a complex combination of individual intentional activities. But it is far from clear that a credible reductive account of collective intentional agency is in the offing" (pg. 5).

It's hard, though, to spell out a general form of what non-tokenisation looks like. Much seems to come down to authors' intuitions about what joint action entails—suggesting perhaps an empirical rather than theoretical starting point should be preferred.

The second way of framing weakly characterised joint action focuses on individuals' motivations to participate in the joint activity. In particular, a need to steer away from the possibility of what we might call *purely instrumental* possibilities of joint action. Elizabeth Pacherie generalises this by suggesting a "joint goal requirement" as a key feature distinguishing intentional joint action in a strong sense from a weak sense. This requires that participants "should view their own actions as contributions to a we-goal (or joint-goal) rather than viewing their and the other agents' actions as contributions to some agent unmarked goal they happen to share (whether they are aware of it or not)" (Pacherie, 2013: pg. 1822). And part of this is that the joint-ness of their actions is not just a means to this goal but is, instead, part of the goal itself. For an activity to be a jointly intentional action in the strong sense, then, the individuals must think they are bringing about the outcome together, and part of what this means is that the joint goal requirement is met. Accepting the joint goal requirement helps paint a clearer picture of shared intention by, for example, ruling out certain cases of collective activity. Coercive or deceptive behaviour is the easiest and often first to be excluded, given questions about the goal's desirability to the agent under duress as

[16] Schmid, to be clear, appears unconvinced by not only reductive (what he calls 'singularist') accounts, like Michael Bratman's, which see shared intention as reducible to individual attitudes only, but also certain popular alternatives which regard shared intention as irreducible in this way, like Margaret Gilbert's. He says, e.g., that

> "… whatever self-knowledge there is in, or of, "we intend", it cannot differ from the way "I intend" is self-known, because whatever else "we intend" involves other than you yourself obviously concerns other people, and you can't self-know other people: they're not you, and you is all you can self-know. Cast in terms of determination rather than knowledge, the contradiction becomes even more blatant: about self-determination, you have to keep it to yourself—you can't self-determine other people. Thus it seems obvious in this line of reasoning that "we intend" does not involve a plural subject in the way "I intend" involves a singular subject" (Schmid, 2018: pg. 236).

well as fragility of any shared motivation to reach the goal. Purely instrumental motivations to participate in the joint activity also violate the joint goal requirement, an idea that draws on a rich tradition in the literature on collective agency concerning treating others only as a means to one's ends, what Pacherie describes as treating one's partner as a mere 'social tool'. The joint goal requirement is thus also a one way of excluding from an account of shared intention situations involving purely instrumental collaboration.

I have framed these as separate concerns partly because these arguments are sometimes advanced separately, by different authors, and partly because I don't see a reason why they can't come apart. It's plausible we can develop a non-tokenistic characterisation of shared intentionality that is ambivalent about whether agents treat one another as social tools or not. For example, Schmid's (2018) basic form of first-personal plural awareness is based on claimed psychological differences between individual versus joint intentional action. While he deals here with what I've called tokenism of shared intention, it doesn't seem to require a commitment to a particular requirement regarding self-interested behaviour.

That said, tokenistic and instrumentalist concerns are often entangled, with many philosophers appealing specifically to kinds of mutual, non-instrumental treatment to explain what gives a particular collective activity its distinctly 'shared' flavour. It is, as the reference to Pacherie above showed, exactly those unselfish motivations which are used to identify paradigmatic examples of shared intention. This is the route that Bratman takes, and though he isn't responding directly to the authors described above, at different points in his work we see explanations for how his proposal addresses both problems of tokenism and instrumentalism. In his later work, in particular, Bratman shows a clear sensitivity to these, going to trouble to emphasise how so in quick succession:

> "The thesis is that shared intention and modest sociality consist, at least in central cases, in appropriately interrelated public structures of individual planning agency. These interrelated planning structures go beyond the merely cognitive interrelations involved in knowledge of each other's minds and present in standard forms of merely strategic interaction.
>
> (…)
>
> They thereby go beyond merely cognitive links among the participants to capture an important way in which each is treated by the others as an intentional co-participant.

(…)

> The basic thesis works, in part, by building appropriate reference to the other into the contents of the intentions of each. While it acknowledges the potential roles of various unarticulated commonalities of sensibility, it does not just appeal to "a background sense of the other as a candidate for cooperative agency".
>
> (… and, finally …)
>
> A central thought of this discussion is that modest sociality, while consisting in appropriate forms of interconnected planning agency, is not merely strategic interaction within a context of common knowledge" (Bratman, 2014: pg. 87–92).

The last quote is illuminating. It tells us that Bratman thinks shared intention in the strong sense requires more than individuals simply best-responding to one another in contexts of shared knowledge—and, furthermore, that his account explains how. We will see shortly that it is precisely down to differences in purely instrumental motivations agents have in strategic interaction (to treat one another as social tools), on the one hand, and non-strategic behaviour in modest sociality, on the other, which sets these types of social interaction apart. In addition, evidence of these non-instrumental motivations is to be found in the kind of interpersonal commitment which characterises agents' relations. It is thus the presence of this commitment which will ultimately identify shared intention in a strong sense.

*

To understand what gives rise to social commitments, we need to understand the difference between strategic and non-strategic interaction. Bratman's structure of shared intention is built up from "basic norms of individual intention rationality", including norms of consistency, agglomeration, coherence and stability. Continuity between individual and joint activity means, he says, there are plausible, corresponding norms of *social* agglomeration, *social* consistency, *social* coherence and *social* stability, which constrain and guide collective behaviour and whose failure to satisfy usually undermines the ability of multiple agents to coordinate their actions and settle on roles and contributions (Bratman, 2014: pg. 27). Structures of *interrelated* planning agency are thus understood by appeal to the underlying norms core to individual intention rationality, and it is the interaction of these norms along with agents' interrelated intentions which anchors the social norms described and which causes agents to internalise them (Bratman, 2014: pg. 143). This is how social normativity in

shared intentional activity "involves the "emergence" of an explanatory role of norms of social rationality of intention" (Bratman, 2014: pg. 149). As the foundational norms are first and foremost a suite of *functional* norms and principles for supporting action, any 'social rationality' emerging from their interaction with multiple individuals' intentions should, we expect, be similarly grounded. Bratman emphasises this, describing social normativity as, for example, "a basic structure for explaining the main contours of socially rational shared intentional activity, including coordinated action and planning in the pursuit of a common end, and associated bargaining and shared deliberation" (Bratman, 2014: pg. 109). So who does what and when, how to align actions, plans and subplans, et cetera, anything which requires collective decision making or action in pursuit of the joint venture.

Highlighting this functional ethos means acknowledging that an important motivation for engaging in joint activity is, indeed, precisely because it enables us to achieve valuable outcomes we might otherwise struggle to achieve on our own. But given the aversion to instrumental motivations described earlier, there's now a possible tension between accounting for shared intention's functional role versus emphasising a separate social dimension which sees intrinsic value in simply doing something together with others. Distinguishing the value of acting together from the value of the shared goal thus presents a challenge for Bratman. To address it, he expands on his notion of social normativity by spelling out, in his set of sufficient conditions for shared intention, some specific dispositions and behaviours it entails. This includes intentions in favour of meshing subplans, dispositions to help if needed, the tracking of the joint activity and mutual responsiveness.

There's a problem with this approach, though. The dispositions and behaviours described are also consistent with what a rational, purely self-interested agent would likewise do well to adopt. Indeed, Elizabeth Pacherie (2013: pg. 1824) points out that they plausibly fall out of a coupling of certain, non-controversial norms already associated with individual planning and two of Bratman's key conditions for shared intention, namely:

1) Intentions on the part of each in favour of the joint activity, and

2) Interlocking intentions: each intends that the joint activity go in part by way of the relevant intentions of each of the participants.

Parsing Bratman's account to these two conditions (against a background of typical norms of agency) is useful because it tells us that this is exactly where we must find the 'sharedness' in

his account. Looking at them, we see it is not to be found in the first condition, because to avoid his account being circular, Bratman is clear that condition (1) is to be understood in a way that is neutral with respect to shared intentionality. The reference to the joint action featuring in each party's personal intention should not be thought of as presupposing any notion of sharedness. Thus, as Pacherie correctly notes, it is the second condition which is central to Bratman's explanatory view of shared intention. She says that

> "[i]t is the fact that for each participant, the content of their intention refers to the role of the intentions of other participants that, for Bratman, captures the intentional joint-ness of their actions.
>
> (...)
>
> [B]y conceiving of shared intentions as an interlocking web of intentions of individuals, [Bratman's account] moves away from the classical reductive analyses of collective action, since it maintains that the crucial link among the attitudes of agents involved in joint activity is not just a matter of mutual belief or mutual knowledge" (2013: pg. 1824–1825).

What we need from condition (2), then, is a characterisation of how this network of attitudes is different to, and goes beyond, mere 'strategic interaction' between agents or the kind of tokenistic action under mutual knowledge to which Schmid and Pacherie refer.

*

It is thus in the nature of the interlock between agents' intentions that we will find the aspects of Bratman's account that make it intentionally joint in the strong sense. What he has in mind looks nicely articulated in the following passage:

> "The basic thesis provides a model of the social glue that ties together the participants in modest sociality. According to this model, this social glue is not solely a cognitive glue of common knowledge, though it does involve a form of common knowledge. This social glue also includes the forms of intentional interconnection and interpersonal support…, beliefs about success and interdependence…, actual interdependence…, mutual responsiveness in sub-intention and action…, and the normative pressures of social rationality that emerge from these structures given relevant norms of individual plan rationality" (Bratman, 2014: pg. 87).

If the above is correct, the key to getting at the heart of the interlocking intentions in condition (2) is to understand this essential 'social glue' as binding agents through 'intentional interconnection and interpersonal support'. How are we to understand what this refers to or looks like? Something that looks like intentional interconnection appears in some of Bratman's early work, notably in his response to the Velleman-Baier challenge. As discussed earlier, settling and having control over matters is usually taken to be the purview of the intender. Bratman's response, recall, is that we're not confused in talking about joint settling if we accept that agents' intentions are persistent interdependent, and that individual's settle matters about and retain a sense of volition towards intended joint action by their intentions partially mediated via their partners'. If what Bratman means by intentional interconnection builds on this idea, then an intentions-via-intentions feature could be one way of understanding the social glue concept.

I think this is exactly what Bratman does in his work on shared agency over two decades later. He starts with this idea of intentional interconnection but fleshes it out with greater normative requirements: not only must agents see their intentions (in favour of the joint action) going by way of their partner's intentions, but taking this to be true means, in addition, they must intend that their partner's intentions are effective as well. This therefore transforms into the idea that agents who are truly interconnected in the right way each *intend* their partner's role and contribution in the joint action. Crucially, this is different to simply *expecting* them to make their contribution, or *predicting* that they will. Furthermore, because intention is characterised by commitment, this intending and not merely predicting places certain demands on participants to be committed to their partners, including to be willing to support and help them if required and to respond to them in ways that support their collective participation and acting together. An agent acting intentionally but based only on expectations of how a partner will act is not characteristically committed in the same way. They face no such demands and do not necessarily expect their partners to be committed.

*

This is the broad stance that I see Bratman as taking in defending against the critiques of tokenism and instrumentalism. Some of this he formulates in response to specific authors, including the following challenge from Björn Petersson:

> "Suppose I want the window smashed. When I note your presence on the street, I think that if you act in a certain way, the window can be smashed as a result of both

> our acts, and I form an intention accordingly. What I intend in that case is merely to get the window smashed, while predicting that your actions will be components in the process leading to that result. This prediction may rest upon my knowledge that your intentions are similar to mine, and that our subplans are likely to mesh in a way that enables me to reach my goal. There is mutuality and interdependence, in line with Bratman's requirements. Still, I would say, nothing in this picture captures "sharedness" or "collectivity" in any sense distinct from what we can construe in terms of standard individualistic theory of action" (Petersson, 2007: pg. 140).

Petersson's point is two-fold: that here, I am behaving strategically, as in I am merely best-responding to what I expect you to do; and that strategic behaviour is insufficient grounds for shared intention in the strong sense. We should not, he concludes, in this case say that we are breaking the window together. Bratman, in his response, agrees with Petersson on the second point but essentially rejects the first. He rejects that the example undermines his account on the grounds that his criteria for modest sociality are not met in the first place, as "Petersson's description appeals at crucial moments to prediction when what is required by the basic thesis is intention; and it is a central theme of the planning theory that these attitudes differ in systematic ways" (Bratman, 2014: pg. 93). This is a clear example of the difference between intending and predicting a partner's intentions from the previous section.

To better understand the distinction between these, we need to know what intending entails that prediction does not. The core idea, as discussed at the beginning of this chapter, lies in the commitment associated with the former. So we need to get a better sense of what social commitment looks or is experienced like in shared intention. Interestingly, Bratman doesn't provide a direct characterisation of it. We can, however, glean what this might be from various examples he gives of its manifestation in shared activity. For example, he says:

> "Though in Petersson's example I expect that you will act in ways that promote the smashing of the window, it is not clear from the description of the example that I intend that. Perhaps I have no disposition at all to help you if you need it, or to reason about means to support you in your role, or to filter options incompatible with your playing your role. And though I expect your intention to be effective it is not clear in the example that I intend that. So it is not clear in Petersson's example that I intend our joint window smashing in part by way of your intention.
>
> (...)

> To be sure, in each of these cases of strategic interaction, the participants intend to act in certain ways given that, as they expect, the other will act in certain ways. But in each case we should resist the inference from S intends *A, given that (as she expects) the other will B*, to S intends *(the joint activity of A and B)*. After all, S may intend *A, given that (as she expects) the other will B* without any disposition to filter out options incompatible with the other's performance of B or to take the other's performance of B as an end for her means end reasoning or to act in order to support the other's performance of B. And when we resist this inference, and insist on the distinction between intending and expecting, we are in a position, in the words of Gold and Sugden, "to differentiate collective intentions from the mutually consistent individual intentions that lie behind Nash equilibrium behavior."" (Bratman, 2014: pg. 94–96, author's emphasis).

Of course, intending and not only predicting a partner's actions must be *mutual*. For us to share an intention not only must I intend the effectiveness of your intention, but you must intend the effectiveness of mine. We see this in another of Bratman's examples, involving walking alongside a stranger:

> "Suppose you are walking alongside a stranger, and you are each acting in ways that are in strategic equilibrium in a context of common knowledge. Each knows what the other intends to do and does; each pursues what he wants or values in the light of this knowledge of the other, knowing that the other is reasoning in a parallel fashion; each knows that if both so act there will be a coordinated concatenation of their walking actions; and all this is out in the open. And now the important point is that such public strategic interaction need not satisfy the conditions of the basic thesis.
>
> First, though each believes that there will be the cited coordinated concatenation of walking actions, it does not follow that each *intends* that. To intend the coordinated concatenation each would need to be disposed to take that complex of activities both as an end for his own means-end reasoning and to be guided in action by this end; and each would need to be disposed to filter potential options for deliberation with an eye to their compatibility with this end. But it may be that none of this is true of you or the stranger. Perhaps the stranger does not *intend* (though he does expect) that you will act in these ways, and has no disposition to help you if you need it. Indeed, perhaps he is looking for ways to thwart your progress down the street

without physical violence, even though he sees that you are indeed progressing down the street and he is doing what he thinks best given that you are. This stranger does not intend that the two of you walk together down the street. Given what he knows to be the limits on his powers, he does expect that you will in fact walk in the way you are walking. And he intends to respond to that, and so expects that there will in fact be a coordinated concatenation of the walking actions of each. But this is not yet to intend that coordinated concatenation.

Again, perhaps the stranger expects that your walking will be the issue of your relevant intentions and yet does not *intend* that. Perhaps he is keeping his eyes open for a preferred mechanism that would issue in your walking in a way that bypasses those intentions of yours. Being a realist and not especially strong, however, he does not believe that this is what will happen; he expects that you will walk by way of your relevant intention, and he does what he sees as best given that. But he does not intend that your intention be efficacious, and is set to thwart this if an appropriate opportunity should arise. So his intention does not appropriately interlock with yours. It follows that this case of walking alongside a stranger does not satisfy the conditions set out in the basic thesis. So the basic thesis can say this is a case of strategic interaction that is not a case of modest sociality. And that is what we wanted" (Bratman, 2014: pg. 92–93, author's emphasis).

These examples provide a non-exhaustive list of what the commitment aspect of Bratman's emergent social rationality looks like. It includes dispositions to help one's partner should she need it, not attempt to thwart her intentions, not attempt to exploit her should the opportunity become available, not be acting only in one's own interests (as one thinks best) given what's available to do; in short, to not behave strategically.

5.3 Commitments, not only intentions, can settle matters

The idea of interpersonal commitment provides a plausible solution to the problem of how there can be shared intention when there's motivational uncertainty. But it requires a significant shift in how we think intentions and commitments are connected, which I'll get to.

The main idea I'm proposing is that if an agent is in a position to rely on her partner to be committed to the joint activity—and to her contribution to and participation in it—then it's possible that her intentions can settle matters in the way required for shared intention *even*

in contexts where she is uncertain what her partner intends. First, she can settle matters about what he as well as they together intend and will do. Second, if these commitments are mutual and are common knowledge, then she also knows that he regards her as committed in the same way, and she also knows that he knows that she regards him as being committed in the same way, et cetera; each knows that they and the other are committed to the joint action, and furthermore each knows that they are together committed to the joint action. This means each can also settle matters about what they personally will do and, importantly, what they will do together. And so they are in a position to jointly settle matters.

Importantly, interpersonal commitments between you and me ground joint settling in two ways. First, in terms of supporting each of our expectations of how I and you will act. Second, in terms of supporting the volitional sense for each of us that the joint action is up to us both individually and together. They facilitate the kind of mutual responsiveness and persistence interdependence between our intentions that supports me seeing my intention (that we *J*) as partly mediated by your intention (that we *J*), and vice versa. This provides the control-like element that we sensed was missing in the direct application of Bratman's Asymmetry Thesis at the end of Chapter 3.

In line with this idea, Fernandez-Castro and Pacherie (2022) argue that commitments shape the sense of joint agency that multiple agents experience when doing something together. They first reflect how the experience of joint action differs from individual action in several ways, including an expanded complexity in predicted action consequences, asymmetries in roles, expertise, hierarchy, et cetera, among participants and the distinctive emotional experiences and affective states they have as the joint action unfolds (pg. 3). The authors then reflect on the role of commitments in each of these areas, and of particular interest for our discussion on settling, one argument they make is that

> "[c]ommitments exhibit an important normative element that is manifested in the fact that … each party is entitled to holding the other party responsible for their duties and, for instance, to engage in regulative actions … when a commitment-based expectation is frustrated. Such a normative status results in a greater capacity for exerting control over the co-agent, and consequently, in a stronger feeling of control over the joint action …. Furthermore, the normative element of commitments may also shape the sense of joint agency by counterbalancing or reducing the possible disruptive effect that various asymmetries among members of the group may cause … [repairing] the

> feeling to the extent that they endow the individuals potentially affected by such asymmetries with the capacity to exert normative control over the actions of others" (pg. 12).

Interpersonal commitment, if in place, boosts the feeling of control among the parties and so enhances their sense of joint agency. This feeling of control also plausibly fulfils part of what's required for joint settling. It is therefore the combination of predictive and agentic qualities that, as discussed in this and the previous chapters, provides a basis for joint settling and which the presence of mutual commitment seems to satisfy.

Bratman's argument is that shared intention in the strong sense requires that all involved intend and not only predict their partners' contributions, which entails them being committed to supporting them. This commitment is driven by normative pressure grounded in a combination of the norms of practical reasoning available from individual intentional action and the contents of the agents' intentions. This provides a way of reconciling shared intention with the presence of substantial uncertainty about intentions. When agent's have reasons to doubt their partner's intention to play their part in a shared activity, if interpersonal commitments are present and public then it's possible that these can provide alternative reasons for relying on them. Social commitments are therefore one tool for reducing motivational uncertainty, precisely what was posing the problem.

*

As mentioned earlier, the idea that commitments can reduce motivational uncertainty is not new. Bratman's idea of social commitment is part of a broader, popular view in the joint action literature that such commitments are a useful mechanism for providing assurance in cooperative activities, notably those in which there are alternative, potentially preferable options available. As shown in this chapter, though, Bratman's notion of social commitment is tightly connected to his notion of intentional attitudes; social commitment emerges from the network of interdependent individual intentions.

This raises a concern. If interpersonal commitment is purely intention-linked, in that intentions and commitment cannot come apart, then this presents a potential problem for the solution I've just proposed. For if it's your intentions I am uncertain about, then I must, in some sense, also be uncertain about your commitment. If this is true, it's not clear why I should be able to rely on your commitment rather than your intentions in the first place.

Indeed, on our reading of it, Bratman's view of social commitment does seem conditional in this way: I am committed to supporting the effectiveness of your intentions *given that I intend* the joint activity, and you are committed to supporting the effectiveness of mine *given that you intend* the joint activity. I should therefore only *expect* you to be committed to me if you already intend the joint activity, and so if I am uncertain that you intend our joint activity then I should also be equally uncertain that you are committed to supporting my part in it. To roughly summarise the concern: because social commitment emerges *from* the network of collective intentions and not the other way around, we don't have grounds for thinking how interpersonal commitments can shape intentions, which is what I've proposed as a solution to the problem of motivational uncertainty in the previous section.[17] As it stands, without intentions we are left without commitments.

This has big implications for the possibility of there being shared intention in cases like BEACH. Most importantly, it must mean, for interpersonal commitment to provide the solution I suggest it does, then it must be the case that even in a situation where I am uncertain about your intentions, it must be possible that I can simultaneously *not* be uncertain about your commitments. This implies that intentions and commitments can come apart, even if we take it that they ordinarily do not. This opens many avenues for analysis, but I want to focus on the topic at hand, which is the relevancy of this for joint settling—or more specifically how there can be joint settling in contexts involving uncertainty about intentions. In BEACH, it's still required that Mya must see his and Iva's intentions as collective, and so their intentions must settle the matter collectively because the matter can't be settled twice. But now, what's going to settle for Mya that he and Iva go to the beach has to be through Iva's commitment, even if he is uncertain that she intends to go. Mya must therefore see his intentions as persistent interdependent with Iva's intentions or commitment—that is, he sees his intention that they go to the beach as dependent on her commitment that they go to the beach (and where he knows she knows that he is aware that she is committed in this way). The idea is therefore that it is the *commitments*, and not the intentions (or not only the intentions), which settle the matter. This doesn't work if we think of commitments as consequent on intentions, because then if you don't have the intentions then you also don't

[17] It's important not to conflate uncertainty about whether you're committed, my focus here, with uncertainty about whether you'll follow through on your commitment, given that I know you are committed. The latter is interesting to analyse in social contexts, but raises questions of willpower and agency rather than questions about motivation and why agents should think others will feel committed in the first place.

have the commitments, and so commitments can't settle the matter; uncertainty about intentions, in this case, just ought to mean uncertainty about whether it's settled at all.

Is the idea that intentions and commitments can come apart a credible one? Well, if Mya and Iva share intentions, what are they settling? Going to the beach together. What settles the matter? As shown in Chapter 4, the way we originally understand joint settling is that our intentions have to settle the matter, and they have to do so collectively. The question now is, is that changing: do Mya and Iva's intentions have to settle their going to the beach together, or is it reasonable that, say, it's their intentions *and* interpersonal commitments which together settle the matter? I think the latter is plausibly true. If there *is* shared intention in BEACH, then intentions and commitments can't be one thing, because uncertainty about the former would mean uncertainty about the latter. But we needn't think they are entirely separate, arbitrarily related things either, because then it wouldn't make it reasonable to think of commitments as settling anything to do with intentions at all. So we are not suggesting the role of individual intentions in settling is redundant. An obvious question then is what the relationship between intentions and commitment is, and a straightforward answer is that commitment should *characteristically* drive intentions (as Bratman says they do), even though in this particular situation Mya doesn't know whether Iva's commitment to him will guide her intention in favour of the joint activity. So they are often importantly interwoven, but they are not glued together.

Another reason that pursuing an analysis of the separation is useful is because it pushes us to develop a theory about commitments in shared intention in order to accommodate motivational uncertainty. In accounts like Bratman's, social commitments and the associated obligations to meet them are by-products of an initial, narrower focus on the more psychological aspects of joint action. They are not seen as essential to shared intention and so are stripped out of any primary analysis of the phenomenon, an approach that Bratman describes as a 'division of philosophical labour' and which I challenge in the next chapter. If we avoid treating commitments as closely-related to but distinct from intentions, I will argue: first, we lack a credible explanation for why commitments motivate agents to act, in particular under conditions of uncertainty; and, second, commitments have little power in explaining what it is that makes shared intention different from other kinds of collective activities—which involve intentional agents but which lack a genuinely shared character—as authors like Bratman argue they have.

5.4 Conclusion

In Chapter 4, I argued that two quite different accounts of shared intention find it difficult to explain joint settling in conditions of motivational uncertainty. This is because background assumptions of cooperativity, which I've said are more taken for granted than argued for, don't hold up in these cases—and agents can neither 'go ahead' and simply rely on their partners to make their contribution nor, following this, settle matters about what they will do. Substantial uncertainty about another's willingness to contribute, I said, means a joint settling requirement on shared intention is not yet explained. In turn, this means that we are missing a good defence for why motivational uncertainty doesn't undermine the possibility of shared intention, as in cases like BEACH.

The aim of this chapter was to explore a possible solution to this problem in two steps: first identifying and then filling in the gap of what's required for Mya and Iva to genuinely share intentions. The first step was to look for insights into what exactly is still missing by picking up on the intuition we had at the close of Chapter 3; namely, that the AT opened the door to there being shared intention in situations where agents intend a joint activity based only on predictions of their partners' actions. Using Bratman's view of interpersonal commitment helps us understand why prediction alone might be an insufficient basis for shared intention. In this view, being committed to, for example, supporting one's partner should they require it is essential to what it means for a shared activity to be intentionally joint in the strong sense. We realise that the AT, while perhaps a valid norm of rational action, doesn't tell us anything about whether agents are or are not committed in the right way.

The chapter's second step was to focus on this overlooked featuring and role of interpersonal commitment to see if it could act as a foundation for the kind of joint settling required in shared intention identified in the previous chapter. Given that social commitments are, in the literature of joint action, a popular tool for thinking about reducing motivational uncertainty, it seems reasonable that they might. Indeed, Bratman's interpersonal commitment looks like it provides a plausible explanation for why agents can rely on their partners and depend on them to make their contribution to the joint activity, despite having other reasons to be uncertain whether they will. If agents are committed to one another, and this is common knowledge, then each may be in a position to reliably predict how their partner will act and so, as per the previous chapter, they are in a position to jointly settle

matters despite the sense of uncertainty. I say 'may' as the point is there's no *guarantee* one's partner will perform their part; rather, commitments must be weighed up against additional considerations why one's partner may abandon the joint activity. Nonetheless, the gist of this view is that if Mya counts on Iva being committed to both of their roles in the shared intention, he can rely on her to make her contribution though he has reasons to believe she may no longer intend to do so. And so Mya can plausibly settle matters about their acting together, and shared intention can perform its characteristic functional role of leading to shared agency.

In the last section of the chapter, I discuss the ramifications of this approach. In Chapter 4 we saw that both Bratman and Roessler have a problem explaining joint settling in the face of motivational uncertainty. I have suggested that the same idea provides the solution for both of them; namely, we have to think about commitments as distinct from intentions, and things that can exist even when there are no intentions—although this is plausible only if we agree that they do characteristically generate intentions and we have a theory about their connection to intentions. So even if they are usually connected, the solution requires accepting that intentions and commitments can come apart, and that it is intentions, commitments *or* both which can settle matters in joint action. This places an important emphasis on commitments for explaining joint activity, notably in contexts where there's motivational uncertainty. Whether the view of commitment in question credibly does so is the topic of the next chapter.

CHAPTER 6

Are Theories of Commitment in Shared Intention Credible?

The previous chapter found a possible solution to the problem of uncertainty about a partner's intentions. Reasons I have to be uncertain about your intentions, generated by the emergence of attractive alternatives and perceived changes in your interests, can be balanced with other reasons I may have to believe you will remain committed to our joint activity. Because, in the normal course of things, our shared intention means that we are both committed to supporting one another's participation in the joint activity, weighing up these reasons I might decide to rely on you to make your contribution and so settle matters about our acting together. This chapter will argue, however, that a straightforward use of Bratman's version of interpersonal commitment to solve our problem is, on closer inspection, not well-founded. This is because the kind of commitment he proposes is not credible in situations involving motivational uncertainty—a surprising insight given that it is precisely in these sorts of contexts that social commitments are thought to perform a useful function.

While this bears on the question of the possibility of shared intention in BEACH, there are more general questions raised by commitment credibility worth addressing. Problems of commitment credibility in contexts with motivational uncertainty are not, though, limited to Bratman. I briefly look at Margaret Gilbert's account of joint commitment in shared intention and Berislav Marušić's recent work on trust to show that similar concerns apply. None of these authors have a ready answer as to how Mya and Iva can share intentions despite one or both of them being uncertain what the other intends.

6.1 Credibility concerns with Bratman's version of interpersonal commitment

For them to work, commitments must be credible. The receiver must be able to rely on the commitment to provide normative guidance to the maker to make the latter more likely to adhere to what they have committed to doing than prior to committing. The major problem we face is that, for all his descriptions of what committed behaviour *looks* like, what Bratman

hasn't given us are good reasons *why*, in the first place, agents would be moved to act committed. Because his version of interpersonal commitment emerges purely from norms of planning in his individualist account, features that guide Iva's decisions in terms of her own interest don't give us a reason why Iva will resist reconsideration based purely on features of the situation related to her relationship to Mya. In a recent paper Fernández-Castro and Pacherie (2020: pg. 8–9) make this point, arguing that Bratman's account of shared intention doesn't explain exactly why people are motivated to meet their commitments (their criticism extends to Margeret Gilbert's account too, which I discuss later in this chapter).

The primary issue is that Bratman gives us little insight into why agents wouldn't abandon their commitments when it suits them. Norms of intention rationality have an instrumental normative significance—that is, they are useful insofar as complying with them helps us attain intended and desirable ends—but the problem Fernández-Castro and Pacherie point out is that the norm of stability, which underlies intention's commitment to action, doesn't say anything about why agents would be motivated to comply with this rather than reconsider previously formed intentions. Notwithstanding cognitive constraints, it's still the case that practical rationality would demand that we reconsider our intentions were there valid reasons for doing so, as in case of an obvious beneficial change in interests. This is a worry for an account of individual agency but it is magnified in the shared case, as the intentional interconnection described in the previous chapter means one's own intentions can fall apart when others reconsider theirs, as Fernández-Castro and Pacherie (2020) discuss:

> "… 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 … 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.
>
> (...)
>
> In other words, persistence interdependence may create a domino effect and lead to the unravelling of the whole structure of interrelated intentions, without this involving irrationality" (pg. 8–9).

This tension comes from seeing persistence interdependence coupled with an understanding that people can have different reasons for acting together. This makes the shared intention vulnerable to the possibility that one or more members may reconsider their reasons for participation. Perhaps even more pressing, while in individual agency this may be less of an issue, given that both the benefits and costs of any reconsideration are borne by the agent herself, in shared agency the costs and benefits may not be spread equally. This means that, to understand how shared activity is supported, it's even more important to know how agents weigh up their options. The authors conclude that

> "to solve the [commitment] 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 implausibly strong claim and certainly a claim that Bratman does not explicitly endorse … 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" (Fenandez-Castro & Pacherie, 2020: pg. 9).

*

Bratman's account of shared agency thus gives us little insight into why any sort of underlying *social* motivation would lead one to avoid resistance to reconsideration. It does not suggest that there is any *socially-derived* normativity to underpin commitment. This leads to the credibility problem described. This is, though, perhaps too narrow a reading of Bratman's view, and we can be more generous in our interpretation of what he's referring to with interpersonal commitment. More specifically, he's given us specific examples of dispositions and behaviours which might be considered the result of other guiding normative principles. From these we might infer certain underlying motivations of agents who adopt them. If we get this right, we might conclude that implicit in Bratman's account are factors which give us an insight into why, based on social concerns, an agent might remain

committed; that is, which provide a response to Fernández-Castro and Pacherie's point above regarding lack of clarity about motivations.

For instance, Bratman's notion of social rationality includes dispositions to help one's partner should he need it, avoid taking opportunities to exploit him and generally support his intention to achieve his specific goals. We might infer from this that there's a sense in which agents are motivated, at least partly, to act in their partners' interests. Such social preferences are hardly controversial, and modern rational choice theory is full of different ways to incorporate them. It's plausible to think—and reasonable to think Bratman would be fine with this—that these preferences might feed into agents' motivation to be committed, including by influencing whether intentions are reconsidered when interests change. A wider interpretation of Bratman's ideas therefore provides possible social reasons for why agents might be motivated to meet their commitments, partially addressing concerns about credibility. This could, in turn, open up a pathway to explaining how commitment might address uncertainty about intentions, for example if an agent can rely on her partner's care for the agent's own interests to motivate him to make their contribution.

However, additions with important explanatory power like this rarely come for free. If we ask how it is that an agent can rely on the social preferences of others, it requires explaining why, in the first place, these social preferences are triggered in this particular context and directed towards that particular agent. An explanation of this looks like it will either require an appeal to some pre-existing sense of sharedness, group identity, or the like, and so lead us in a circle, or an appeal to some irreducible form of sociality, and so violate the continuity thesis.

What's required, then, is an explanation of how the dispositions and behaviours of interpersonal support, proposed by Bratman as examples of his claimed social rationality at work, *emerge*. Does an investigation into possibilities for the emergence of these supportive dispositions justify us saying we no longer have a problem of commitment credibility? A positive answer would provide us with a grounds for agents' reliance on one another, even in contexts of uncertainty.

*

Consider the table below, which explores, in different interaction contexts, whether or not agents acting with others are likely to be disposed to support an interaction partner should

they need it. Table rows differentiate situations in which agents are motivated purely by personal material interests—that is, strategically—and in ways which diverge from purely instrumental motivations—that is, non-strategically. Table columns differentiate situations in which agents' material interests are or are not aligned. So we can think of Col. 1 as involving perfect or near-perfect alignment and Col. 2 as involving known attractive alternatives for one or more of the agents, which may or may not tempt them to abandon (or even fail to begin) the shared activity. (The binary presentation of interest alignment is not necessary but makes the following argument clearer). The introduction of Col. 2 to a table containing only Col. 1 is equivalent to the idea of introducing motivational uncertainty into joint action contexts. It's the fact that Iva has an attractive alternative which gives Mya pause to consider whether Iva truly did or still intends that they go to the beach together.

Table: Strategic interaction versus interest alignment in various joint action contexts

	[Col 1] **Aligned material interests**	**[Col 2]** **Non-aligned material interests**
[Row 1] **Strategic interaction**	[a] Supportive dispositions and behaviours	[b] Supportive dispositions and behaviours depends on outside option
[Row 2] **Non-strategic interaction**	[c] Supportive dispositions and behaviours	[d] Supportive dispositions and behaviours (possible, depends on relative value)

Our core question is whether Mya can rely on Iva to intend that they go to the beach together when he is aware of the attractive alternatives available to her (so they are in Col. 2). From the previous chapters, we think that Mya can settle matters only if he is in a position to be justifiably confident that Iva will meet him at the beach. What we're exploring is whether Bratman's account contains within it a solid grounding for a kind of interpersonal

commitment to support Mya settling matters. Can we strengthen our understanding of what makes commitments credible by, crudely, backwards engineering pro-social preferences from the forms of interpersonal support Bratman is proposing?

I think that we cannot. This is partly because, as already discussed, his account of individual intention rationality from which the idea of commitment stems makes no mention nor requires no role from this. But it's mainly because, first, we cannot be sure *why* agents have supportive dispositions and behaviours towards their partner. This might emerge through non-strategic interaction (as Bratman wants it to) but it cannot be assumed to, given that such dispositions and behaviours are also consistent with strategic behaviour. And second, because of the above, a generalised notion of social commitment that's said to be marked by these dispositions and behaviours is not credible. This interpersonal commitment cannot be assumed to emerge from contexts in which interests are not perfectly aligned.

To see why, first consider the case where interests are perfectly aligned (Col. 1). See that we are either in Row 1[a], with agents acting purely strategically, or Row 2[c], where agents are not (or not only) acting in purely strategic terms. We're asking where the supportive dispositions and behaviours stem from—or, in Bratman's terms, why these norms of sociality emerge. In [c], these supportive behaviours are driven by motivations related to social preferences and concern for others' interests. There are normative pressures here to support one another in making contributions and participating in the joint action, including when it's not in our material interests to do so. However, agents in [a] who are driven purely by self-interest can act in ways which look on the surface like the sorts of supportive dispositions and behaviours as in [c]. Most obviously, if it is in my interests to achieve our joint goal, and if our joint goal requires both your and my contributions, then it is in my interests to support you in making your contribution should you struggle to do so. In Petersson's window smashing example, I want the window smashed and so do you (though you don't know that I want this too). I notice that the rock you are holding is not likely to be large enough to do the job, so as you search for another object I secretly toss a larger brick into your path, expecting you'll see it and use it to smash the window. In this case, it's in my interests to assist you, so this supportive behaviour can be present even without us sharing an intention to smash the window and me acting based only on expectations of what you will do.

The idea that social norms can emerge from purely strategic behaviour is not new. David Lewis' canonical treatment of the emergence and persistence of conventions argued

this succinctly. Lewis showed how two strategic agents engaged in a coordination problem can successfully coordinate given only a background of agent rationality and a basis for common knowledge. Shared norms for coordination can emerge simply from the rational behaviour of ordinary agents, with only minimal background requirements that include the "mutual ascription of some common inductive standards and background information, rationality, mutual ascription of rationality, and so on" (Lewis, 1969: pg. 56), a system of agents preferences, a system of "concordant mutual first- and higher-order expectations" (and an assumption about their cut-off point) and normal background understandings that preclude, for example, sabotage or extremely unpredictable events. Assurance that one's partner will select the right action or make the right contribution is effectively guaranteed by these features, meaning that "common knowledge of rationality is all it takes for an agent to have reason to do his part of the coordination equilibrium. He has no need to appeal to precedents or any other source of further mutual expectations" (Lewis, 1969: pg. 71). If it's plausible to extend Lewis' ideas beyond only giving an agent a reason to make their own contribution and towards giving them a reason to support their partner in making theirs, then supporting one's partner should be rational in order to boost the chance of coordination. Pointing at the kinds of mushrooms to look for when foraging in the forest will help both you and I in making our famous autumn stew while avoiding getting poisoned.

This is not to directly compare Lewis and Bratman's accounts. Rather, the former suggests one reason for thinking that in the kinds of simple coordination activities Bratman has in mind, supportive dispositions and behaviours can be supported by purely instrumental motivations in contexts *where interests are aligned*. Turning back to the table, it's plausible that the kind of social rationality Bratman sees as characteristic of shared intention can actually emerge in both [a] and [c]. From the outside, the types of interpersonal support could look very similar despite important differences in what motivates agents to provide them. This suggests that the presence of supportive dispositions and behaviours is not particularly helpful for uniquely identifying whether [a] or [c] is the case when agents' interests are aligned, that is, for identifying whether or not an agent's commitment to her partner is motivated by purely instrumental concerns.

This implications for thinking again about the credibility of the theory of commitment in Bratman's account, which emerge when we introduce into an interaction context attractive alternatives that test agents' commitment to their partners. This is tantamount to a shift to cases in which interests are no longer perfectly aligned in Col. 2, with the problem now that

we don't know whether, after this move, we are in [b] or [d]. There are many ways such a shift may occur, as with Mya becoming aware of Iva's attractive football-watching option. Or the environment might change for both: arriving at the beach and seeing some large swell, Mya, a good surfer, might instead hire a board and head out to surf by himself. While before we may have gotten away with being indifferent as to whether we were in [a] or [c]—given that behaviours supportive of the joint activity likely emerge in both—it's now crucial to the success of the shared activity to know whether we are in [b] or [d], as interpersonal commitment is no longer guaranteed to emerge—at least not without making additional assumptions.

6.2 The risks of a weak theory of interpersonal commitment in shared intention

The fundamental point from the previous section is that our inability to determine the motivation underlying commitment when interests are aligned means we don't know whether this commitment is credible in cases where interests no longer align. This is primarily because the supportive dispositions and behaviours Bratman uses as evidence of interpersonal commitment in shared intention do not, it seems, correctly identify the kinds of non-strategic behaviour that would make it credible. We don't know if agents are simply committed because it's in their interests, and so we are not in a position to assess whether and in what contexts agents are motivated to keep their commitments when interests are no longer aligned.

One obvious implication is that it doesn't give us an explanation for when and under what conditions people are motivated to remain committed to a joint action. This means that we do not yet have a reason why Mya can rely on Iva to be committed to their joint activity, given that Iva may be neither motivated to make her contribution nor disposed to support Mya should he need it. Absent an understanding of what makes commitment credible, we are back at the point where we don't have a clear idea as to how to overcome the problem of how there can be shared intention in contexts involving uncertainty about intentions.

Moreover, while this is a problem for our analysis of BEACH, it's crucial to point out possibly more widespread risks to Bratman's account—and perhaps to accounts of shared intention more generally—that the conclusion of too weak an account of interpersonal commitment would pose. First, a lack of social commitment credibility leads to possible confusion within Bratman's account itself. Second, it raises questions about how

generalisable Bratman's proposal for shared intention is, if, as seems to be the case, commitment isn't really playing any role in supporting and explaining an ongoing joint action. And third, we might lose sight of what it is that makes a joint activity intentionally joint in the strong sense required.

*

If we care about developing sound theoretical and descriptive accounts of shared intention, then we'll be worried by our inability to properly establish social commitment credibility. At best, we're peering through a fogged up window, with hints and glimpses of what 'glues' people together but no clear explanation. We can articulate this fogginess by drawing on Bratman's own examples to show where these exact questions about the need for interpersonal commitment and support lead to contradictions in his own proposals of a minimal account of shared intention.

Consider Bratman's well-known example of the unhelpful singers who intend that they perform together. These two first crop up in Bratman's description of what he calls *Shared Cooperative Activity* (SCA) (1992) and are then mentioned in a footnote in his later book on shared agency (2014). In their first showing, Bratman describes how these singers intend to perform together in spite of the fact that each would prefer the other to mess up (perhaps because it would cement them as the best), as simultaneously neither expects this to happen and so goes ahead and firms up their intention to sing it. On whether this counts as a case of shared intention, he says:

> "Return now to feature (iii) of SCA: the commitment of each agent to support the other's attempts to play her role in the joint action. Do the attitudes cited so far ensure this feature?
>
> To some extent they do. Suppose I intend that we sing the duet together. I am committed to pursuing means and preliminary steps I believe to be necessary for our so acting. That follows from demands of means–end rationality on my intentions. So I am committed to helping you play your role in our joint action to the extent that I believe such help to be necessary.
>
> But what if I believe that you will not need my help? I might then intend that we sing together and still not be at all prepared to help you should you unexpectedly need it. Consider, for example, the case of the unhelpful singers: You and I are singing

> the duet. I fully expect you to get your notes right, and so I intend to coordinate my notes with yours so that we sing the duet. But I have no disposition at all to help you should you stumble on your notes; for I would prefer your failure to our success. Were you unexpectedly to stumble I would gleefully allow you to be embarrassed in front of the audience — as I might say, "One false note and I'll abandon you to the wolves." And you have a similar attitude: You fully expect me to get my notes right, and so you intend to sing your notes in a way that meshes with mine. But were I to stumble you would not help; for you prefer my failure to our success. We each intend that we sing the duet in the world as we expect it to be, and we each intend that we do so by way of meshing subplans. But we do not have commitments to support each other of the sort characteristic of SCA. If we, as unhelpful singers, do in fact sing the duet together our singing may be *jointly intentional*; but it is not a SCA" (Bratman, 1992: pg. 103–104, author's emphasis).

This last line makes it clear that, according to Bratman, commitments to support one's partner are not essential to the singers sharing intentions. He goes on to say that it's very likely that at least some nominal requirement to provide help if required—in some 'cooperatively relevant circumstance'—will plausibly be minimally required to establish some degree of cooperativeness. But, he argues, "the mere presence of intentions that we J (by way of meshing subplans) need not by itself ensure satisfaction of this requirement. That is the lesson of the case of the unhelpful singers" (Bratman, 1992: pg. 105). To belabour the point, Bratman makes it clear this conclusion is not the result of chance but is, in fact, a *desirable* feature of any account of shared intention. He says, for example, that John Searle's insistence that collective intentionality implies cooperation is too strong. So, while

> "both jointly intentional action and SCA will involve somewhat similar webs of intentions concerning the joint activity … if a joint-act-type were to be loaded with respect to joint intentionality but still not, strictly speaking, cooperatively loaded, we would still not want to appeal to [cooperation] in specifying the intentions essential to SCA" (Bratman, 1992: pg. 104, fn 18).

However, Bratman seems to change his mind two decades later. In *Shared Agency* (2014), he presents a revision of the singers example (as referenced in Bratman, 2014: pg. 185, fn 38), this time focusing on two individuals going to New York City:

> "Suppose that I intend that we go to NYC in part by way of your intention that we go and meshing sub-plans, and in ways that cohere with the connection condition … This puts rational pressure on me both to track necessary means to this intended end and to filter further intentions accordingly … However, my cited intention that we go to NYC does not see your contribution to our joint activity as merely an expected precondition of our going to NYC … Your contribution to our going to NYC is, rather, a part of what I intend … This means that the demands of means-end coherence and of consistency apply to my intention in favor of, *inter alia*, your playing your role in our joint activity: I am under rational pressure in favor of necessary means to that, and in favor of filtering out options incompatible with that. I am under rational pressure in the direction of steps needed as means if you are to play your role in our joint activity. And I am under rational pressure not to take steps that would thwart your playing your role. This mean that, insofar as I am rational, I will be to some extent disposed to help you play your role in our going to NYC if my help were to be needed.
>
> Granted, I can intend our going, and so your role in our going, and still be willing to bear only a limited cost in helping you … But if I intend our going then I am under rational pressure to be willing to some extent to help you if need be. This is in part because I need to be set not to thwart you; and so I need to be set to help you at least to the extent of refraining from thwarting you. But, further, if I intend our going, and do not just intend to go given that, as I expect, you will go, I will be under rational pressure to be willing to some extent to provide some (perhaps limited) positive support for your role in our going. And I am under such rational pressure even if I expect that you will in fact not need such help" (Bratman, 2014: pg. 56–57).

Here we see Bratman raise the intentions-via-intentions (the settling-control mediation) requirement, the demands this places on intending versus expecting a partner's contributions and the rational pressures to intend that their intentions are effective. What before, in his account of SCA, was formerly reserved for shared cooperative activity in which cooperativeness was in effect 'added on top of' joint intention is now built directly into it. These are no longer 'nice to have' norms that improve the odds of successful shared activity; they are, instead, a direct consequence of rationality itself. This conceptual shift may not have been much of a problem except for the fact that, as I argued earlier, the interpersonal commitment now baked into shared intention is not credibly motivating. The problem then is

this: commitment to support one's partner wasn't necessary before, in SCA (1992), so we didn't need to unpack its origins in such detail; but now that it's necessary, in Shared Agency (2014), as a response to the Petersson-like objections we are faced with having to explain its emergence from individual practical and intention rationality and find it difficult to do so.

*

If this particular conceptualisation of commitment lacks credibility, then we have little insight into why agents won't be motivated to reconsider their intentions in the face of new options. In expanding the possibility of shared intention to situations involving imperfectly-aligned interests, we don't yet have grounds to justify if and why social norms of commitment will emerge. One response Bratman might make, however, is to be content to limit his account to encompass only contexts where interests are aligned. He says, for example:

> "Finally, consider competitive activities. We might be engaged in a shared intentional activity of playing chess together, even though—since we are in competition—neither intends that there be mesh [sic] of sub-plans all the way down. This limits the extent to which what we do together is a cooperative activity. It does not, however, block a shared intention to play chess together, and it allows that our chess playing is a shared intentional activity. So there will be shared intentions that involve intentions on the part of each that only favor mesh in sub-plans down to a certain level. Nevertheless, given our interest in sufficient conditions for modest sociality, I will focus on cases that involve intention-like commitments to mesh all the way down" (Bratman, 2014: pg.55–56).

It's possible, then, that Bratman might disagree with my interpretation of his notion of strategic interaction. From the examples and his explanations I have inferred that he means it to be shorthand for acting with purely instrumental, or self-interested reasons. This allows us to stratify this type of behaviour (Table Rows) from interest alignment (Table Columns), so that one can behave strategically or not, when interests are or are not aligned; hence our table with four scenarios. Bratman might, though, be using the term strategic interaction as shorthand for non-aligned interests. In this case, it's only possible to have shared intention if interests are aligned; if not, then he'd be happy to preclude the possibility of shared intention.

This would however involve a substantial pre-commitment to a particular view of shared intention that I am not sure Bratman would want to make. First and foremost, it would

severely limit his claim to provide a truly general account of shared intention. Beyond this, choosing to limit his account to situations with aligned interests runs counter to the ethos that guides Bratman's work. An important rationale for his view of the norms and principles involved in planning is that these provide a platform for agents to bargain and negotiate about who does what, when, how, crucially in a temporally extended way, such that plans and sub-plans are expected to be filled in over time. This is a direct consequence of being forced to deal with unpredictability in the future environment, which is at odds with a presumption that we should ignore the fact that interests change and may no longer align.

This is not to say there isn't a unique form of social normativity essential to shared intention which aligns with Bratman's conceptualisation of it, but only that we haven't got the explanation of its source quite right. Getting this explanation in is proving to be quite complex but the following are some brief examples of possible avenues to explore. We could say, for example, that agents' supportive behaviours just *are* driven by non-strategic preferences, that we just *are* in [c] and not [a]. But very quickly we realise this is not enough, as it either begs the question or leads in a circle: if we select specific contexts because they are shared, how do we explain the criteria for sharedness on which they were selected, if not via appeal to the sharing of intentions? Another idea is to focus on ways in which [a] and [c] could differ in specific characteristics other than those we've used as indicators of commitment; that is, abandon the use of interpersonal commitment as a distinguishing feature of shared intention. This could include unique differences in the psychology (e.g., Searle, Tuomela) or phenomenological experience (e.g., Dan Zahavi) of those involved, to then be used as alternative markers to differentiate strategic from non-strategic behaviour. Alternatively, if we bind ourselves to Bratman's continuity thesis, then perhaps differences in the cognitive processes involved might explain that, though on the surface the supporting behaviours look the same, agents in [c] are making decisions based on a collective optimisation problem, about who is best placed to do what and when, versus in [a] where they're optimising based purely on their own outcomes (the team reasoning in Gold & Sugden, 2007). Or we might take it that agents in [a] and [c] have different reasons for their behaviour. Relational models theory, for example, suggests that people can view acts of giving and reciprocity through either communal or transactional modes of engagement (for the original proposal, see Fiske, 1992; for a review, see Haslam & Fiske, 2005). These different modes are, perhaps, one way of seeing different reasons agents may have for the same kind of supporting behaviour that manifests in various types of social interaction.

These ideas are all familiar from a large body of work on the economics and psychology of cooperation and coordination. The next chapter engages with some of these in more detail but, for now, there are three points to note if we're to rely on any to explain interpersonal commitment. First, their grounds for why we get commitment in social activity is different from Bratman's. They provide alternative reasons for thinking why agents might be in [c] and not [a] and provide different reasons as to why the commitment that ensues is credible. Second, they are themselves open to challenge and if used to characterise shared intention, then the concepts they propose need work done to explain how they avoid being circular. Finally, it's not clear we can simply add something like this to an account like Bratman's while still meeting the continuity thesis. In particular, if we're introducing a new explanation of commitment which draws on an essential, unique or *sui generis* kind of sociality then (provided its passed the circularity test) there's a strong chance the continuity thesis won't hold.

*

Limiting Bratman's project to contexts involving aligned interests (Col. 1) impacts his claim to be providing a paradigmatic account of shared intention. And even if we constrained his account in this way, it would still leave us with the job of explaining why, when interests are aligned, [c] and not [a] would hold. Still, if we're happy to stick to these situations (Col. 1), then identifying the source of interpersonal commitment might not matter, as the kinds of supportive dispositions and behaviours being sought are, in a sense, guaranteed either way.

We might, though, be worried by an inability to distinguish whether we're in [a] or [c] for another reason. For if it's essential to an account of shared intention that agents' attitudes are non-strategic, then it's vital that we know what is driving supportive dispositions and behaviours which may be on show when analysing whether a particular situation counts as an example of the phenomenon. This cuts close to the bone of Bratman's account. Recall that he is proposing a non-tokenistic/non-instrumental account of shared intention by clarifying that, in his view, agents sharing intentions are bound together by a glue that's irreducible to common knowledge and exhibit a reciprocity grounded in intentional interconnection. Crucially, this social glue and mutual instrumentalism are evidenced by a commitment to both one's own and one's partner's participation in and contribution to the joint activity. This commitment manifests in being supportive of one's partners and attempting to align, organise and coordinate intentions and actions. However, because these forms of interpersonal support

are consistent with both strategic and non-strategic behaviour, we cannot use this feature as an identifying mark of the social glue Bratman has in mind. His version of social commitment thus plays no role in supporting ongoing shared activity or explaining it. And if we can't establish how exactly to characterise this social glue using the resources available, then we no longer have an adequate explanation for what defines shared intention in the strong sense required. Ultimately, this is not to say it's impossible to reconcile Bratman's account—and, possibly, other reductive accounts facing the same challenge—with these concerns, only that they require further research in order to be confident that we can.

6.3 Mutual obligations and joint commitment

The discussion in the previous section is relevant not only to a very specific feature of Bratman's conditions—using interpersonal commitment to identify his social glue—but also to the tenor of his proposed approach to studying shared intention in general. Bratman is well known for what he describes as 'an attractive division of philosophical labour' between, on the one hand, characterising the norms essential to a basic thesis of shared intention and, on the other, describing additional social norms of moral obligation which often, though importantly not always, support persistence interdependence of intentions. What he's after is to show how shared intention provides normative guidance in practical reasoning (i.e., how it performs its functional role in guiding collective decisions on how and what to do) through pressures generated purely by norms of rational agency with no additional, alternative normative pressures sourced in moral or ethical questions about how one should treat others. Persistence interdependence, recall, is present when each continues to intend in favour of the joint activity only if their partners continue to do so as well. It is an element of Bratman's account which appeals to a

> "generic interrelation captured by our abstract characterization of such interdependence … which is motivated in part by reflection on the settle condition. And the basic thesis can appeal to this generic condition of interdependence without making an essential appeal to the special case in which this interdependence is based on mutual obligations" (Bratman, 2014: pg. 112).

He does note that persistence interdependence "can be realized by different kinds of interpersonal structures", involving instances of feasibility-based, desirability-based and obligation-based interdependence (Bratman, 2014: pg. 112). These forms may, but need not

overlap, meaning that "there can be desirability-based and/or feasibility-based interdependence that does not depend on the presence of relevant mutual obligations" (Bratman, 2014: pg. 112). A party may, for instance, caveat their expression of intention with the right to change their mind, or may recognise the presence of moral obligations but secretly intend not to comply with them. Mutual obligations are, on their own, thus seen as insufficient to ensure shared intention, and persistence interdependence does not therefore depend solely on the recognition of mutual obligations. Bratman acknowledges, though, that for adults engaged in 'temporally extended interaction', obligation-based interdependence is both common and useful. He says:

> "After all, in many such interactions each of the participants will have, in effect, assured the other that she will intend the joint activity, and/or intentionally encouraged the other to rely on this and/or intentionally reinforced the other's reliance on this. Though the details are a complex issue in moral theory, it seems that such forms of assurance and/or intentionally induced or reinforced reliance will frequently issue in moral obligations of each to each to continue so to intend. And in such cases, the participants' recognition of these mutual moral obligations will frequently help explain why there is persistence interdependence. It is because each recognizes these mutual obligations, in a context of common knowledge, that each is set, other things equal, to retain her intention so long as the other does. And so the resulting interdependence will be obligation-based (though it may also be desirability-based and/or feasibility-based)" (Bratman, 2014: pg. 110).

This view has the added advantage of avoiding tricky conversations about when exactly certain interactions do, in fact, ground social obligations. Bratman sees these as the purview of moral philosophy and not necessary for an account of intention and planning to answer, saying that

> "we can leave it to substantive moral theory to articulate and defend detailed principles concerning such moral obligations … We can simply note that sometimes a temporally extended aetiology of persistence interdependence induces mutual moral obligations whose recognition by the participants is part of the explanation of that persistence interdependence" (Bratman, 2014: pg. 111).

He says that there are several theoretical advantages to taking this approach:

> "First, it allows for an attractive division of philosophical labor: we can defend the basic thesis while leaving for further normative inquiry the precise principles of relevant moral obligation. Second, it acknowledges the important role that morality can sometimes play in our modest sociality without making obligations essential to modest sociality … The third advantage is that this understanding allows us to retain a model of modest sociality that is broadly continuous with the planning theory of individual agency while making room for the possible role of distinctive interpersonal norms of moral obligation" (Bratman, 2014: pg. 113).

To reiterate: it's crucial for Bratman's account and the continuity thesis on which it's founded that his thesis makes no appeal to the role of such mutual obligations, moral or otherwise.

*

That mutual obligations should not be seen as essential to shared intention is a feature of Bratman's account often used to draw a sharp divide between his, and accounts which take a similar view, and those who see an irreducible social obligation as required. However, the lack of a substantive account of social commitment undermines the idea that mutual obligations are not essential. One thing the investigation into commitment credibility has shown is that the forms of social rationality Bratman has in mind—including interpersonal commitment—cannot be assumed to emerge from the basic underlying norms of practical and intention rationality when interests are not aligned. This means that he can only partial out moral factors when explaining joint action by constraining his perspective to cases where agents' interests are aligned. But there's a catch in doing this, besides the impact on account generalisability: the notion of interpersonal commitment now no longer seems to be needed, as it's already in agents' interests to be disposed or take steps towards supporting their partners; interpersonal commitment may be present, but it no longer does any explanatory work. This paradox is highlighted only because we're possibly expanding the scope of shared intention to include cases with motivational uncertainty, in which there are now reasons to be uncertain about a willingness to contribute.

To see this, note how Bratman's division of philosophical labour implies that mutual obligations are tools we can *choose* to use, but don't have to: "we can nevertheless see how, once such forms of interpersonal moral obligation are available to us, they can be put to work in the creation of a form of the persistence interdependence that is an element of our modest sociality" (Bratman, 2014: pg. 113). It is in reflecting on the settle condition and norms of

rationality that support it that allow him to rely on these norms. Take for example the defeasible statement made upfront: 'no obligations'. Bratman says that "even if this caveat on the part of each blocks relevant obligations, it need not block relevant predictability of each to each; and so it need not block the kind of mutual rational support that lies behind persistence interdependence" (2014: pg. 111). So obligations per se don't need to ground the settle condition. This is missing the point of these obligations, though, for if agents are in a position to predict how others will act and so settle matters, then it must be the case that their interests are, to some extent, aligned. But if interests are aligned, then there is no useful role for mutual obligations. They are not useful tools in this particular situation, the caveat 'no obligations' is superfluous and they are not needed to ground settling anyway. To imply so would be insincere treatment of their role in joint action or a lack of theoretical parsimony.

To give obligations their due, we have to ask *why* and *when* they are useful—with shared intention's functional role being our normative standard. For example, when there's common knowledge of intentions and beliefs, as is commonly assumed (see Chapter 2), then as a consequence there's no special role for norms of obligation in enabling us to rely on other people to make predictions about how they will act. Any question about whether one's partner is going to act is redundant: that question has already been settled by the psychological states of those involved via common knowledge. To be sure, it's important that the structure of intentional knowledge does imply that certain norms will be in force, but that appears as a consequence of common knowledge; it's not an essential part of what's providing support for agents' actions, or rationalising their actions and so on. This is because, if there is common knowledge in the situation, in the way Bratman envisages, then there's no room for any concern related to how likely it is that my joint action partner will act with me.

Moreover, Bratman suggests that interpersonal obligations—as a tool for reducing motivational uncertainty and stabilising expectations—are frequently used, but he doesn't provide a systematic overview of the conditions under which this occurs. This is perhaps outside the scope of his account, but as I've hoped to show in this chapter, his proposed division of philosophical labour only works by imposing additional, quite rigid constraints on what we take to be shared intention; that is, limiting us to cases involving aligned interests.

*

Margaret Gilbert's approach to developing an account of social commitment contrasts strongly with Bratman's ideas about a division of philosophical labour. As a popular account

that places the role of interpersonal obligation in supporting shared intention at it's centre, it might provide an alternative option for the problems still faced in explaining if and how it's possible for there to be shared intention in BEACH. In Gilbert's view, it is the presence of certain forms of obligation between agents which constitute many of the social groups of which they are part and which bind their members together. This is true for groups performing shared activities as well. Mutual obligations, she thinks, are essential to explaining how shared intention performs its, again, functional role of guiding collectively intentional action, and not just that they are a useful tool to use.

One interesting aspect of Gilbert's view is that the structure of obligations she proposes is partly explained in reference to the process required for members wanting to dissolve them. Recall that Bratman's explanation of interpersonal commitment is as part of a social rationality emerging from individual rationality plus, as argued in this chapter, a plausible implicit assumption of aligned interests. Gilbert, on the other hand, explains the normative pressures associated with the joint commitments she proposes primarily through reflection on instances where one or more parties might consider abandoning their commitment; that is, situations in which individuals' interests are patently no longer aligned. Though neither author is explicit about this being a particular feature of their approach, it leads us to think that concerns about commitment credibility encountered earlier might be different for an account with a specific focus on commitment dissolution.

According to Gilbert, people share an intention when they are "jointly committed to intend as a body to do J" (Gilbert, 2009). Joint commitments emerge when two or more people express a personal readiness to jointly, with others, commit all of them to this action, *J*. Gilbert's argument that shared intention involves joint commitments can be called an argument from rights and obligations (Michael & Pacherie, 2015). The presence of obligations is what gives joint commitments their 'normative force' in and of themselves, creating obligations for their authors and corresponding entitlements for their recipients. In her rights-based approach, an agent is only relinquished from an obligation to carry out an action she committed to when the recipient of the commitment allows this, releasing her from her commitment. More specifically, individuals are jointly committed and share an intention when three 'criteria of adequacy' are met and which Gilbert claims any reasonable, adequate account of shared intention must satisfy. These are:

- *Disjunction criterion*. It is not necessary that for every shared intention the individual parties involved have correlative personal intentions (i.e., personal intentions aimed at the satisfaction of given shared intention).

- *Concurrence criterion*. Absent special understandings, concurrence of all parties is required to change or rescind a shared intention, or to release a given party from participation.

- *Obligation criterion*. Parties to a shared intention are obligated to each act as appropriate to the shared intention in conjunction with the rest.

Joint commitments involving two or more people can therefore only be created and rescinded by those individuals together. The origin of this idea begins with her noting 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: pg. 180). (This echoes our earlier discussion on settling and control being the sole domain of the intender). While one may rescind one's own personal commitment by simply changing one's mind, because joint commitments are not built up solely 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. So even if one individual no longer intends the collective activity, the joint commitment doesn't fall away.

The subject of the joint commitment is therefore not the individual. Rather, Gilbert views all those who have jointly committed in the way described as the *plural subject* of the joint commitment. Unlike Bratman, this is not to be seen as an interwoven complex of individual commitments, but rather the plural analogy to the individual case in which personal intentions entail personal commitments to act a certain way. This derives from Gilbert's "observations on the way people think and talk about shared intention in everyday life" (Gilbert, 2009: pg. 171), in particular that we usually ascribe shared intention to two or more people at a time and not to each individual alone.

One implication of this is that the normative force of the obligations she describes do not have their root in norms present in solo intentional activity, in which the subject of the intention is the individual. Certainly, Gilbert assumes that agents regard themselves and their partners as rational, but she does not explicate, as Bratman does, social parallels of individual forms of rationality which, if available, could be ascribed to the plural subject to guide 'its' actions. At the same time, Gilbert sees the normative force exerted by mutual obligations as vital to causing jointly committed agents to act as they committed to doing.

It's therefore important for Gilbert's account to have explanatory power that she figures out how to properly locate the social normativity underpinning joint commitment. This must originate in her three criteria, and reflection on them sees the obligation criterion as the one doing the work. We thus need to know what is required for it to be satisfied, as doing so will provide the reasons why these commitments credibly motivate action. One way to do this would be to "list one or more other conditions from which the pertinent obligations of the parties did not follow … and then explicitly posit, in addition, the existence of such obligations" (Gilbert, 2009: pg. 177). However, Gilbert says, this approach is unsatisfactory as we would not yet have explained the grounds for such obligations. Indeed, tacking on additional criteria from which mutual obligations flow only pushes the question back a level (perhaps a veiled reference to Bratman's use of Thomas Scanlon's principles as sources of obligation).

Apart from her intuition on the matter, it's not clear why Gilbert is so motivated to avoid taking this route. If we can locate the *original* source of the obligation then what's the problem? One (unexpressed) reason might be that delegating normative authority to some other, external set of norms or standards means the core concepts she's using to explain shared intention are not giving us anything new, and, consequently, that her account is not explaining what's particularly unique about shared intentional activity. As Roth (2018) puts it: "A concern … is whether joint commitment provides anything like a philosophical account or explanation of mutual obligations, or whether it merely redescribes them". This is perhaps also the source of Gilbert's aversion to seeing mutual obligations in shared intention as *moral* obligations in any way, something I'll discuss shortly.

Continuing on, what Gilbert needs is to ensure there's an intrinsic form of social normativity that gets reflected in her account, such that interpersonal obligations of the type she has in mind are seen as *sui generis* to shared intention. Simply stating this as the case

begs the question, though, so Gilbert's solution is to make the relevance of the obligation criterion endogenous to the other two criteria she proposes,

> "such that the conditions [the obligation criterion] explicitly posits—without explicitly positing the obligations—are such that *it follows* from [the disjunction and concurrence conditions] that the parties have these obligations. That way, the obligations would be explicable on the basis of other conditions" (Gilbert, 2009: pg. 177, author's emphasis).

How best to characterise mutual obligations can consequently be understood by reflecting on the two criteria. Of particular interest is the concurrence criterion which, in Gilbert's view, doesn't face any questions in explaining its own origin and why it's satisfied. This is because Gilbert claims that it's the shared intention *itself* which provides the answers to them:

> "I take it as read that the account should be such that the parties to the shared intention will understand that their concurrence is required as stated, and that, in addition, they will understand that this is a matter of *what shared intention is*" (Gilbert, 2009: pg. 173, author's emphasis).

Hence agents can justify the need for concurrence by simply referring to the shared intention as such. If you and I plan on going for a long walk to the end of the beach and I suddenly stop halfway, a natural justification for you saying to me "You can't just stop here" is by appealing to the shared activity itself, rather than for you to come up with some ethical reason for why I should continue.

*

Mutual obligations are thus partly grounded in the concurrence criterion—but where do claims and rights fit in, which is how Gilbert's account was described at the beginning? A clue comes from the observation that in the cases of shared intention Gilbert takes as paradigmatic, it's not simply anyone's concurrence which must be sought; it is specifically the concurrence of one's partners to the joint commitment. This means Gilbert is talking about obligations of a particular type in which individuals are obligated *to* other people to act in ways they've previously committed *to them* to do. These are obligations directed towards specific agents with whom one is jointly committed and are neither obligations in general (e.g., to give away a proportion of one's salary) nor alternative kinds of obligations one may have to specific others (e.g., to look after one's mother in general). There's a lot more that can

be said about this but it's enough for now to say that Gilbert follows this line of thought, to propose that theories of claim-rights can explain this 'nature of directed duties' she sees as characteristic of shared intention. As per Gilbert, rights theorists generally explain a directed duty as follows:

> *A's right against B to an action of B's is said to be B's obligation (or duty) to A to perform the action.*

Gilbert argues that obligations in the context of shared intention should be similarly interpreted, that is, as claim rights in which parties owe each other action appropriate to the shared intention. We implicitly recognise this, she says, in the special standing we afford those jointly committed the right to issue a rebuke to one another in case of failure to perform an appropriately agreed upon action at the appropriate time where concurrence has not been requested ("You can't just stop here"). It is thus Gilbert's intuition that jointly committed agents are claim-right holders to the actions of their partners to which they have a right: "each "owns", in some intuitive sense" (Gilbert, 2009: pg. 176) their partners' future actions. These rights have their source in the public and voluntary generation of their joint commitment. Agents therefore have obligations to perform their parts—that is, to meet their commitments—because others have valid claims over their future actions. Conversely, agents are relinquished from obligations to perform their part, to which they have committed, only when the recipient of the commitment, or all parties to the joint commitment, allows this.

*

There are several positive features of Gilbert's account. One is the notion just described that directed obligations lie at the heart of shared intention. The idea that one's interaction partners have a special standing in this respect is indeed a useful component of an account of commitment in joint action, as I'll pick up in the next chapter. On the surface, Gilbert's account also looks like it provides a pleasing counterpoint to the issues we saw in Bratman's. First, the concern of tokenistic shared intention is addressed, as if there's shared intention on her view then it involves something distinctively shared (the joint commitment) compared to solo intentional action. Second, the concern about making room for purely selfishly-motivated interaction is partly addressed, as joint commitments guide agents away from doing what's wholly in their self interest, such as abandoning the joint activity when it so suits. Third, her account seems to have greater generalisability as making it work doesn't require limiting it to specific cases where, for example, interests are aligned or agents'

cooperativeness is assumed. Finally, many researchers typically find it difficult to think of joint action as devoid of any kind of social obligation, so her account aligns with this general intuition. More than that though, Gilbert's intuition of how we generally experience these obligations also feels familiar. It's certainly possible that at times there are ethical issues that come with not meeting our commitments, but it also seems reasonable that letting others down in many small-scale interactions doesn't invoke any heavy ethical or moral questions. It could easily be that my stopping on the beach is, for example, not in any way disrespectful of your moral character or standing as an equal—and your rebuke would not reflect this (if it was, perhaps you would do more than just issue a mild rebuke!).

These features make Gilbert's joint commitment look like a plausible candidate for overcoming the problems in explaining shared intention under motivational uncertainty with which we've been engaged. Specifically, joint commitments seem to provide a reason why agents can jointly settle matters when they have reasons to be uncertain about their intentions. If agents expect their partners to remain committed and meet their social obligation to contribute, then they can rely on them to do so. Moreover, they know that their partners expect the same, and, given common knowledge of the joint commitment, each knows that each is committed in the same way, knows the other knows they are committed in the same way, and so on. So even if there is uncertainty about intentions, if the joint commitment is common knowledge—which it is, by definition—then this provides a counterpoint to worries about one's partner being attracted by alternative options and abandoning the joint activity.

We also don't have the worry that we had when dealing with Bratman that uncertainty about intentions must surely mean uncertainty about commitments, which would make it difficult to see how commitments can settle matters without intentions doing so. Gilbert's very clear that intentions can come apart, as in her example of one person in a group currently climbing a hill who decides to himself that he no longer intends to walk to the top. Because it is not merely a combination of personal intentions, but requires concurrence to be revoked, Gilbert sees the joint commitment persisting despite that climber's change in attitude. Being able to separate intentions and commitments therefore opens up the possibility of relying on commitments to settle matters even though there may be uncertainty about intentions.

At the same time, the proposal I'm making is not exactly the same as Gilbert's, as I take it that we can't see intentions and social commitments as two very separate things either, which is an important driver behind her disjunction criterion. If they were, then it wouldn't

make it reasonable to think of commitments as settling the matter. After all, settling would seem to be divorced from anything we've covered so far, and have nothing to do with intentions at all. There are, furthermore, additional issues concerning Gilbert's overall view which means it's unable immediately to solve the problem of motivational uncertainty—at least not without making additional assumptions that go beyond the scope of her account. The first has to do with Gilbert's proposal that the subject of the joint commitment is a plurality of the agents involved and not each individual on their own. About this plural subject, in earlier work she says that

> "I have argued that those out on a walk together constitute the plural subject of a particular goal, roughly, the goal that they walk along side by side for a certain roughly specified period. Let us say that a given set of people have a … 'shared' goal when they are the plural subject of a goal.
>
> (...)
>
> [I]n my view, human social groups are plural subjects. That is, in order to form a social group, it is both logically necessary and logically sufficient that a set of human beings constitute a plural subject. Clearly this is a thesis about concept, namely our intuitive concept of a social group" (Gilbert, 2006: pg. 7–9).

While Gilbert's therefore clear that she's making a conceptual and not metaphysical claim here—that is, her proposal of the plural subject should not be interpreted as an agent in its own right—she's also not especially clear how we should otherwise interpret this plural subject of the shared intention. She provides some clues in her work, saying for example:

> "I have argued that going for a walk together with another person involves participating in an activity of a special kind, one whose goal is the goal of a plural subject, as opposed to the shared personal goal of the participants". (Gilbert, 1990: pg. 9)

and

> "Given the meaning of my technical phrase ''plural subject'' I dub the account of shared intention I am discussing the plural subject account of shared intention. Though its ontological commitments are, I believe, unexceptionable, it does imply

> that a shared intention is not constructed out of singularist-intentions, contrary to the assumption of many" (Gilbert, 2009: pg. 182).

This shows that what we get from Gilbert about the plural subject is really a set of direct descriptions about what it's *not* (especially as compared to other reductive accounts), rather than a direct description of what it is.

Still, not having an exact characterisation of the plural subject may not be a deal breaker for thinking about motivation and uncertainty, as Gilbert still makes room in her account for the role of individual intentions:

> "In addition to the individual commitments derived from the joint commitment at the heart of a given shared intention, on the plural subject account, the parties are likely to develop a variety of concordant *personal* intentions. These will arise under the guidance, so to speak, of the foundational joint commitment and the joint commitments involved in any shared sub-plans, along with individual commitments derived from these.
>
> (...)
>
> Though the plural subject account does not itself posit any particular personal intentions, then, one can predict that shared intentions on that account *will be accompanied by a variety of meshing personal intentions of the parties*, when those parties act appropriately in light of their shared intention and any shared sub-plans they have consequently developed" (Gilbert, 2009: pg. 184–185, author's emphasis).

This means we still have individual intentions and beliefs which develop and guide individual behaviour and on which we can focus, despite not having a clear picture of the plural subject.

Another concern with Gilbert's account is that it looks circular, as she seems to appeal to a collective willingness or commitment to form a joint commitment in the first place. I've already discussed issues of circularity in this thesis, but suffice it to say that Gilbert's account is not immune to the problem of figuring out the 'upstream' source of the sharedness which avoids using shared intention as an explainer. And similar ideas and solutions as before—for example, collective identification, awareness, belief, et cetera—might very well provide the foundation for Gilbert's joint commitment, so I don't see this as a deal breaker for our question of motivational uncertainty either.

However, even given the workarounds of the above concerns, Gilbert's account of joint commitment doesn't provide a credible answer as to why individuals should rationally rely on their partners when they're uncertain about their intentions. This is because Gilbert neither goes further in discussing the extent to which the *sui generis* type of obligations are expected to motivate individuals to meet them (in the face of competing considerations, which have rational demands of their own to be met), nor says anything about what exactly individuals should take into account when requesting a partner's concurrence. Indeed, a strict reading of her account suggests that one's partner may not be able to refuse such concurrence when requested. More charitably, she has perhaps just taken this process for granted, focusing on the normativity of joint action rather than its practicalities (Chennells & Michael, 2022). What Gilbert has in mind as the type of obligation therefore looks very much up to intuition, as is, more importantly, how it's supposed to provide normative guidance.

*

Perhaps, however, we can make more effort to augment Gilbert's account to give her obligations 'bite' while remaining confined to the minimal conditions she proposes. Whether this boosts credibility remains to be seen, especially as substantiating her account looks like it will contradict what Gilbert is after in the first place—additions such as allowing for a role for social emotions likely undermine her claim that it's her three criteria per se which do the work of grounding shared intention. Still, let's dig into Gilbert's claim that obligations essential to joint commitment are non-moral in nature. She's adamant, recall, about a *sui generis* grounding for obligations on the shared intention itself. That these obligations are not to be confused with moral obligations, whose inclusion would require adding criteria to Gilbert's minimal account, is likely one reason for the unanswered questions about how they motivate behaviour. We lack familiar benchmarks from ethical discussions, such as when we are allowed to let others down and break promises, which we might otherwise turn to for answers. Because what must satisfy the obligation criterion must flow from reflection purely on the disjunction and concurrence criteria, recall that this leads Gilbert down a path in which obligations are the result of and response to claim-rights an agent's partners have over their action performance, with only these agents having a special standing to make this demand. Agents owe the future performance of an action to their partners. She says that,

> "as far as the theory of shared intention goes, I would argue that the interpretation we need is in terms of owing, an interpretation given by two distinguished rights

theorists, H. L. A. Hart and Joel Feinberg. In this construal, the parties owe each other action appropriate to the shared intention.

(...)

Feinberg refers at one point to a right-holder's demanding what he has a right to as his. This implies that if I owe someone a certain action, in the sense of ''owe'' in question here, he already in some intuitive sense owns that action … In what sense can one own the future action of another person?

(...)

I take from Feinberg the point that to be in a position to demand something from someone is for it already to be in some intuitive sense one's own. That is because demanding in the relevant sense is demanding as one's own. This, Feinberg implies, is something any claim-right holder can do with respect to an action to which he has a right. This suggests that there is an important and closely linked family of concepts here … [which] can be displayed as follows: one who has a right to someone's future action already owns that action in some intuitive sense of ''own''. Until the action is performed he is owed that action by the person concerned, thus being in a position to demand it of him prior to its being performed and to rebuke him if it is not performed. If it is performed, it has finally come into the possession of the right-holder, in the only way that it can.

This all suggests a way of interpreting the obligation criterion that fits the observable facts about shared intention and offers a plausible interpretation of them. Consider again the case of Rom and Queenie. Rom both rebukes Queenie (albeit mildly) for going too slowly for the satisfaction of their shared intention, and demands that she speed up if she can. Queenie implicitly accepts his standing to issue such rebukes and demands when she says ''Sorry!'' In so doing she acknowledges, in effect, that at the time he spoke Rom had a right against her to actions in accordance with the shared intention; and that she owed him such actions, which he already in some sense owned. In other terms, she has the corresponding directed obligation to perform such actions. Intuitively the same goes, with appropriate changes, for the parties to any shared intention" (Gilbert, 2009: pg. 176–177).

Note that here Gilbert is characterising the *nature* of the obligation relation between jointly committed agents, but not telling us *why* agents own and have a special standing to claim action performance; that is, what ultimately justifies them being in this position. One obvious avenue to explore for answers is to dive into the work of the authors Gilbert quotes and on which she bases her ideas, notably Feinberg (1970) and, to a lesser extent, Hart (1955). What do they have to say about the grounds for claim-rights?

In Feinberg's (1970) *The Nature and Value of Rights*, he says that legal claims and rights are linked together in the following way (across various pages, my emphasis):

> "When a person has a legal claim-right to X, it must be the case (i) that he is at *liberty* in respect to X. i.e., that he has no duty to refrain from or relinquish X, and also (ii) that his *liberty* is the ground of other people's duties to grant him X or not to interfere with him in respect to X.
>
> (…)
>
> To have a right is to have a claim against someone whose recognition as valid is caused by *some set of governing rules or moral principles*.
>
> (…)
>
> To have a claim … is to have a case meriting consideration, that is, to have reasons or grounds that put one in a position to engage in performative and propositional claiming.
>
> (…)
>
> What then is the relation between a claim and a Right? … As we shall see, a right is a kind of claim, and a claim is "an assertion of right".
>
> (…)
>
> It is an important fact about rights (or claims), then, that they can be claimed only by those who have them.
>
> (…)
>
> Having rights, of course, makes claiming possible; but it is claiming that gives rights their special *moral* significance.

(…)

This is made possible by the fact that claiming is an elaborate sort o*f rule-governed activity*. A claim is that which is claimed, the object of the act of claiming.

(…)

… for it remains true that not all claims put forward as valid really are valid; and only the valid ones can be acknowledged as rights.

(…)

“Validity," as I understand it, is justification of a peculiar and narrow kind, namely justification within a *system of rules*. A man has a legal right when the official recognition of his claim (as valid) is called for *by the governing rules*. This definition, of course, hardly applies to moral rights, but that is not because the genus of which moral rights are a species is something other than claims. A man has a moral right when he has a claim the recognition of which is called for—not (necessarily) by legal rules—but by moral principles, or the principles of an enlightened conscience.”

We can draw three insights from these brief quotes. First, the kind of claim-rights with which Feinberg is concerned certainly seem to have a special moral significance despite Gilbert’s claim otherwise. Second, Feinberg focuses closely on legal rights. Though he claims parallels with moral rights and claims, he doesn’t elaborate on this, and whether his ideas map to other domains is left unanswered by him. As one commentator notes,

> “although the title of Feinberg's paper suggests that he is talking about rights in the generic sense, the only species of rights he discusses at length are legal rights. Are there nonlegal performances of making a claim parallel to the activity of making a legal claim? If so, in what ways are these nonlegal performances similar to and different from their legal counterpart? If not, how are we to interpret the traditional language of moral rights and human rights” (commentary by Carl Wellman, in Feinberg, 1970: pg. 258).

Third, while Feinberg gives us a good idea of the nature of the relation between the rights claimant and the ower, he doesn’t provide much detail about what grounds the former’s special standing to make a claim in the first place. As a second commentator notes,

> "it should be pointed out that while Feinberg tells us a great deal about the "nature" of rights, he doesn't really tell us very much about their value. He says that the claims involved in rights are to be validated by reference to a "set of governing rules", but he doesn't tell us which of the infinite variety of conceivable governing rules are the right ones. Until we know this, though, I don't think we should rest satisfied" (commentary by Jan Narveson, in Feinberg, 1970: pg. 260).

Interestingly, this is exactly the problem we faced with Gilbert's account and why we turned to Feinberg in the first place! We've gone up a level to find the source for the satisfaction of the obligation criterion and encountered, in the ideas to which Gilbert herself appeals, the same problem.

I have not focused on Gilbert's other source, given her greater reliance on Feinberg, but in the paper she references it's even more obvious that Hart (1955) is making no effort to separate moral and legal rights at all. In fact, at the outset he says: "I shall advance the thesis that if there are any moral rights at all, it follows that there is at least one natural right, the equal right of all men to be free."[18] And highly relevant to thinking about joint commitment:

> "The most obvious cases of special rights are those that arise from promises. By promising to do or not to do something, we voluntarily incur obligations and create or confer rights on those to whom we promise; we alter the existing moral independence of the parties' freedom of choice in relation to some action and create a new moral relationship between them, so that it becomes morally legitimate for the person to whom the promise is given to determine how the promisor shall act.
>
> (…)
>
> The simplest case of promising illustrates two points characteristic of all special rights: (i) the right and obligation arise not because the promised action has itself any particular moral quality, but just because of the voluntary transaction between the parties; (2) the identity of the parties concerned is vital—only *this* person (the promisee) has the moral justification for determining how the promisor shall act. It is

[18] Feinberg (1970) actually suggests moral grounds not far off this. He says: "The activity of claiming, finally, as much as any other thing, makes for self-respect and respect for others, gives a sense to the notion of personal dignity, and distinguishes this otherwise morally flawed world from the even worse world of Nowheresville." And: "Having rights enables us to "stand up like men," to look others in the eye, and to feel in some fundamental way the equal of anyone. To think of oneself as the holder of rights is not to be unduly but properly proud, to have that minimal self-respect that is necessary to be worthy of the love and esteem of others."

his right; only in relation to him is the promisor's freedom of choice diminished, so that if he chooses to release the promisor no one else can complain" (Hart, 1955: pg. 183–184, author's emphasis).

Drawing solely on these works by Feinberg and Hart makes it difficult to separate obligations from other moral questions, despite Gilbert's claim to do. In one sense, her approach to thinking about the essential normativity in shared intention is not too dissimilar from Bratman's division of philosophical labour. She tries to separate moral reasons agents have to meet their commitments out from some other guiding force or set of principles. The problem is, by doing so we don't have much idea what lies at the heart of the obligations Gilbert proposes, and turning to her original sources hasn't helped much either.

Turning back to the problem of uncertainty about intentions, this makes Gilbert's account difficult to rely on to explain the necessary basis for reliance and joint settling. Either Gilbert's account runs the risk of that which she feared, namely, that additional criteria are needed to satisfy the obligation criterion. We would have to ask questions about the extent to which these motivate agents to meet their obligations—but then it's not clear we need Gilbert's account at all. Or we could continue with her account's appeal to non-moral obligations but then have no reasons yet for why jointly committed agents should meet the obligations they have to their partners. We have no insight into what makes them credible.

6.4 Exploring trust as a grounds for settling

Both Bratman and Gilbert's accounts of interpersonal commitment have issues which make them risky to use as a conceptual basis for joint settling. Both struggle to explain individuals' motivation to meet their commitments to their partners, a problem that's particularly acute when assessed in light of the kind of motivational uncertainty that's present in contexts like BEACH. My analysis, however, has been limited to the potential for a particular kind of social relation in the form of social commitment to provide the basis for reliance, responsiveness and joint settling that shared intention requires.

If we keep the remit narrow and continue searching for a solution specifically in the nature of the social interaction between Mya and Iva, there may be an alternative kind of social relation that could help. One possibility is a role for trust and trusting relations, a field of research that's grown in recent popularity. A rough application of trust for our specific purposes is to use it to replace the background assumptions of ordinary predictability that

supported the arguments for joint settling in shared intention in Chapter 4. A basic premise is that if we can trust our partners to perform their part in a joint activity, we can rely on them to participate and so settle matters.

One attractive view for thinking about the role of trust directly in relation to my question of motivational uncertainty can be found in Berislav Marušić's (2015) book *Evidence and Agency*. In it, he addresses questions closely related to my project, concerning the possibility of intentional action and uncertainty about success, and in particular questions about control and self-determination. His primary focus is solo action, but he also briefly expands his ideas to the joint case, to some extent paralleling Bratman's individualistic methodology. This thus provides an interesting alternative which can either help us find solutions to how to address uncertainty about intentions, shed light on why the problems faced are not limited to Bratman's account, or both.

The core question Marušić asks in his book is how an agent can commit to an action she has good reason to believe she may not follow through with, when this following through is entirely up to her and not due to factors outside of her control. He starts off with an observation about the world that he takes as true: namely, that sometimes we commit to doing things we have good reason to believe we may not end up doing. The book is then dedicated to answering a dilemma he claims this situation gives rise to: on the one hand, it would seem that an agent cannot sincerely commit to doing something she doesn't believe she will actually do, while on the other it would appear irrational for her to believe she will do something she has reason to believe she will not do.

These questions sound familiar from Chapter 3's discussion which, drawing on Bratman, led us to adopting the Asymmetry Thesis. However, Marušić's proposal to resolve the dilemma he has posed is different, with his general argument going as follows. First, unlike Bratman, he argues that intentions and beliefs are not different kinds of attitudes, something he also picks up on in later work in which he and a co-author argue that

> "intentions are beliefs—beliefs that are held in light of, and made rational by, practical reasoning. To intend to do something is neither more nor less than to believe, on the basis of one's practical reasoning, that one will do it … we identify intentions with beliefs, rather than maintaining that beliefs are entailed by intentions or are components of them" (Marušić & Schwenkler, 2018).

Second, and following this, he argues that the kind of belief involved in intention should not be thought of as a credence about or subjective likelihood of success. Intention, he says, involves a 'practical belief' that one will do what one intends to do—mirroring Bratman's discussion of 'flat-out' belief. Third, notwithstanding the above, an agent can form the practical belief that she will *A* without knowing that her intention to *A* will be effective.

Crucially, what's required to reconcile these last two points is the agent's knowledge that *A*'s execution is up to her. So, Marušić's argument goes, an individual can commit—rationally and sincerely—to the future performance of an action she is uncertain she'll perform provided that she knows that that action is up to her to perform; that is, she knows that it's *within her control*. This can get tricky if we consider situations in which we might intend something but have doubts about whether we have the right level of commitment to see the project through. We might, for example, know that we're susceptible to attractive alternatives, or we might not know the amount of energy or resources required to be successful. Still, as Marušić argues, provided it is 'up to ourselves to control', there should not be a problem in accepting that we can nonetheless rationally and sincerely intend an activity despite these uncertainties.

*

The need for control and settling conditions in intention strongly echo the earlier discussions in Chapters 2–4 on how to theoretically explain the sense of joint settling and control that's essential to joint action. So if we take Marušić's view at face value, can it help us if we apply it to questions of motivational uncertainty in shared intentional activity—which, to note, is not the focus of Marušić's work? One apparent and relevant difference between settling matters in the individual and joint case is that in the latter agents are reliant upon their partners' action performance. When acting alone, the action's performance is entirely 'up to myself' to settle on and perform, whatever the likelihood of success. Conversely, when acting with others, the joint action's performance is not up to me to settle and perform on my own, as the outcome is a joint project whose success requires all of our contributions. This is the exact point made by Annette Baier about there being important limitations in the control each of us has over settling the actions of our partners. So any solution to uncertainty that places one's own control over the outcome at the centre runs up against the problem that, in the case of shared activity, it is specifically the fact that it's *not* up to oneself to settle the whole joint action that is at issue.

This doesn't yet mean that Marušić's ideas fail to apply in cases of shared activity, only that barring the invocation of a supra-personal or plural agent, we need to find some feature of the joint action which mediates this sense of control between agents, so that each agent knows the outcome is still up to them as a group to perform together. This is what I think Marušić has in mind when he briefly turns his attention to joint activities in which, because we cannot have the same knowledge of our partner's intentions as we do our own, there's an inherent element of uncertainty in what we can know about them. The question Marušić faces is how to accommodate this uncertainty within a framework that demands the kind of control over future action that's central to his solution to the problem of uncertainty in individual intentional action. In response, he argues that it's possible only if we introduce some interpersonal characteristic of the situation which mediates this control, and he says trust plays this role. For me to trust that you will perform your part is to know that you will perform it, and so to know that my control of our joint action is mediated by both my and your intentions, and vice versa. In Marušić's view, trust is therefore an essential feature of our joint action, one which supports you and I each having the kind of control—the sense in which uncertainty is fine provided it's up to us to perform when the time comes—required by shared intention.

This gives us a possible solution to the problem of uncertainty about intentions, if we see a role for trust to bridge the control gap between agents' intentions. If agents trust each other to perform their part, despite having reasons to be uncertain about whether or not they will (e.g., because of the presence of attractive alternatives), then we can reasonably conclude that they have grounds for settling matters about future activity and seeing those matters as up to them as a group to perform when the time comes.

There are, however, some important stumbling blocks to this straightforward application. First, the new, essential feature we are introducing into the account is by nature a *social* characteristic, if we think that trusting others differs in relevant ways to something parallel when acting alone—would it mean anything to think of this as trusting our future selves? This would violate any narrow continuity, conceptual or otherwise, between individual and shared intention. Second, even if we accept that trust is now necessary in contexts of motivational uncertainty, we still don't have adequate grounds for why agents *should* trust one another. It's not obvious yet why trust should emerge as part of the general social rationality that's part of shared intention. Hence, while we might have an idea of what

a trusting relationship is like, more work is needed to make a case for why trust is both non-circular and well-founded in a minimal account of shared intention.

Interestingly, this result is a close analogue to the issue of commitment credibility raised for Bratman's account. This brief turn to Marušić thus seems to broadly reinforce the findings from the previous sections in this chapter. Across Bratman, Gilbert and Marušić's accounts, we have some ideas for how to address uncertainty about intentions in a minimal account of shared intention. But their theoretical grounds appear to be too fragile for us to rely on when we introduce into the situation factors that generate motivational uncertainty.

6.5 Conclusion

Chapter 6 unpacks an important problem with Chapter 5's application of Bratman's notion of interpersonal commitment as a solution to the problem of settling under motivational uncertainty. For commitments to work, they must be credible: the receiver must be able to rely on or trust that the commitment provides normative guidance to the maker such that the maker is more likely to adhere to what they have committed to do than prior to committing. To understand if commitments are credible, we need to understand the source of this normativity. The problem we discover is that Bratman's view of a 'social rationality', which is said to emerge purely from norms of individual rationality, should plausibly *not* be expected to emerge in contexts with motivational uncertainty. This result is brought to the surface by focusing on situations with competing interests but it perhaps highlights a more general issue. As several authors have discussed, Bratman's account of shared intention doesn't do a convincing job of explaining why people would be motivated to meet their commitments (as he envisions them) to their partners. It's his ideas about cognitive and informational constraints which rationalise his original notion of commitment—and which underpin the norm of stability—but in extending this to the shared case he gives us no reason why agents would be motivated to meet their commitments rather than reconsider previously formed intentions when they face attractive alternatives, as in our case at hand.

The implications of this are two-fold. First, it means we can't use Bratman's commitment as a way to understand how there can be shared intention in situations like Mya and Iva's in which motivational uncertainty is present. While the discussion of social commitment has proved fruitful, touching on ideas that I'll draw on in the upcoming and final chapter, Bratman's specific version of it doesn't look like it's going to help us here. It may be

that adding additional requirements to his original conceptualisation can get us further but, as I've pointed out, it's not obvious how we can do this without the risk of either proposing something circular or straying away from his core continuity thesis. The second implication concerns Bratman's account more generally, beyond cases like in BEACH. Bratman uses interpersonal commitment—and the supportive behaviours and dispositions which flow out—as evidence of non-tokenistic and non-instrumental social interaction; that is, to defend a strong sense of jointness for his account of shared intention. But if these behaviours do not actually uniquely identify non-strategic behaviour, as Bratman argues they do, then this isn't available to him to use to justify why his account excludes the forms of strategic interaction he says it does, something several critics argue against. This finding also potentially undermines the 'division of philosophical labour' approach Bratman proposes for developing an account of shared intention.

Returning to the case at hand, we see that it is not only Bratman's account that struggles to explain shared intention in contexts with motivational uncertainty. In turning first to Margaret Gilbert, and her theory of shared intention and the joint commitment and mutual obligations she sees as essential to it, I've argued that though her account has several positive features it nonetheless likewise lacks a credible explanation for why individuals are motivated to meet their commitments. Part of the reason is because, like Bratman, Gilbert doesn't give us reasons why agents should be more motivated by their joint commitment than by selfish reasons they have for choosing an attractive alternative. And another part is because, in reflecting on her source material, it's not obvious that the obligations she thinks agents are motivated to meet really are separate from other moral and ethical reasons we have for meeting our commitments. De-linking them from these concerns and proposing instead an obligation unique and derivable to shared intention removes these reasons for acting which may have been more familiar to us. Replacing them with an intuition of why these *sui generis* obligations should motivate agents to meet commitments they have made doesn't provide the robust explanation for shared intention under uncertainty that we need.

Finally, briefly exploring Berislav Marušić's work on individual agency and committing to actions under uncertainty leads us, I argue, to the same conclusions as with Bratman. Looking at Marušić's extension of his ideas to cases of joint activity, in which he presents trust as a mediating factor, we are again left without a clear reason why agents who face competing interests will be motivated to keep their partner's trust or believe their partner will be motivated to keep their trust. Adding additional social reasons into the picture can

help to explain this, but again it looks like an approach that's circular or which oversteps a normative, metaphysical and conceptual continuity between individual and shared agency.

CHAPTER 7

The Sense of Commitment and Motivational Uncertainty[19]

Why social commitments motivate people to meet them is something several existing accounts of shared intention which give a role to such commitments find difficult to explain. This is highlighted when we explore the possibility of shared intention in contexts where agent interests are not aligned and where there's therefore substantial uncertainty about intentions. Various proposed interpersonal mechanisms for reducing motivational uncertainty—Bratman's interpersonal commitment, Gilbert's joint commitment, Marušić's trust—struggle to provide a credible explanation for why, in the presence of non-aligned interests, individuals can rely on their partners to be motivated to meet their commitments. This is not to say these proposals are without merit; indeed, the alternative account of commitment proposed in this chapter draws on several of their features. They all suffer, however, from two issues to do with motivation:

- *Proximal motivation issue.* They lack an explanation of how the norms they propose which guide committed agents to meet their commitments should be factored in the context of other norms and principles part of their practical reasoning.
- *Ultimate motivation issue.* They struggle to explain what the underlying normative source is of these social commitment-guiding norms; that is, where the value in meeting interpersonal commitments comes from.

I argued in the last chapter that the reason these issues tend to be overlooked is because of an almost exclusive focus in analysing shared intention against a backdrop of assumed cooperativity. This doesn't credibly identify the nature of the social glue that's supposed to set genuinely shared agency apart from other, weaker forms of joint action. Of course, this isn't a problem if we simply rely on intuition that agents are justified in relying on their partners for cooperation. This would suggest, though, that interpersonal commitment

[19] A first-author publication that draws extensively on and reworks selected text from this chapter can be found at: Chennells, M., Michael, J. (2022) *Breaking the right way: a closer look at how we dissolve commitments.* Phenomenology and the Cognitive Sciences. https://doi.org/10.1007/s11097-022-09805-x

isn't playing a meaningful role in supporting and explaining shared activity. At the very most, this approach says we're working on the assumption (though often without making it at all!) that people around us are cooperative, and that thinking about commitment is simply helpful for diagnosing either when things break down or which responses are appropriate to support joint action success. But if we think it's important to have a robust concept of social commitment—which we should if it's used to pick out shared intention—then this is a bigger issue: a commitment that doesn't motivate the committer to act committed precisely when this is needed is not a useful conceptualisation of commitment at all. The analysis in the last chapter suggests that a better test of commitment credibility—why it's present and what behaviour it promotes—can be found by shifting to situations with non-aligned interests. In doing so, we're able to tease apart motivations for acting committed because, put crudely, an agent who supports her partner when it's not in her interests signals better evidence of a credible, non-instrumental form of commitment to him than when it is in her interests.

While the majority of the previous chapters have been negative, identifying incompatibilities between motivational uncertainty and certain features of existing accounts of shared intention, this chapter concludes the thesis with a positive proposal. It presents an account of how people experience social commitments which doesn't suffer from the same credibility issues as those identified previously. This makes it better suited as a basis for shared intention in contexts involving motivational uncertainty—and so a possible route for resolving our problems in BEACH. The proposal is based on a particular understanding of what motivates people to keep commitments they have made; namely, to meet the reasonable expectations others have of them. It also takes a holistic view of commitment generation, emphasising implicit processes in addition to the explicit processes traditionally focused on.

7.1 What do we care about when we think of social commitment?

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 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 late night; his wet clothes damaged his phone) and therefore has no way of contacting Giulia?

In keeping with the theme of this thesis, I take it that in everyday life, we often find ourselves confronted with situations in which we would like to extricate ourselves from commitments that we've made—because our interests have changed, because we're tempted by some alternative that has arisen, because it is no longer feasible, because there is a conflicting commitment which we value more, et cetera. What factors or principles do we, or should we, appeal to in such situations, and why might we be motivated to keep commitments we have made? These questions target what was found wanting in existing theories of commitment in the last chapter. Moreover, as earlier chapters demonstrated, joint settling requires not only my own willingness to contribute but also expectations about your willingness to contribute. And these expectations are grounded in precisely these same questions regarding when and how you might prefer to dissolve your commitment to contribute to and play your role in our joint action, meaning that we're thinking about these questions in the context of joint action.

As we shall see, the type of scenario with James and Giulia, as with Mya and Iva, turns out to be a revealing test case for theories of commitment. For it's easy to be lured into thinking that commitment dissolution 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. I'll 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), which say much about how commitments are generated but little about how they are dissolved. More generally, as I've discussed in detail, 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. The essence of the previous chapter is, recall, what Fernández-Castro and Pacherie (2020) called "the credibility problem for commitments". Existing theoretical approaches to commitment in the philosophy of joint action don't provide a sufficiently robust account of the motivational basis that fully explains why agents abide by their commitments and, so, why their commitments should be seen to be credible. The shortcoming appears to be in theory rather than in practice: we observe that people do, in

reality, seem to meet commitments they have made, but find it hard to explain why if we take as a starting point existing theoretical accounts of joint action.

Besides the issue of credibility, another concern I've not yet raised is that current accounts of shared intention don't fully capture the myriad, and often subtle, ways in which we are motivated to meet our commitments. More specifically, the link between, on the one hand, a purely normative account of commitment in joint action and, on the other, a more phenomenological account which pays attention to the ways in which we actually experience, and are motivated to meet commitments, is weakly explicated in existing accounts. Part of this chapter aims to fill this gap 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 is modulated by situational cues of that partner's expectations. The studies described in this chapter 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. I contrast this with a simplified account of the dynamics of commitment dissolution inspired by the existing normative theories of joint action I've discussed thus far in this thesis.

As alluded to earlier, the minimal account proposed in this chapter is not incompatible with, nor does it seek to replace, any existing theoretical conceptualisation of commitment. Rather, the various views discussed in Chapters 5 and 6 provide a theoretical starting point by grounding what we might 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 to whom we have made commitments. Further to this, in proposing this minimal account, I am 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. The hope, instead, is to fill in the gaps about credibility and value identified before, as well as 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 this chapter hopes to achieve. Rather, another motivation is to look at the implications for a more general understanding of the psychological processes at play in

situations involving commitments in joint action. In particular, though a fully-fleshed out account that captures the full range of such processes is beyond the scope of this chapter, the 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. Recent empirical work points towards this need, as does what looks like the common theoretical problem identified in previous chapters, namely that to avoid circularity accounts must incorporate some basic psychological processes to initiate, or situate at the heart of, the sense of collectivism or sharedness that comes with sharing intentions. Moreover, it is the simple view—which draws directly on purely normative accounts of commitment—that has informed much of the empirical research that has been undertaken so far concerning the dissolution of commitments. It's therefore important to focus on contrasting this view with an alternative which sees possibly non-reflective processes, for which there's an increasing body of evidence, as integral to requests like this.

Overall, the simple view, which sees commitment dissolution as a straightforward matter, in which one simply requests release, tells us little about the different forces that motivate for and against this release. This way of thinking does not provide an adequate explanation capturing the actual dynamics that unfold in such situations. 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'll see, careful consideration of the dynamics of commitment dissolution also turns out to generate important insights into 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 think about commitments. Most importantly, careful examination of the dynamics of commitment dissolution enables us to provide answers to the following four key questions:

1) What principles do we appeal to when we want to dissolve a commitment?

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?

7.2 The simple view of commitment dissolution

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., 2016; 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. Let's 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.*

This view should feel familiar from preceding chapters, so to illustrate where it comes from, let's briefly recap the two well-known theoretical accounts of shared intention I've explored in detail, those of Margaret Gilbert and Michael Bratman. We can use their insights 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.

For Margaret Gilbert, recall, 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 are collectively created and so must be collectively rescinded. Without concurrence on release, all parties are obligated to one another to make their contribution to the joint activity. As discussed, her concurrence condition 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). Unlike Bratman's description of a social rationality that naturally emerges from typically cooperative individuals, Gilbert sees shared intention as involving essential robustness checks provided by a social normativity founded on a kind of obligation *sui generis* to shared intentional activity, justified by a shared understanding of and reference to the shared intention itself. Yet, as I've argued, Gilbert does not go further in discussing either the extent to which these obligations are expected to motivate individuals to meet them, particularly in the face of competing considerations, or the factors that individuals should take into account when requesting a partner's concurrence—notable as these are not familiar moral considerations. Gilbert's account therefore doesn't give us much in the way of understanding when it's appropriate or not to dissolve commitments.

In Bratman's 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, remember, Bratman doesn't regard joint action as involving a unique form of social normativity that creates special obligations between the agents involved. Rather, the commitments that are present are those that are distinctively characteristic of intentions as they feature in his planning theory of individual agency. 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 relation to the question of when it is appropriate to dissolve commitments, two features of Bratman's account are noteworthy. 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 and as I've emphasised several times

now, 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.

There's one area of Bratman's research which I have thus far not paid attention to in the thesis. Despite his methodological commitment to separating out moral versus rational drivers of shared intention, he doesn't totally ignore the relevance of ethical reasons agents may have to meet their commitments and expect others to meet theirs. In later work, Bratman actually does address, primarily 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. But he does not construct his own theory for what these might be. Instead, he outsources this component by drawing on Thomas Scanlon (1998). For example, discussing the moral requirement to meet expectations one has voluntarily and intentionally created in another (and on which they have come to rely), and in "the absence of special justification", I must do *X* unless You consent to *X* not being done (as Scanlon's 'Principle of Fidelity' concludes), or otherwise take "reasonable steps" to prevent or compensate for Your possible losses in cases where reasonable expectations you had of me are violated. Though Bratman says nothing (and Scanlon very little) about the factors and processes that characterise how I go about this and what I consider when requesting Your consent, this provides us with an early indication of how one 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.

To summarise, the simple view just presented is based partly on an interpretation of the way commitments emerge from these two purely normative accounts of shared intention. I've argued that neither account, in isolation, gives us much insight as to when it's appropriate to dissolve commitments. Moreover, if these accounts are reflective of more general views of shared intention, the common problem seems to be 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.

*

It is worth emphasising 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're 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, the argument made by Fernández-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. But the analysis in Chapter 6 concluded that there are reasons to doubt that either norms of intention rationality (Bratman) or social normativity (Gilbert) provide the kind of motivation needed for an agent to remain committed and eschew more attractive alternative options that may be in their interest—at least while remaining within their minimal conceptions of shared intention.

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 Gräfeinhain 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 behavioural 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 behaviours 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, we can 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 principles do we appeal to when we want to dissolve a commitment?), 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 simple 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, the simple view is inadequate. It does not explain why we may sometimes deem it inappropriate or awkward to request release. And it tells us nothing about the principles or 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, several authors have taken a different approach to thinking about commitment. In the following section, I'll 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 where 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 it comes to 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.

7.3 The Sense of Commitment framework

Recently in the psychological literature, Michael et al. (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 (or a related emotion) if she does not perform the action. To capture this, they introduce the notion of a 'sense of commitment', and following them I 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.[20] 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.[21] To borrow an example from Michael et al. (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, when an assurance has been given verbally in a promise-like form 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 communication (verbal assurance or non-verbal, e.g., Siposova et al., 2018). 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

---

[20] 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.

[21] It's interesting that Bratman leaves open the possibility that commitment might come in degrees, though he doesn't provide a reason for why this would be functionally useful and doesn't discuss it in detail. In his paper on Shared Cooperative Activity (SCA), he says:

> "SCA involves commitments to support the other that go beyond those of the unhelpful singers. How much beyond? Some participants in a SCA may be willing to incur what would normally be seen as fairly high costs in helping the other; others may be willing to help only if the costs thereby incurred are of a sort that would normally be seen as minimal. Willingness to support the other comes in degrees" (Bratman, 1992: pg. 104).

The rationale I am presenting for the sense of commitment would not contradict Bratman's theoretical view.

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. 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 as in this 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 (Michael et al., 2016b) has also shown that coordination in joint action can generate or enhance a sense of commitment.[22] 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

[22] Chennells et al. (2022) investigate broader hypotheses related to the impacts of coordination on interpersonal relations. In two experiments, designed to be as similar in task structure as possible, they separately test whether coordinated decision-making boosts altruistic motivation towards a partner and whether it increases trust in a partner. The results show that repeated coordination with the same partner increases people's resistance to tempting outside options—a prediction the sense of commitment theory would also make—but does not increase trust, supporting the first but not the second hypothesis.

as diverse as evolutionary theory and experimental economics and psychology, investigating the origins of human cooperation (e.g., Henrich & Henrich, 2007; Nowak, 2012; Tomasello, 2009; Skyrms, 2004; West et al., 2007). This has led to significant progress in specifying 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., 2007; 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. 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 shared intention and his reference Thomas Scanlon's work, mentioned earlier. Unsurprisingly, then, a focus on expectations therefore also provides insight into the kinds of things philosophers seem to care about when they discuss obligations that arise in collective activity.

Another interesting direction of research posits a possibly proximal, more basic need to belong (Fernández-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). In a related vein, Michael et al. (2021) explored the impact of social acceptance on commitment in a study focused on participants with borderline personality disorder (BPD). BPD patients face interpersonal problems due to a heightened sensitivity to cues of acceptance or rejection in their relationships, and the authors investigated certain psychological processes underpinning this

heightened responsiveness. An initial induction process was designed to trigger either acceptance or rejection followed by a coordination task, and the study analysed participants' commitment to their partners by sometimes presenting them with attractive opportunities to defect for their own benefit. The results showed that participants in the BPD group were less committed than participants in the control group when exposed to the rejection manipulation.

Overall, 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 what this chapter presents is a broadly 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 about proximal motivations to honour commitment provides a reason to expect people to honour commitments, and thus also a justification for relying on them to do so. It gives us a reason why people find commitments valuable, both in proximal terms, as an explanation for the kind of normative force commitments exert in real time, and in ultimate terms, in why they have value in the first place.

The reason why the three differences—graded motivation, the role of situational cues and the desire to meet expectations—between the simple view and the minimal account of a sense of commitment are particularly relevant to the discussion of commitment dissolution is that these differences 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[23].

[23] 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, one idea is to take a step back and introduce a further bit of theoretical cognitive machinery—namely, the concept of virtual bargaining. See Chennells and Michael (2022) for more on this.

7.4 Concerns about incommensurability and instrumentalism

Before considering the implications of using the sense of commitment to answer questions from BEACH about shared intention, motivational uncertainty and commitment, I need to address two important challenges to the framework.

The first concerns (in)commensurability of costs and benefits which the sense of commitment framework suggests affects agents' motivation to act committed. What has perhaps been taken for granted thus far in analysing both the simple view and the new proposal is the assumption that individuals have the ability to identify, measure and compare—accurately or otherwise—different utility costs and benefits associated with available actions. This raises the question of commensurability, given that various types of costs and benefits need to be compared. 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 the sense of commitment view, more so than for the simple view as it extends beyond the latter by including a possibly diverse and wide range of factors that affect individual judgements about whether or not to meet commitments.[24] 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 from going for a walk, let alone when contextual factors, like weather or sleep deprivation, are accounted for.

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 internally but also against 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, it's therefore 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 the research questions around motivation and uncertainty, is beyond the remit of this thesis, but both intra- and inter-personal benefit and cost utility comparisons—across multiple action-options and indeterminate action-outcomes—pose a problem for any psychological model that involves individuals making judgements about optimal courses of action. It's not straightforward how

[24] Models based on the assumption that agents act to maximise their utility have broad validity, as discussed in Chapter 3.2, but the more specific predictions made by the sense of commitment framework arguably make questions of commensurability more relevant and need to be tackled.

we evaluate relevant factors and arrive at correct judgements about generating and maintaining commitments—it's even plausible that agents are sometimes actually unable to arrive at a judgement about whether or not to honour a commitment at all! Yet, as discussed in the previous section, research does suggest that an agent's commitment towards a partner appears 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. This idea 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 between different actions. Emotions thus act as a neural common currency for choice (see Levy & Glimcher, 2012; Sescousse et al., 2015).

*

The second challenge to the sense of commitment framework is familiar from our earlier discussion (in Chapter 5) of instrumental versus non-instrumental motivations as a basis for intentionally joint activity. In particular, on the surface, the sense of commitment framework doesn't seem to align with the idea that individuals sharing intentions do not treat their partners in instrumental terms, that is, as social tools for achieving desired outcomes. The focus on shared intention supported by the need to meet expectations others have of us and expecting others to meet our own, or face the consequences, runs the risk of looking like the kind of strategic behaviour Bratman, in response to Petersson, was trying to avoid. I have

several things to say on this, on the way we think about what is or should be involved in joint activity and the kinds of cases we take to be paradigmatic of the phenomenon.

There's a risk that philosophers are overly averse to the idea that agents sharing intentions can be highly motivated to participate by the ends they expect to result. The idea that a theoretical view of joint action must preclude the possibility that agents treat their partners simply as a means to an end has had a big influence on minimal accounts proposed—as we've seen across the previous chapters. This discussion also tugs at some old and deep ideas in philosophy which have to do with the possibility and ways in which social interaction can undermine agency, autonomy, respect and the dignity of those involved. But is it right to shy away from these motivations? It's true that purely instrumental motivations do run counter to what we expect in some types of relationships. But whether the same applies to shared intentional activity should not be taken for granted, particularly because the paradigmatic cases authors traditionally focus on—certainly many of those discussed in this thesis—involve, at least in part, a distinctly instrumental motivation for acting with others. We act together to achieve something more easily than on our own or which would otherwise be unavailable to one or more of us. The challenge, then, is to formulate an account that reconciles (a) the need for joint action to involve a kind of 'sharedness' that is stronger than mere strategic interaction, or coordinated and goal-directed action in which agents see their partners purely as means to an end, with (b) the fact that cases of shared activity presented as typical usually revolve around a joint performance of actions specifically geared towards achieving desired goals.

Is it possible, therefore, that there are conditions under which it's appropriate to see one's partner as instrumental in achieving a (joint) goal while respecting the need to see the simple act of their participation as valuable in and of itself? One view is that this will depend on the nature of the relationship between the individuals involved. For example, an intuitive and widely held notion is that friendship shouldn't involve the kinds of instrumental motivations just described—this is precisely what friendship is *not*, some argue. Philosophical takes on friendship usually begin with something like it being "a distinctively personal relationship that is grounded in a concern on the part of each friend for the welfare of the other, for the other's sake, and that involves some degree of intimacy" (Helm, 2021: pg. 1). So, while Aristotle (Nicomachean Ethics, Book VIII) distinguished between three kinds of friendship—of pleasure, of utility, and of virtue; each of which provides us with its own reason for loving our friend—most modern writers see a friction between the last and the

first two; between concern for a friend for her sake (or, more specifically, 'motivated by the excellences in your friend's character') and the pleasure and utility that result from such concern: "If you benefit your friend because, ultimately, of the benefits you receive, it would seem that you do not properly love your friend for his sake, and so your relationship is not fully one of friendship after all" (Helm, 2021: pg. 2). This has led to the classification of pleasure and utility friendships as 'deficient modes' of friendship and their contrast with 'genuine, non-deficient' friendships based on virtue.

Of course, while intimacy and mutual care are important markers of a friendship, these alone may not be sufficient for having this kind of relationship. What's needed, in addition, are pointers both to the fact that friends routinely *do* things together and to their motivation to do so based on the perceived intrinsic value of their interaction. This immediately takes us back to tension I noted at the start:

> "A final common thread in philosophical accounts of friendship is shared activity [is which] friends engage in joint pursuits, in part motivated by the friendship itself. These joint pursuits can include not only such things as making something together, playing together, and talking together, but … for these pursuits to be properly shared in the relevant sense of "share," they cannot involve activities motivated simply by self interest… Rather, the activity must be pursued in part for the purpose of doing it together with my friend, and this is the point of saying that the shared activity must be motivated, at least in part, by the friendship itself" (Helm, 2021: pg. 8).

Though we're concerned with characterising joint action and not friendship, these ideas hint at a plausible way of thinking about how to reconcile the two needs of genuine sharedness and instrumental concerns. The simple idea is that there's an important difference between having *purely* instrumental reasons for participating in shared activity versus *some* instrumental considerations, in the latter attributing some value to doing something or bringing something about *together*. There's an intrinsic value in the action being joint, and maybe even joint with *this* person.

*

This isn't a satisfactory end point, as it raises the obvious question about how 'much' instrumentality is allowed. Is it appropriate to exploit one's partner's participation a little or a lot? Perhaps this is best answered empirically, as it seems hard to shift away from personal

intuitions on the matter, and there's an enormous literature and wealth of empirical research in psychology and economics that can provide insight. This includes analysis of behaviour in social contexts where outcomes are the result of the contributions or decisions of all involved and where we have evidence about attitudes towards exploitation, such as punishment in the face of unfair treatment given what's expected, which helps us identify normative standards.[25]

Ultimatum Games (UG) provide a widely employed and crude test of norms of fairness. In its original version, one person (giver) is gifted an amount of money and given the option to share some of their windfall with a second person (receiver), with the caveat that it is up to the receiver whether to accept the division. Givers can give zero, some or all of their cash to the receiver who either accepts the split, in which case they get the agreed amounts, or rejects it, in which case both get nothing. Given the rationale that it's strictly better for the receiver to receive anything, no matter how little, than nothing at all, we should expect that she will be willing to accept any allocation the giver proposes greater than zero, including the minimum possible (e.g., 1%). The receiver, knowing and anticipating this, should offer only this minimum amount if he is motivated purely by material self-interest and seeks to maximise his outcomes.

That these experiments lead to behaviours which differ in reality from those predicted is well documented, and generate several norm-related insights that are useful for us here. First, givers typically offer more than the minimum but not an equal split (typically a 20%–40% proportional cash split in their own favour). Second, receivers generally accept less-than-equal splits falling within a 20–40% band, suggesting that they somewhat tolerate the giver's lean towards self-interest. Third, that receivers typically reject offers below this range suggests that *excessive* self-interest is perceived as unfair, even punishable at a personal cost. Crucially, several studies demonstrate how this type of punishing behaviour is a reaction to the giver's inferred intention rather than purely an aversion to unequal payoffs and outcomes. Falk et al. (2003) show that identical offers in an UG are rejected at different rates depending on the giver's other possible offers, with rejection rates higher when a more equal offer split was available. The same authors also test the influence of full control versus allocation by an independent mechanism, a type of computerised random assignment (Falk et

[25] There are many ways to frame the concern about exploitation, and I don't need one way in particular to make the point, so I call it exploiting, rather than: using someone as a means to an end; unfairly exploiting one's collaboration partners (specifically, exploiting their contributions) for personal gain; benefiting oneself at the expense of the other (whether they incur real or opportunity costs); letting others down when their outcomes and they are reliant on our contributions as well; etc.

al., 2008; earlier studies using this approach include Blount, 1995, and Offerman, 2002). They find that punishment of unequal allocations is largely driven by perceived unfair intentions, behaviour interestingly consistent when outcomes are both positively or negatively in the receiver's favour (so intentions matter for both punishment and reward). Bereby-Meyer and Fiks (2013) explore costly punishing behaviour as a function of age, while also testing the influence of motive versus outcomes by having subjects play the UG either with a human proposer or a random machine. Five-year olds proposed and accepted divisions regardless of the source or split, but eight-year olds and twelve-year olds tended to reject unfair offers from a human proposer but accept them from the random device. This suggests that by age eight, children have fairness norms as well as the ability to infer the intention behind an unfair offer. Sutter (2006) likewise finds that children (seven to ten years) and teens (eleven to fifteen years) react to perceived intentions like university students. However, the first two groups reject unequal offers much more often than the latter, suggesting that the relative importance of intentions increases with age. Lastly, Cushman et al. (2009) disentangle what they take to be a methodological issue in several of the studies like the ones just described: the separate whether the allocation was intended by the giver and whether they had any control over the allocation. They introduce a game in which monetary allocations are made with a "trembling hand", so that givers have partial control such that intentions and outcomes are sometimes but not always mismatched. They demonstrate that controllability is at least as important as intent in determining punishing behaviour, hypothesising that punishment only makes sense if the other party can influence the probability of outcomes in future interactions. They also found that both intention and control effects were in force for selfish versus but not generous and fair behaviours. In short, there is good empirical evidence that perceived *intended* self interest is punished over and above the outcomes that result.

The fourth insight from UGs is that givers appear to be aware of and anticipate these responses to selfish intent, hence it's not by chance that most offers fall within the non-punishable range. For example, Güth and van Damme (1998) find that in an UG the kind of information on the allocation that's common knowledge matters, and that subjects act more in their own self-interest when they can get away with it. Specifically, the lower the information content of a message sent to the receiver—when the receiver learns nothing about their proposed allocation—the more selfishly the giver behaves, making more unequal and personally favourable offers. It's also not the case that fairer allocations are simply the result of other-regarding preferences rather than anticipating punishment. While an offer in an

UG significantly above the minimum might be an indicator of altruistic motivation, a comparison with Dictator Games, in which the receiver cannot reject and is forced to accept the giver's split, typically shows substantially lower offers, implying that care for receivers' interests alone cannot fully explain the typical allocation in ultimatum games. The type of punishment need not even be costly, as Dana et al. (2006) showed in experiments using Dictator Games (where the receiver has no opportunity to reject the offer). They found that subjects were willing to pay a cost to exit the game provided the receiver never knew a game was being played, a result not seen when the game was automatically private. This suggests that the receiver's beliefs were a main driver of the decision to pay to exit—an attempt to avoid violating social expectations they anticipate they would otherwise face.

The last insight from UGs comes from modifications to these experiments which enable a watching third party (who receives no payout and is unaffected by the split) to accept or reject the offer. These are useful to test whether there are general social norms about what behaviour is expected and appropriate, as third-party punishment addresses the confound that individuals punish because they have been personally affected. Results typically indicate that third parties sanction low-ball offers by rejecting them on behalf of the receiver, even in cases where it's personally costly for the third party to do so (e.g., Fehr & Fischbacher, 2004a; see Krueger & Hoffman, 2016, for an emerging neuroscience on this). There is even evidence that these norms, hypothesised to sustain large-scale cooperation, are culturally transmittable. Salali et al. (2015) test a simplified version of the third-party punishment game with children aged three to eight years, and find that all children imitate costly punishment for both equal and unequal offers, with rates of imitation increasing with age, but that only older children imitate not-punishing for equal and unequal offers.

Findings from UGs thus give us some insight into how shared activity can accommodate self-interested behaviour. First, there is widespread understanding, and at even a young age, that an intention to benefit at another's expense should and will face costly repercussions (Casari & Luini, 2012, even present evidence that punishing behaviour is proximal, which we are hard-wired to mete out; and see Fehr & Fischbacher, 2004b, for an additional review of the evidence). And second, that some self-interest is tolerated, in that there appear to be important fairness benchmarks against which behaviour is assessed and whose violation results in punishment.

*

Of course, UGs aren't quite joint actions in the traditional sense. They involve an obvious power imbalance, generally present a zero-sum set of payoff options, don't require any sort of close-knit personal interaction, tend to focus on decisions rather than actions and don't make much room for intention alignment, all things commonly taken to be part of joint activity. Perhaps the most important issue with UGs, however, is that while research into them gives us some idea as to 'how much' self-interest is acceptably, it still emphasises this behaviour as an undesirable feature of collective activities. There's a potentially more radical route. An alternative approach is to explicitly *embrace* the idea that people who act in concert to enact a joint outcome, knowingly and approvingly serve as means to their partner's goals, rather than seek to squeeze this out. It's possible that an account of shared intention can be compatible with instrumental motivations made explicit, perhaps under a certain set of conditions.

This idea stems from the theme of a recent series of papers by Orehek and colleagues (Orehek & Forest, 2016; Orehek & Weaverling, 2017; Orehek, Forest & Barbaro, 2018; Orehek, Forest & Wingrove, 2018; Chandler et al., 2021), who explore a 'people-as-means' perspective on interpersonal relationships. Their broad idea is that we should acknowledge the fact that goal pursuit is often attempted together with others, and that there's therefore a close connection between relationships and achieving one's goals. Their theoretical approach is to extend ideas from traditional goal-systems theory (see e.g., Kruglanski et al., 2002) by incorporating the notion that people serve as means to the goals of others—"helping other people to reach their goals in a variety of ways, such as by contributing their time; lending their knowledge, skills, and resources; and providing emotional support and encouragement" (Orehek, Forest & Barbaro, 2018: pg. 1). Furthermore, the authors propose that people find it rewarding to assist others in their goal pursuit as well as experience feelings of satisfaction and commitment when they, in turn, receive assistance.

The primary insight from goal-systems theory on which these authors draw is the relatively uncontroversial idea that agents typically value objects that serve as means to their goals more than objects that don't. Applied to relationships, agents are hypothesised to be attracted to partners whom they perceive to be instrumental to their active goals (Orehek & Forest, 2016). This leads to several predictions, including that we find partners attractive when this signals goal congruence (similarity) or when they are able and willing to serve as a means to goals not already personally satisfied (complementarity). They also generate new predictions, including that people are satisfied (and may even desire) being evaluated according to instrumentality for a goal they would like to serve towards, and are likely to be

dissatisfied when they stop being perceived in this way or are not recognised for this despite their efforts (Orehek & Forest, 2016). Moreover, goal-systems theory would also suggest that a person's happiness with being perceived as instrumental is boosted if she feels that she is less fungible—that is, less interchangeable with other people or objects—and if she acts as a unique means to a single goal or if she is instrumental to multiple goals (Orehek & Weaverling, 2017).

Some predictions generated by this theory have been tested. Fitzsimmons and Sha (2008) found that active goals bring to mind significant others who are helpful towards this goal (Experiments 1 and 3) and active goals cause people to evaluate instrumental others more positively than non-instrumental others (Experiments 2-5). Orehek, Forest and Wingrove (2018) find support for two hypotheses related to goal multifinality: first, that partners who serve more goals are evaluated as more interpersonally close, supportive and responsive than those who serve fewer goals; and partners who serve more goals are less common in social networks than those who serve fewer goals. They also found a person's average level of instrumental evaluation across their social network was associated with a stronger link between the number of goals served by a particular network member and their relationship evaluation.

It may be obvious to state that people are unhappy to serve as means to goals they do not value. But a people-as-means perspective can plausibly encompass certain types of social interaction in which agents objectify one another, using, exploiting or manipulating them to reach particular goals. It's possible that adopting it for, say, shared intentional activity would therefore licence accounts to include cases involving possible objectification and agents treating one another *only* in terms of their usefulness. Given the lengthy treatment in Chapter 5 of how philosophers seek to differentiate shared intention from purely self-interested behaviour, how can we go ahead and build in these positive ideas about instrumentality? This is not an easy question to answer, but perhaps we have no choice. Orehek and Weaverling (2017), for instance, argue that objectification is an "inevitable psychological process of evaluation" and, as an automatic process, cannot itself be immoral. They do say, though, that the goal for which another is evaluated *can* be judged to be, which provides a possible route to an answer to the question just posed: that is, that the goals towards which an agent is seen as instrumental will matter for whether or not they are happy to be objectified. Possible studies could therefore manipulate whether the agent judges the goal to be immoral or whether it's a goal for which they do not want to be instrumental. These findings might help

us make progress were we to consider incorporating a people-as-means view in an account of shared intention.

*

To summarise, a people-as-means view helps us envision how an intentional activity can be shared even while agents have instrumental motivations for participating. What the description above presents is a plausible, minimal (though non-exhaustive) set of conditions under which instrumental motivations are compatible with the kind of sharedness many writers on joint action seek. It's possible that all parties can be satisfied when, for example, they experience "mutual perceived instrumentality" (Orehek & Forest, 2016: pg. 1), in which each feels instrumental to their partner's valuable goals and perceives their partner as instrumental to their own, and, importantly, where these goals are judged to be desirable. For the purposes of this chapter, this discussion on partner instrumentality provides enough, I think, to counter the worry that the sense of commitment theory—and its focus on a sensitivity to meeting expectations—is incompatible with the kind of sharedness most people think of when characterising joint activity. In addition, it's possible that proximal psychological sensitivities to both wanting to be instrumental to others and to wanting to meet their expectations are highly compatible and need not compete. In fact, we might consider that a sensitivity to expectations is an excellent mechanism for accurately inferring what goals a partner desires to achieve, and meeting these expectations a good way of helping them achieve their goals. Moreover, it's plausible that a proximal mechanism linking positive feelings to being instrumental to one's partner serves the same ultimate reasons presented for the sense of commitment earlier, including a desire to maintain a good reputation, support interdependence and positive relationships and satisfy a basic need to belong. More work could be done to tease these apart and to explore whether mutual instrumentalism could or should form part of a minimal account of shared intention.

7.5 Conclusion

At the outset of this chapter, I 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? This was partly to begin developing an account of interpersonal commitment that's credible and which better captures the multiple, diverse reasons why agents are motivated to generate and maintain them—and so when others might

also expect them to do so. This addresses the main topic of this thesis, which concerns how, like in the circumstances in BEACH, agents can rely on their partners' intentions to contribute or participate despite having good reason—given the presence of attractive alternatives—to doubt they will. Focusing specifically on contexts of potential commitment dissolution illuminates what might be required for this to be true, and I set out to develop an account which would incorporate answers to the following four key questions:

1) What 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?

I want to conclude the chapter by turning back to and answering these questions and by exploring how we might use the sense of commitment framework to address motivational uncertainty within a minimal account of shared intention.

I started out by considering what could be 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 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 interpersonal commitments.

Having identified several problems with this simple view, I presented a recently developed theory based on the sense of commitment in joint action. This framework includes the proposal 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.

Crucially, this proposal can provide a basis for answering the four key questions identified above. In answer to the first, the proposal is that the factors/principles we appeal to when considering commitment dissolution are those which affect our sense of commitment towards a partner; that is, 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, the 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, et cetera. This is partly why simply proposing that agents communicate to change plans is often not a solution, as this process is itself subject to pressures in signalling the extent to which one is experiencing trust or commitment. 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, the 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 of and reliance on us. 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, time, emotions or other costs.

Finally, the proposal implies that what we actually care about when negotiating commitments is maintaining 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 there's no denying that we often care about meeting our obligations, the proposal is that we care about this to the extent that this helps us to maintain our relationships and our reputation—perhaps mediated by the extent to which we feel we are supportive of and useful for achieving our partner's goals. Indeed, the proposal implies that our sense of commitment might even, 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 (especially insofar as our relationship with that agent is important to us).

The approach the 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. This is exactly what we're after when thinking about the case of Mya and Iva agreeing to meet at the beach and the uncertainty that emerges in the meantime. 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 centre 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 sceptical 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 this connection, it bears emphasising that the proposal presented 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 honour 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.

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, what's been sketched is instead a procedure in which different principles and factors may figure differently 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, 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 given the issues described in the previous chapters. Moreover, these contexts provide a fruitful space for thinking through the kinds of basic, proximal psychological mechanisms—such as the desires to meet others' expectations and be instrumental to goals of theirs we support—that might support shared intentional activity.

Conclusion

In this thesis I've explored the (in)compatibility of traditional accounts of shared intention with contexts involving motivational uncertainty. I began by proposing a hypothetical situation in BEACH, involving the specific case of Mya and Iva who initially each intend that they go to the beach together but where, along the way, Mya becomes aware that Iva faces new attractive alternatives—in the form of her favourite team's football match—that make him uncertain, as he waits by the tower at the beach, about whether or not she still intends to join him. The question this posed is whether Mya and Iva can still share the intention that they go to the beach in spite of Mya being uncertain of Iva's intentions and her motivation and willingness to join him. Despite this being a relatively simple and straightforward situation, reflecting experiences we may often encounter when doing things with others, the question is surprisingly difficult to answer on the basis of several existing, leading theoretical accounts of shared intention.

The first three chapters of inquiry detailed why, by identifying three significant challenges to the possibility that shared intention is compatible with one or more agents being uncertain about partner motivations and intentions. In response to each challenge, I proposed solutions aimed at reconciling the tension that motivational uncertainty introduces.

I first explored how common knowledge immediately goes missing in Mya and Iva's case, as it's incompatible with there also being substantial uncertainty about intentions. Drawing on Olle Blomberg's work, I suggested that it's plausible that the lack of common knowledge doesn't preclude the possibility of certain kinds of intentional social interaction. However, I also argued that common knowledge typically serves an important role in accounts of shared intention, in enabling all parties to settle matters about what they will do together. Therefore, if it's plausible there is shared intention in BEACH, then we must (1) accept that common knowledge may be sufficient but not necessary for intentions to settle matters in joint settings, and (2) find alternative reasons for why individuals can settle what they will do together when there is motivational uncertainty.

I then addressed the second concern, about the prospects of an overly weakened belief requirement on intention should we open up the possibility of shared intention in BEACH. I found it useful to turn to the literature on individual agency to search for an answer, given the

robust debate on a similar topic there. Drawing on Michael Bratman's planning theory of intention, and his Asymmetry Thesis in particular, I found a possible way to reconcile an agent's intending yet being uncertain about her joint action's potential for success. However, there's a subtle problem with the solution presented: though a valid norm of practical rationality, the straightforward application of the Asymmetry Thesis seems to require that we see agents who share intentions as treating their partners' intentions no differently to other facts about the world relevant to their joint activity.

I parked this intuition and tackled what looked like a third hurdle to the possibility of shared intention in BEACH. What's still absent, in contexts involving uncertainty about intentions, is the important sense in which agents' intentions settle matters about the shared activity. To see what's missing, we need to better understand this requirement, so I presented how two authors—Michael Bratman and Johannes Roessler—have argued what this essential settling characteristic looks like in the case of shared rather than individual intentional activity. Both authors propose their own theoretical account of joint settling, explaining how each individual is in a position to settle matters about what the group will do. What's interesting is that, despite their very different methodological approaches, both authors rely on similar general assumptions about cooperativity and ordinary predictability to argue why a joint settling requirement is met. The problem we face in cases like BEACH, however, is that these background assumptions that support joint settling are plausibly not met in cases where there's motivational uncertainty. If Mya and Iva are able to share intentions, we thus have not yet solved how in their particular situation their intentions can still settle matters.

In light of this unanswered question about how a joint settling requirement is met, the intuition about what was previously wrong now makes more sense. A lens of joint settling helps pinpoint what was wrong with the direct mapping of the AT to solve motivational uncertainty in shared activity. I highlighted this by turning to Bratman's argument that shared intention involves a commitment to one's partners that comes from each party intending, and not only predicting, their partners' intentions and actions. What the AT gives us seems to fall short of what's required for settling precisely because it opens up the possibility of agents acting based merely on predictions of how their partners will act.

I therefore explored whether Bratman's notion of an emergent interpersonal commitment can be of use, given that commitments like this are, in the literature of joint action, seen as a popular tool for reducing motivational uncertainty. Interpersonal

commitment, in Bratman's eyes, demands of those jointly engaged that they act and be disposed towards supporting one another in pursuit of the joint action should it be required. This therefore provides a plausible explanation for why agents can rely on their partners and depend on them to make their contribution to the joint activity, despite having other reasons to be uncertain whether they will. If agents are committed to one another, and this is common knowledge, then each may be in a position not only to to reliably predict how their partner will act, but to see one another as mutually responsive and exerting a degree of control over each other in ways that support the persistence interdependence of their intentions as well as their sense of shared agency, and so provide a basis for them to jointly settle matters despite the motivational uncertainty present.

But there are ramifications to taking this route, given that it relies on the idea that both intentions *and* commitments can settle matters. This approach makes no sense to adopt if we think intentions and commitments cannot come apart, otherwise uncertainty about intentions surely implies uncertainty about commitments. We must therefore think of commitments as distinct from intentions, though of course they cannot be totally disconnected as then there wouldn't be a reason why they settle anything at all. This is plausible if we accept that intentions do characteristically generate intentions and that we have a theory about this connection, which we do from the earlier discussion on belief requirements on intending.

If social commitment provides a solution to the query of how there can be joint settling in BEACH, then it's crucial to know why commitments guide agent behaviour, not only that they do. In fact, Bratman's theory of commitment, which initially looked promising, is, on closer inspection I argue, not credible; specifically, it doesn't credibly explain why people are motivated to meet their commitments to their partners. This is primarily down to the fact that it's his ideas about cognitive and informational constraints which rationalise his original notion of commitment in intentional action, but which do not therefore provide us with any *socially*-relevant reasons why they should be kept. This means that Bratman's view of commitment is not going to be useful for situations with motivational uncertainty like BEACH. However, in addition, this finding poses risks to Bratman's account more generally. One comes from the fact that Bratman uses interpersonal commitment—and the supportive behaviours and dispositions which are said to flow out—as evidence of what I called non-tokenistic and non-instrumental social interaction. His claim is that the presence of this kind of commitment ensures that collective actions driven by purely strategic, self-interested motivations on the parts of those involved are not encompassed by his account, as several

critics have challenged they are. Bratman therefore uses the requirement for interpersonal commitment to defend a strong sense in which an action is intentionally joint within his account of shared intention. But if commitments aren't credible and the behaviours described do not in truth uniquely identify non-strategic behaviour, as Bratman argues they do, then this defence isn't available to him. A second, related risk, is that weak commitment credibility also undermines his proposed 'division of philosophical labour' between describing a minimal account of basic rationality, that supports shared intention, and relying on obligations, moral or otherwise, to provide the necessary support. This approach now only looks possible because we've implicitly partialled out situations in which commitments might have any role or power in explaining what it means to share intentions.

We're left with the idea that though commitments might be a useful tool, Bratman's characterisation isn't enough for us to go on. To see whether this issue is limited to Bratman, and to give interpersonal commitments their due, I turned to an account that places them front and centre: Margaret Gilbert's theory of shared intention and the joint commitment and mutual obligations she sees as an essential feature of it. Her account has several positive features which address some of the concerns from the investigation into Bratman. But, I argued, ultimately we're left with the same issue, in that Gilbert too lacks a credible view of why individuals are motivated to meet their commitments. One of the reasons for this is that Gilbert's proposal for a *sui generis* kind of obligation in shared intention leaves us without a benchmark for assessing their normative (motivational) strength and without a proposition for how commitments influence an agent's practical reasoning process. It is left up to our intuition as to how effectively they may or may not promote shared activity. To better understand what Gilbert is after, I looked into the sources of inspiration for her claim-rights perspective, unpacking whether she's correct in her proposal that mutual obligations in shared intention should be understood as non-moral in nature— and found that the same challenges are raised against those authors on which she relies. Finally, I reverted back to 'reductive'-style accounts by exploring Berislav Marušić's work on individual agency and committing to actions under uncertainty. Looking at an extension he provides to cases of joint activity, in which he presents trust as a mediating factor, I argued that despite his interesting corollary to commitment, we end up with the same issues as Bratman—and that making additional background assumptions can help here, but that this approach might be circular or oversteps the continuity between individual and shared agency.

The conclusion of these investigations is that we're still missing a robust explanation for interpersonal commitment/trust in shared intention which is both credible and which remains within the modest constraints of a minimal and general account. All three authors therefore struggle to explain how there can be shared intention in contexts with motivational uncertainty. The last contribution of this thesis was therefore an attempt to provide one idea for how to conceptualise social commitment in joint action which

1) is credible, in providing a sound rationale for why we are motivated to meet our commitments to others,

2) is not circular, in that is doesn't draw on any sense of intention sharing to explain its origin,

3) requires no additional metaphysical claims about supra-personal agents, and

4) realistically describes the nuances of the experience of feeling committed.

In so doing, it built on the previous theories proposed while addressing some of their shortcomings. To start, I framed the problem in the context of commitment dissolution, a revealing test case for theories of commitment because it's precisely in these situations that traditional accounts struggle.

I then set up what can be called a simple view of commitment based on certain existing theoretical accounts and discussed its shortcomings. Using a recent body of work by several authors, I presented their framework for a minimal psychological sense of commitment arising in joint activities. I outlined the theoretical background—in particular, the basic need to meet reasonable expectations others have of us, the need to maintain good relationships with others and the need to belong—and showed how it is this minimal commitment which can perform the function of allowing individuals to rely on others when there is uncertainty about intentions. The proposal is guided by several theoretical viewpoints, including the need to encompass both explicit and implicit commitment generation processes, the need to account for both proximal as well as ultimate psychological processes and the need for a graded characterisation of the experience of feeling committed.

As was evident, the proposal was primarily descriptive, based on recent empirical work which supports some of the theory's predictions. But it's plausible that the account has normative implications for a characterisation of commitment too. By spelling out why people

tend to meet their commitments—namely, to build meaningful relationships, engender a sense of belonging and avoid ruining one's reputation—I also suggested that we have reasons why it is sometimes justified to expect and rely on people to honour their commitments—that is, why we should honour commitments and expect others to honour theirs.

The proposed framework does, however, lend itself to a utilitarian lens for thinking about commitment, which raises two concerns which I addressed. The first has to do with tradeoffs, how agents are thought to weigh up the costs and benefits in reasoning about whether or not to keep their commitments. There is no ready answer to the problem of incommensurability, though I noted that this is a problem for any account of psychological decision making. The second concern had to do with making room for the kind of instrumentality that many authors tend to preclude when sharing intentions, as laid out earlier in the discussion on intending versus predicting partner actions. This is an important concern, which I responded to by proposing a middle-ground that sees both instrumental and non-instrumental concerns as typically characteristic of shared intention. On the one hand, we want to rule out purely exploitative behaviour but, on the other, we need to take seriously the idea that we often undertake joint actions for desirable ends. I presented selected empirical work to show how we can understand this in practice, discussing studies on perceived intentional exploitation, the role of punishment in modulating behaviour and the potential for a mutual instrumentalism that sees us as happy to act as means towards goals we care about for people we care about.

In the end, we have what I think is the plausible idea that shared intention is possible in contexts with motivational uncertainty, even though traditional assumptions concerning what agents know and believe are not met, provided some minimal social commitments are instantiated. This, though, requires a view of shared intention that moves beyond the narrow approach of focusing primarily on the psychology of those involved, which tends to screen out normative questions regarding why we are motivated to do things with others. Nonetheless, if we take this route, then it's plausible that Mya can settle matters about him and Iva going to the beach, despite having reasons to be uncertain that she intends to meet him there, provided he can rely on her having at least a basic sense of commitment to him.

Bibliography

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, *100*(401), 464–477. https://doi.org/10.2307/2234133

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. https://doi.org/10.3982/ECTA7384

Anscombe, G. E. M. (1963). *Intention: Vol. 2nd edition*. Blackwell.

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. https://doi.org/10.1017/S0305000900012101

Baier, A. C. (1997). Doing things with others: The mental commons. In L. Alanen, S. Heinämaa, & T. Wallgren (Eds.), *Commonality and Particularity in Ethics* (pp. 15–44). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-25602-0_2

Baker, C., Saxe, R., & Tenenbaum, J. B. (2006). Bayesian models of human action understanding. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (pp. 99–106). MIT Press. http://papers.nips.cc/paper/2815-bayesian-models-of-human-action-understanding.pdf

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Bereby-Meyer, Y., & Fiks, S. (2013). Changes in negative reciprocity as a function of age. *Journal of Behavioral Decision Making*, *26*(4), 397–403. https://doi.org/10.1002/bdm.1768

Blomberg, O. (2016). Common knowledge and reductionism about shared agency. *Australasian Journal of Philosophy*, *94*(2), 315–326. https://doi.org/10.1080/00048402.2015.1055581

Blount, S. (1995). When social outcomes aren′t fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, *63*(2), 131–144. https://doi.org/10.1006/obhd.1995.1068

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., Michael, J., & Heintz, C. (2022). Perceiving commitments: When we both know that you are counting on me. *Mind & Language*, *37*(4), 502–524. https://doi.org/10.1111/mila.12333

Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge: Cambridge, MA: Harvard University Press.

Bratman, M. E. (1992). Shared cooperative activity. *The Philosophical Review*, *101*(2), 327–341. https://doi.org/10.2307/2185537

Bratman, M. E. (1999a). *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge Core; Cambridge University Press. https://doi.org/10.1017/CBO9780511625190

Bratman, M. E. (1999b). I intend that we J. In *Faces of Intention: Selected Essays on Intention and Agency* (pp. 142–162). Cambridge University Press. https://doi.org/10.1017/CBO9780511625190.008

Bratman, M. E. (1999c). Shared intention. In *Faces of Intention: Selected Essays on Intention and Agency* (pp. 109–129). Cambridge University Press. https://doi.org/10.1017/CBO9780511625190.006

Bratman, M. (2014). *Shared agency: A planning theory of acting together*. Oxford University Press.

Bratman, M. (2022). *Shared and institutional agency: toward a planning theory of human practical organization.* Oxford University Press

Broome, J. (1999). Normative requirements. *Ratio*, *12*(4), 398–419. https://doi.org/10.1111/1467-9329.00101

Broome, J. (2013). *Rationality through reasoning*. Wiley-Blackwell.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage (Vol. 4)*. Cambridge University Press.

Butterfill, S. (2012). Joint action and development. *The Philosophical Quarterly*, *62*(246), 23–47. https://doi.org/10.1111/j.1467-9213.2011.00005.x

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. (2018). A resource-rational analysis of human planning. *In CogSci.*

Casari, M., & Luini, L. (2012). Peer punishment in teams: Expressive or instrumental choice? *Experimental Economics*, *15*(2), 241–259. https://doi.org/10.1007/s10683-011-9292-6

Chandler, K. R., Krueger, K. L., Forest, A. L., & Orehek, E. (2021). Interested and instrumental: An examination of instrumentality regulation with potential romantic

partners. *Personality and Social Psychology Bulletin*, *49*(2), 197–214. https://doi.org/10.1177/01461672211061942

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), 1–9. https://doi.org/10.1038/s41598-018-34096-1

Chennells, M., & Michael, J. (2022). Breaking the right way: A closer look at how we dissolve commitments. *Phenomenology and the Cognitive Sciences—Special Issue: The Phenomenology of Joint Action: Structure, Mechanisms and Functions.* https://doi.org/10.1007/s11097-022-09805-x

Chennells M., Woźniak M., Butterfill S., & Michael J. (2022) Coordinated decision-making boosts altruistic motivation—But not trust. *PLOS ONE, 17*(10): e0272453. https://doi.org/10.1371/journal.pone.0272453

Clark, H. H. (2020). Social actions, social commitments. In *Roots of Human Sociality,* 126–150. Routledge.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "Trembling Hand" game. *PLOS ONE*, *4*(8) e6699. https://doi.org/10.1371/journal.pone.0006699

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

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Dean, M., Kıbrıs, Ö., & Masatlioglu, Y. (2017). Limited attention and status quo bias. *Journal of Economic Theory*, *169*, 93–127. https://doi.org/10.1016/j.jet.2017.01.009

Eidelman, S., & Crandall, C. S. (2012). Bias in favor of the status quo: Bias and the status quo. *Social and Personality Psychology Compass*, *6*(3), 270–281. https://doi.org/10.1111/j.1751-9004.2012.00427.x

Falk, A., Fehr, E., & Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry*, *41*(1), 20–26. https://doi.org/10.1093/ei/41.1.20

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness—Intentions matter. *Games and Economic Behavior*, *62*(1), 287–303. https://doi.org/10.1016/j.geb.2007.06.001

Fehr, E., & Fischbacher, U. (2004a). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4

Fehr, E., & Fischbacher, U. (2004b). Social norms and human cooperation. *Trends in Cognitive Sciences, 8*(4), 185–190. https://doi.org/10.1016/j.tics.2004.02.007

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

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. https://doi.org/10.1162/003355399556151

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).

Feinberg, J. (1970). The nature and value of rights. *The Journal of Value Inquiry*, *4*(4), 243–260. https://doi.org/10.1007/BF00137935

Fernández Castro, V., & Pacherie, E. (2020). Joint actions, commitments and the need to belong. *Synthese, 198,* 7597–7626. https://doi.org/10.1007/s11229-020-02535-0

Fernández‐Castro, V., & Pacherie, E. (2022). Commitments and the sense of joint agency. *Mind & Language (online early view)*. https://doi.org/10.1111/mila.12433

Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, *99*(4), 689–723. https://doi.org/10.1037//0033-295X.99.4.689

Fitzsimons, G. M., & Shah, J. Y. (2008). How goal instrumentality shapes relationship evaluations. *Journal of Personality and Social Psychology*, *95*(2), 319–337. https://doi.org/10.1037/0022-3514.95.2.319

Fleming, S. M., Thomas, C. L., & Dolan, R. J. (2010). Overcoming status quo bias in the human brain. *Proceedings of the National Academy of Sciences*, *107*(13), 6005–6009. https://doi.org/10.1073/pnas.0910380107

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. https://doi.org/10.1126/science.aac6076

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 for joint action. *Philosophical Psychology*, *26*(4), 588–603. https://doi.org/10.1080/09515089.2012.670905

Godman, M., Nagatsu, M., & Salmela, M. (2014). The social motivation hypothesis for prosocial behavior. *Philosophy of the Social Sciences*, *44*(5), 563–587. https://doi.org/10.1177/0048393114530841

Gold, N., & Sugden, R. (2007). Collective intentions and team agency. *The Journal of Philosophy*, *104*(3), 109–137. https://www.jstor.org/stable/20620005

Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development: Bayesian models of child development. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 75–86. https://doi.org/10.1002/wcs.1330

Gräfenhain, M., Behne, T., Carpenter, M., & Tomasello, M. (2009). Young children's understanding of joint commitments. *Developmental Psychology*, *45*, 1430–1443. https://doi.org/10.1037/a0016122

Grice, H. P. (1971). Intention and uncertainty. *Proceedings of the British Academy: Philosophical Lecture*.

Grice, P. (1989). *Studies in the way of words*. Harvard University Press.

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: Meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, *29*, 24–30. https://doi.org/10.1016/j.cobeha.2019.01.005

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004

Güth, W., & van Damme, E. (1998). Information, strategic behavior, and fairness in ultimatum bargaining: An experimental study. *Journal of Mathematical Psychology*, *42*(2), 227–247. https://doi.org/10.1006/jmps.1998.1212

Harris, C., Fiedler, K., Marien, H., & Custers, R. (2020). Biased preferences through exploitation: How initial biases are consolidated in reward-rich environments. *Journal of Experimental Psychology: General*, *149*(10), 1855–1877. https://doi.org/10.1037/xge0000754

Hart, H. L. A. (1955). Are there any natural rights? *The Philosophical Review*, *64*(2), 175–191. https://doi.org/10.2307/2182586

Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review*, *10*(1), 47–66. https://doi.org/10.1207/s15327957pspr1001_3

Haslam, N., & Fiske, A. P. (2005). Relational models theory: A confirmatory factor analysis. *Personal Relationships*, *6*(2), 241–250. https://doi.org/10.1111/j.1475-6811.1999.tb00190.x

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. https://doi.org/10.1017/S1930297500005581

Helm, B. (2021). Friendship. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/friendship/

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

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

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

Krueger, F., & Hoffman, M. (2016). The emerging neuroscience of third-party punishment. *Trends in Neurosciences*, *39*(8), 499–501. https://doi.org/10.1016/j.tins.2016.06.004

Kruglanski, A. W., Shah, J. Y., Fishbach, A., Friedman, R., Chun, W. Y., & Sleeth-Keppler, D. (2002). A theory of goal systems. In *Advances in experimental social psychology, Vol. 34* (pp. 331–378). Academic Press. https://doi.org/10.1016/S0065-2601(02)80008-9

Kurth-Nelson, Z., & Redish, A. D. (2012). Don't let me do that!—Models of precommitment. *Frontiers in Neuroscience*, *6*, 138. https://doi.org/10.3389/fnins.2012.00138

Kutz, C. (2000). Acting together. *Philosophy and Phenomenological Research, 61(1),* 1–31. https://doi.org/10.2307/2653401

Laurence, B. (2011). An Anscombian approach to collective action. In *Essays on Anscombe's* Intention (pp. 270–296). Harvard University Press. https://doi.org/10.4159/harvard.9780674060913.c11

Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, *117*(3), 785–807. https://doi.org/10.1037/a0019688

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

Lewis, D. (1969). *Convention: A Philosophical Study*. Blackwell.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, *E1*. https://doi.org/10.1017/S0140525X1900061X

Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N., & Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. *Advances in Neural Information Processing Systems*, *27*. https://proceedings.neurips.cc/paper/2014/hash/7fb8ceb3bd59c7956b1df66729296a4c-Abstract.html

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199591565.001.0001

Mant, C. M., & Perner, J. (1988). The child's understanding of commitment. *Developmental Psychology*, *24*, 343–351. https://doi.org/10.1037/0012-1649.24.3.343

Marušić, B. (2015). *Evidence and agency: Norms of belief for promising and resolving*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198714040.001.0001

Marušić, B., & Schwenkler, J. (2018). Intending is believing: A defense of strong cognitivism. *Analytic Philosophy*, *59*(3), 309–340. https://doi.org/10.1111/phib.12133

Michael, J., Chennells, M., Nolte, T., Ooi, J., Griem, J., Network, M. D. R., ... & Montague, P. R. (2021). Probing commitment in individuals with borderline personality disorder. *Journal of Psychiatric Research, 137*, 335–341. https://doi.org/10.1016/j.jpsychires.2021.02.062

Michael, J., Gutoreva, A., Lee, M. H., Tan, P. N., Bruce, E. M., Székely, M., Ankush, T., Sakaguchi, H., Walasek, L., & 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), 89–120. https://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

Michaelson, E. (2018). Lying, testimony, and epistemic vigilance. In J. Meibauer (Ed.), *The Oxford Handbook of Lying* (1st ed., pp. 214–228). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198736578.013.16

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

Moran, R. (2004). Précis of authority and estrangement: An essay on self-knowledge. *Philosophy and Phenomenological Research*, *69*(2), 423–426. https://doi.org/10.1111/j.1933-1592.2004.tb00403.x

Moran, R. (2020). Précis of the exchange of words. *Philosophical Explorations*, *23*(3), 267–270. https://doi.org/10.1080/13869795.2020.1802117

Nebel, J. M. (2015). Status quo bias, rationality, and conservatism about value. *Ethics*, *125*(2), 449–476. https://doi.org/10.1086/678482

Nicolle, A., Fleming, S. M., Bach, D. R., Driver, J., & Dolan, R. J. (2011). A regret-induced status quo bias. *The Journal of Neuroscience*, *31*(9), 3320–3327. https://doi.org/10.1523/JNEUROSCI.5615-10.2011

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*, *437*(7063), 1291–1298. https://doi.org/10.1038/nature04131

Offerman, T. (2002). Hurting hurts more than helping helps. *European Economic Review*, *46*(8), 1423–1437. https://doi.org/10.1016/S0014-2921(01)00176-3

Orehek, E., & Forest, A. L. (2016). When people serve as means to goals: Implications of a motivational account of close relationships. *Current Directions in Psychological Science*, *25*(2), 79–84. https://doi.org/10.1177/0963721415623536

Orehek, E., Forest, A. L., & Barbaro, N. (2018). A people-as-means approach to interpersonal relationships. *Perspectives on Psychological Science*, *13*(3), 373–389. https://doi.org/10.1177/1745691617744522

Orehek, E., Forest, A. L., & Wingrove, S. (2018). People as means to multiple goals: Implications for interpersonal relationships. *Personality and Social Psychology Bulletin*, *44*(10), 1487–1501. https://doi.org/10.1177/0146167218769869

Orehek, E., & Weaverling, C. G. (2017). On the nature of objectification: Implications of considering people as means to goals. *Perspectives on Psychological Science*, *12*(5), 719–730. https://doi.org/10.1177/1745691617691138

Pacherie, E. (2007). The sense of control and the sense of agency. *Psyche*, *13*(1), 1–30.

Pacherie, E. (2012). The phenomenology of joint action: Self-agency vs. joint-agency. In Seemann Axel (ed.) *Joint Attention: New Developments* (pp. 343–349). MIT Press.

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

Pauer-Studer, H. (2014). Rational requirements and reasoning. *Economics and Philosophy*, *30*(3), 513–528. https://doi.org/10.1017/S0266267114000315

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–321. https://doi.org/10.1016/j.cognition.2010.11.015

Petersson, B. (2007). Collectivity and circularity. *The Journal of Philosophy*, *104*(3), 138–156. https://www.jstor.org/stable/20620006

Pettit, P., & Schweikard, D. (2006). Joint actions and group agents. *Philosophy of the Social Sciences*, *36*(1), 18–39. https://doi.org/10.1177/0048393105284169

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(3), 833–838.

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. https://doi.org/10.1016/S0047-2727(03)00021-5

Roberts, G. (1998). Competitive altruism: From reciprocity to the handicap principle. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *265*(1394), 427–431. https://doi.org/10.1098/rspb.1998.0312

Roberts, G. (2005). Cooperation through interdependence. *Animal Behaviour*, *70*(4), 901–908. https://doi.org/10.1016/j.anbehav.2005.02.006

Roessler, J. (2020). Plural practical knowledge. *Inquiry*, 1–20. https://doi.org/10.1080/0020174X.2020.1787221

Roth, A. S. (2018). Interpersonal obligation in joint action. In *The Routledge Handbook of Collective Intentionality* (pp. 45–57). Routledge. https://doi.org/10.4324/9781315768571-6

Salali, G. D., Juda, M., & Henrich, J. (2015). Transmission and development of costly punishment in children. *Evolution and Human Behavior*, *36*(2), 86–94. https://doi.org/10.1016/j.evolhumbehav.2014.09.004

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59. https://doi.org/10.1007/BF00055564

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.

Schmid, H. B. (2014). Plural self-awareness. *Phenomenology and the Cognitive Sciences*, *13*(1), 7–24. https://doi.org/10.1007/s11097-013-9317-z

Schmid, H. B. (2016). On knowing what we're doing together: Groundless group self-knowledge and plural self-blindness. In Michael S. Brady, and Miranda Fricker (eds), *The Epistemic Life of Groups: Essays in the Epistemology of Collectives, Mind Association Occasional Series* https://doi.org/10.1093/acprof:oso/9780198759645.003.0004

Schmid, H. B. (2018). The subject of "we intend." *Phenomenology and the Cognitive Sciences*, *17*(2), 231–243. https://doi.org/10.1007/s11097-017-9501-7

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge University Press.

Searle, J. R. (1990). Collective intentions and actions. *Intentions in Communication, 401(4)*, 401–415.

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

Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, 129–138. https://doi.org/10.1037/h0042769

Simon, H. (1982). *Models of bounded rationality*. Cambridge, MA: MIT Press

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

Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge University Press.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Stoutland, F. (2008). The ontology of social agency. *Analyse & Kritik*, *30*(2), 533–551. https://doi.org/10.1515/auk-2008-0210

Sunstein, C. R. (2014). Nudging: A very short guide. *Journal of Consumer Policy*, *37*(4), 583–588. https://doi.org/10.1007/s10603-014-9273-1

Sutter, M. (2007). Outcomes versus intentions: On the nature of fair behavior and its development with age. *Journal of Economic Psychology*, *28*(1), 69–78. https://doi.org/10.1016/j.joep.2006.09.001

Sylwester, K., & Roberts, G. (2013). Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. *Evolution and Human Behavior*, *34*(3), 201–206. https://doi.org/10.1016/j.evolhumbehav.2012.11.009

Székely, M., & Michael, J. (2018). Investing in commitment: Persistence in a joint action is enhanced by the perception of a partner's effort. *Cognition*, *174*, 37–42. https://doi.org/10.1016/j.cognition.2018.01.012

Tollefsen, D. (2005). Let's pretend!: Children and joint action. *Philosophy of the Social Sciences*, *35*(1), 75–97. https://doi.org/10.1177/0048393104271925

Tomasello, M. (2009). *Why we cooperate*. MIT Press. https://mitpress.mit.edu/9780262013598/why-we-cooperate/

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57. https://doi.org/10.1086/406755

Tuomela, R. (2017). *Non-reductive views of shared intention*. In *The Routledge Handbook of Collective Intentionality* (pp. 25–33). https://doi.org/10.4324/9781315768571-4

Tuomela, R., & Miller, K. (1988). We-intentions. *Philosophical Studies*, *53*(3), 367–389. https://doi.org/10.1007/BF00353512

Velleman, J. D. (1997). How to share an intention. *Philosophy and Phenomenological Research*, *57*(1), 29. https://doi.org/10.2307/2953776

Vesper, C., Butterfill, S., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks*, *23*(8–9), 998–1003. https://doi.org/10.1016/j.neunet.2010.06.002

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

Wallace, R. J. (2020). Practical reason. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/practical-reason/

West, S. A., Griffin, A. S., & Gardner, A. (2007). Evolutionary explanations for cooperation. *Current Biology*, *17*(16), R661–R672. https://doi.org/10.1016/j.cub.2007.06.004

Wittgenstein, L., Anscombe, G. E. M., & Wright, G. H. von. (1969). *On certainty*. Blackwell.

~