ROUTLEDGE FOCUS

THE PHILOSOPHY AND PSYCHOLOGY OF COMMITMENT

John Michael

The Philosophy and Psychology of Commitment

The phenomenon of commitment is a cornerstone of human social life. Commitments make individuals' behavior predictable, thereby facilitating the planning and coordination of joint actions involving multiple agents. Moreover, commitments make people willing to rely upon each other, and thereby contribute to sustaining characteristically human social institutions such as jobs, money, government and marriage. However, it is not well understood how people identify and assess the level of their own and others' commitments.

The Philosophy and Psychology of Commitment explores and explains the philosophical and cognitive intricacies of commitment. John Michael considers how commitments motivate us and their often implicit and tacit nature. To flesh out the philosophical framework of his argument he draws on experimental work with young children, adults and human-robot interaction within the context of joint action, considering the role of the emotions and whether very young children are sensitive to commitment.

Providing an important account of the nature and operation of commitment, this book is essential reading for those working in philosophy of psychology, cognitive science, experimental philosophy, and social and developmental psychology. It will also be of interest to those working in emerging fields such as human-robot interaction and behavioral economics.

John Michael is a senior lecturer in the Department of Psychology at BPP University, London, UK, and affiliated faculty member in the Department of Cognitive Science at the Central European University, Vienna, Austria.

Routledge Focus on Philosophy

Routledge Focus on Philosophy is an exciting and innovative new series, capturing and disseminating some of the best and most exciting new research in philosophy in short book form. Peer reviewed and at a maximum of 50,000 words shorter than the typical research monograph, *Routledge Focus on Philosophy* titles are available in both ebook and print on demand format. Tackling big topics in a digestible format the series opens up important philosophical research for a wider audience, and as such is invaluable reading for the scholar, researcher and student seeking to keep their finger on the pulse of the discipline. The series also reflects the growing interdisciplinarity within philosophy and will be of interest to those in related disciplines across the humanities and social sciences.

The Right to Know
Epistemic Rights and Why We Need Them
Lani Watson

Honouring and Admiring the Immoral
An Ethical Guide
Alfred Archer and Benjamin Matheson

Newton's Third Rule and the Experimental Argument for Universal Gravity
Mary Domski

The Philosophy and Psychology of Commitment
John Michael

For more information about this series, please visit: www.routledge.com/Routledge-Focus-on-Philosophy/book-series/RFP

The Philosophy and Psychology of Commitment

John Michael

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2022
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record has been requested for this book

ISBN: 978-1-138-08549-7 (hbk)
ISBN: 978-1-032-12829-0 (pbk)
ISBN: 978-1-315-11130-8 (ebk)

DOI: 10.4324/9781315111308

Typeset in Times New Roman
by codeMantra

Contents

Acknowledgments

The work on this book was supported by a Starting Grant from the European Research Council (679092: Sense of Commitment) and a Philip Leverhulme Prize from the Leverhulme Trust. I would like to thank the following friends and colleagues who generously took the time to read and offer comments on the work presented here: Stephen Butterfill, Matthew Chennells, Alessandro Salice, Anna Strasser, Alexandra De La Trobe, Simon van Baal, Simon Myers, Martin Dockendorff, Christophe Heintz, Marcell Székely, Luke McEllin, David Dvorkin, Evan Westra, Günther Knoblich, Natalie Sebanz, Wayne Christensen and two anonymous reviewers. I am also grateful to my family, and especially to my husband Thomas Wolf, for the unwavering commitment which has sustained me and inspired me to see this project through.

1 Introduction

Commitment is the glue holding together characteristically human forms of social life. Commitments make individuals' behavior predictable in the face of fluctuations in their desires and interests, thereby facilitating the planning and coordination of joint actions involving multiple agents. Moreover, commitments make people willing to perform actions that they would not otherwise perform. For example, a taxi driver picks up his clients and transports them to their desired destination because they are committed to paying him afterwards for the service, and a construction worker performs her job every day because her employer has made a credible commitment to pay her at the end of the month. Indeed, the taxi driver and the construction worker are willing to accept money as payment only because a network of other agents (notably the central bank) is committed to taking various measures to sustain the currency in question. Thus, social objects and institutions such as jobs, money, government, scientific collaborations and marriage depend for their origin and stability upon the credibility of commitments.

Despite the crucial importance of commitment for characteristically human forms of sociality, it is not well understood how people identify and assess the level of their own and others' commitments. In fact, there has not even been much research conducted with the aim of gaining a better understanding of this – at least not until just recently. I think this is at least in part because commitment is a rather elusive concept. It comes up in many different contexts and plays important roles in various disciplines, such as philosophy, psychology, economics and anthropology, and yet it is rarely defined explicitly. And, as we shall see later on, when it is defined, it is not defined in such a way as to facilitate the testing of hypotheses about factors influencing it or about cognitive mechanisms underpinning it.

DOI: 10.4324/9781315111308-1

The current book is intended to redress this shortcoming, i.e., to illuminate the phenomenon of commitment (what it is, what influences it and what consequences it has). Before undertaking this challenge, though, it is important to acknowledge an elephant in the room – namely, the fact that commitment is not only an inchoate concept but also a heterogeneous one. In everyday life, we experience many different forms of commitment. Consider the following four examples:

a Agnes made a commitment to pick Sam up at the airport tomorrow.
b 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, and she feels committed to showing up.
c Frank was unsure whether to go to the cinema or the theater tonight, but he decided in favor of the cinema and now he is committed to that plan.
d Roger is committed to birdwatching and spends considerable amounts of time and money pursuing this hobby.

There are many differences among these four examples, and they could be used to illustrate a number of distinctions which one might make among different forms of commitment.[1] In the present book, two of these distinctions will be particularly important. The first is the distinction between individual and social forms of commitment. The second is the distinction between normative and psychological aspects of commitment. I will begin by saying just a bit about each of the first two distinctions.

First, consider the distinction between individual and social commitment in relation to the examples above: Agnes and Polly are committed at least in part because the goals in question are ones that are valuable to other people, whereas this is not the case for Frank and Roger. This raises the question: How do individual commitment and social forms of commitment relate to each other? Of course, we should not assume that individual and social commitment have anything interesting in common simply because we sometimes use the same English word to refer to them. However, as we shall see later on, both individual and social commitment can be illuminated by careful consideration of the ways in which they relate to each other. Thus, although the main focus in this book is on social commitment, I will also devote space to the discussion of individual commitment.

Second, consider the distinction between normative and psychological aspects of commitment. This distinction is not independent of the distinction between social and individual commitment. This is because norms come into play differently depending on whether one is talking about social or individual commitment. In the case of social commitment, commitments sometimes have a normative character in that they involve obligations (Agnes and possibly Polly), whereas sometimes they do not (possibly Polly). Some authors, in fact, define social commitments in terms of obligations. Summarizing such views, Michael, Sebanz and Knoblich (2016a: 2) write that commitment can be understood as

> a relation among two agents and an action X, such that one agent has an obligation to some other agent to do X because she has intentionally expressed her willingness to do X under conditions of common knowledge, and this has been acknowledged.

Michael, Sebanz and Knoblich (2016a) go on to argue that there are forms of social commitment which do not involve obligations – we will come back to this in later chapters.

Individual commitment can also have a normative character but in a different sense. To see this, consider Bratman's analysis (1984; 1987) of the role of intentions in individual agency. In his analysis, intentions function to terminate practical reasoning and to structure means-end reasoning about how to achieve goals. In other words, they settle the question of what goal to pursue, and thereby enable one to move on to the subsequent question of how to go about achieving the goal. Taking one of our examples from above: Frank desires equally to go to the cinema and to the theater, but cannot do both because the performances are at the same time, so he finds it difficult to form a plan to do either. But if he forces himself to make a decision in favor of the one or the other, he forms an intention to do the one or the other. Now, he can end his deliberations and use this intention as a basis for forming a plan. In order for the intention to fulfill these functions, it has to have at least some degree of robustness: if Frank decides to go to the cinema but then, when confronted with the need to decide which metro line to take (assuming he would need to take the blue line to get to the cinema and the green line to get to the theater), he again starts deliberating about whether he prefers the cinema or the theater, and his original decision and his resultant intention will not really have served their purpose. In this sense, reconsideration would constitute a violation of norms of practical rationality. In other words, intentions are useful

in part because they involve commitment to a course of action, and because of their link to norms of practical rationality.

In contrast to these normative uses of the term 'commitment', we sometimes also use the term to refer to a psychological state – in particular, to a disposition or a motivational state. Roger, the birdwatcher, for example, is committed in this sense, meaning that he is motivated to go birdwatching whenever possible, and is disposed to invest time, money and effort in birdwatching, and to resist alternative options.

How, then, do these individual and social forms of normativity relate to commitments as psychological states? Again, we should not presume that these phenomena relate to each other in any deep or meaningful way just because we happen to use the same word to refer to them. But, over the course of this book, I hope to persuade you that it is instructive to consider the interrelations among these different forms or aspects of commitment. To be clear, I am not aiming to show that *all* forms or aspects of commitment fit together neatly as part of a single theoretical package; there may well be some forms or aspects which I am leaving out. Instead, my aim is to present a way of thinking about how social and individual forms of commitment hang together, and a way of understanding the relationship between psychological and normative aspects.

***

In the following, my aim is to provide answers to three key questions: How does social commitment relate to individual commitment? How do normative and psychological aspects of commitment relate to each other? What are the cognitive and motivational mechanisms that underpin commitment? By providing answers to these three key questions, I hope to illuminate how commitment can function as a glue holding together characteristically human forms of sociality.

The book is structured as follows. In Chapter 2, I will begin by reviewing three influential theoretical approaches to commitment: individual commitment (Bratman, 1984; 1987; 2013; 2018), social normative commitment, based on speech act theory (Austin, 1975; Gilbert, 1990; Scanlon, 1998; Searle, 1965; Shpall, 2014), and a game-theoretic approach (Frank, 1988; Schelling, 1980). Each of these three approaches provides useful insights to be incorporated into a comprehensive framework. In Chapters 3 and 4, I will present this framework. In Chapter 5, I will provide an overview of experimental research that has been conducted to investigate social factors that trigger or enhance commitment. In Chapter 6, I will discuss some

work in progress that explores the underlying cognitive and motivational mechanisms which are common to instances of individual and social commitment. Chapter 7 sketches a perspective on the development of commitment.

Note

1 There are many other distinctions that could also be drawn among these and other forms of commitment. For attempts to taxonomize heterogeneous forms of commitment, see Löhr (in prep); Michael and Pacherie (2015) and Shpall (2014).

2 A brief overview of existing approaches

2.1 Introduction

In this chapter, I briefly review three approaches to commitment. For each approach, I will consider what answers, if any, it provides to our three key questions: How does social commitment relate to individual commitment? How do normative and psychological aspects of commitment relate to each other? And: What are the cognitive and motivational mechanisms that underpin commitment?

The first of these approaches, drawing on Bratman (1984; 1987; 2013; 2018), takes individual commitment as the starting point: it attempts to conceptualize individual commitment and to use this as a basis for explaining social commitment. It conceptualizes commitment in normative terms – namely, in terms of the norms of practical rationality. It does not address psychological aspects of commitment.

The second is an approach to social commitment which draws upon speech act theory (Austin, 1975; Gilbert, 1990; Scanlon, 1998; Searle, 1965; Shpall, 2014). It also conceptualizes commitment in normative terms, but the norms in question are moral norms (or social norms in Gilbert's case). It is not obvious how it relates to individual commitment. It does not address the psychology of commitment.

The third, a game-theoretic approach, takes social commitment as the starting point (Frank, 1988; Schelling, 1980). This approach invites us to view individual commitment as a derivative of social commitment (Sperber & Baumard, 2012).

I will evaluate these approaches on their own terms, identifying strengths and weaknesses of each. In Chapter 3, I will draw upon aspects of each of these in developing a new framework.

DOI: 10.4324/9781315111308-2

2.2 Individual commitment: Bratman

Bratman's starting point (1984; 1987) is to examine the role of intentions in individual agency. In his analysis, intentions function to terminate practical reasoning and to structure means-end reasoning about how to achieve goals. In other words, they settle the question of what goal to pursue, and thereby enable one to move on to the subsequent question of how to go about achieving the goal. Taking one of our examples from the introduction: Frank desires equally to go to the cinema and to the theater, but cannot do both because the performances are at the same time, so he finds it difficult to form a plan to do either. But if he forces himself to make a decision in favor of the one or the other, he forms an intention to do the one or the other. Now, he can end his deliberations and use this intention as a basis for forming a plan.

In order for the intention to fulfill these functions, it has to have at least some degree of robustness: if Frank decides to go to the cinema but then, when confronted with the need to decide which metro line to take, he again starts deliberating about whether he prefers the cinema or the theater, and his original decision and his resultant intention will not really have served their purpose. In other words, intentions are useful in part because they involve commitment to a course of action.

However, it would also be silly to stick blindly with intentions in the face of important new information.[1] If it turns out that the metro line running to the cinema is under construction and Frank would have to walk, then maybe it makes sense for him to reconsider. In other words, intentions should not commit us *unconditionally*. As Castro and Pacherie (2020: 9) have recently pointed out, this means that intentions actually require us to perform a balancing act between pusillanimity and stubbornness. It is worth highlighting that Bratman's account does not specify any principles which would help to determine when it is rational or functional to persist and when it is not.

A further observation to make at this stage about Bratman's account is that it does not illuminate the underpinning psychological mechanisms which determine whether and to what extent we remain committed to our intentions, nor the mechanisms which then actually do sustain commitment. This is no objection; Bratman's account is not designed to illuminate these mechanisms. Be that as it may, if we are interested in understanding the psychological mechanisms underpinning commitment, we will have to look elsewhere.

What about social commitment then? Bratman (1987) points out that social context may bolster the case for resisting reconsideration in cases in which we have stated our intentions publicly because we may want to maintain our reputation as predictable, reliable agents so that others will be willing to interact with us in the future (Theriault, Young, & Barrett, 2020). Specifically, Bratman makes this point about the social dimension of commitment in relation to cases in which an intention has been publicly stated. But why should this be decisive? Roger's motivation to go birdwatching may be enhanced if other people are aware of his commitment to birdwatching, but this does not require him to have stated it publicly; it may be sufficient for other people to have seen him ostentatiously toting expensive birdwatching equipment or to have heard him pontificating about the migratory patterns of seabirds. Thus, Bratman is right that we may bolster our individual commitments by drawing other people into them. But stating them publicly is just one way of doing this, not a necessary condition. And indeed, in later work, Bratman (2013, Chapter 5) observes that joint action can lead agents to rely on each other, and that this reliance can be a source of obligations (cf. Bratman, 1997; Scanlon, 1998). Building on this idea, we may desire an account that illuminates in general terms how, when and why other people can bolster our individual commitments.

In sum, Bratman's analysis provides a clear and compelling reason for thinking that goal-directed action in general requires a certain degree of commitment. This is because we need to settle some practical questions (e.g., what goal to pursue) in order to get on to other questions (What plan to pursue? What goal to aim for next, etc.). He is also right in emphasizing the fundamental importance of commitment which his analysis reveals for agents like us, who routinely make so many interrelated plans that unfold over variable timescales. Moreover, Bratman provides the starting point for an analysis of how social commitments can build on individual commitments: our relationships with others and our reputations may be affected by the amount of commitment we exhibit. Building on Bratman's analysis, it would be desirable to identify normative principles bearing upon the question of how much commitment is appropriate, and under what circumstances more or less commitment is appropriate.[2] It would also be valuable to develop a better understanding of the psychological mechanisms by which we determine how much commitment is appropriate and by which we implement that level of commitment. Finally – and of particular interest to us here – we wish to identify the social factors that may build upon and bolster individual commitment.

2.3 Social commitment: the standard normative approach

According to a standard philosophical conception, a commitment is a relation among at least one committed agent,[3] at least one agent to whom the commitment has been made, and an action which the committed agent is obligated to perform because she has given an assurance to the second agent that she will do so, and the second agent has acknowledged this under conditions of common knowledge[4] (Austin, 1975; Gilbert, 1990; Scanlon, 1998; Searle, 1965; Shpall, 2014). I will refer to commitment in this standard philosophical sense as 'commitment in the strict sense'. For example, Susie has an obligation to Jennifer to pick up the kids from school because she (Susie) has expressed her willingness to do so, and Jennifer has acknowledged this. In the canonical case, the expression is effectuated by means of the speech act of promising. Of course, one can make a commitment (and indeed perform the speech act of promising) without explicitly saying 'I promise', but whether one says 'I promise' or simply 'yes', the expression 'will count as and will be taken as a promise in any context where it is obvious that in saying it I am accepting (or undertaking, etc.) an obligation' (Searle, 1965: 68).

This conception provides a clear characterization of paradigm cases of social commitment (i.e., commitments arising through promises or other forms of assurance), and I believe that it provides a fruitful starting point for normative discussions about commitments (Bratman, 1992, 1999; Gilbert, 1990, 2009). For example, this definition gives what I take to be a clear and satisfying answer to the question of how commitments relate to obligations. Specifically, commitments give rise to obligations over and above the obligations that one already has anyway (one has the obligation to pay one's taxes and to help drowning children where possible irrespective of any commitment one may have entered into). In other words, commitments in the strict sense are a *source* of *some but not all* obligations. Moreover, this conception also explains nicely why commitments are directed towards specific individuals (i.e., if Agnes is committed to picking up Sam at the airport tomorrow, then this is a commitment that is specifically directed towards Sam). The reason why they are directed towards specific individuals is that they are relations linking at least two agents and an action (See Roth, 2018).

In this book, however, I am interested in illuminating the cognitive and motivational processes that lead people to feel and act committed, and to expect others to do so as well. In pursuing this aim, I hope to contribute to the larger project of articulating 'a cognitive

architecture that addresses the cognitive processes enabling people to perform actions together… [one that] covers planning for immediate actions, action monitoring and action prediction, and ways of simplifying coordination' (Vesper et al., 2010: 998). My contribution to this project is to explore what role commitment may play in joint action *understood broadly*, i.e., as 'any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment' (Sebanz et al., 2006: 70; for similar definitions, see Butterfill, 2012; Vesper et al., 2010).

From a psychological point of view, the standard conception of commitment in the strict sense has a few shortcomings. For one thing, it is clearly tailored to instances in which commitments arise through promises or other verbal assurances. But what about cases in which commitments arise without any verbal exchange at all? It is all very well to say, as Searle does (see above), that commitments can be generated just by saying yes, or even just by nodding or winking or by other means, as long as it is clear from the context that one intends thereby to take on a commitment (and this is common knowledge). But what features do the context need to have in order for this to be the case?

A second shortcoming is that it does not really explain why we are (sometimes) motivated to honor commitments. It simply tells us that when a commitment is in place, we have an obligation to perform the action we are committed to performing. But it does not explain why this motivates us. And indeed, we are not always motivated to do what we are obliged to do. It would be desirable to have an explanation that illuminates why we are sometimes, but not always, motivated to honor our commitments.

Third, the philosophical conception of commitment in the strict sense lures us into a superficially appealing but inaccurate view of what happens when we consider dissolving commitment. Specifically, it leads us to think that 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. This way of thinking has informed the empirical research that has been undertaken so far concerning the dissolution of commitments (Kachel, Svetlova, & Tomasello, 2018). As Chennells and Michael (under review) have argued, however, this way of thinking does not provide a phenomenologically 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. Moreover, 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.

A fourth shortcoming is that it presents us with a binary notion of commitment: either the necessary conditions are fulfilled, and there is a commitment, or they are not, and there is no commitment. To see why this is unsatisfactory, consider the following variants of the example of Polly and Pam from the introduction to this book (also described by Michael, Sebanz, & Knoblich, 2016a: 3; adapted from Gilbert, 2009: 6):

> Polly and Pam are in the habit of smoking a cigarette and talking together on the balcony during their afternoon coffee break. The sequence is broken when one day Pam waits for Polly but she doesn't turn up. In this case, there has been no explicit agreement to smoke a cigarette and talk together every day, and yet one might nevertheless have the sense that an implicit commitment is in place, and that Polly has violated that implicit commitment. This 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, it seems that mere repetition can give rise to an implicit sense of commitment. Similarly, […] one agent's investment of effort or other costs in a joint action may also give rise to an implicit sense of commitment on the part of a second agent. If Pam, for example, must walk up five flights of stairs to reach the balcony where she and Polly habitually smoke together, Polly's implicit sense of commitment may be greater than if Pam only had to walk down the hall.

This example is intended to show that the extent to which one feels committed can be modulated by factors like the prior history of repeated interaction, and the amount of effort that one's partners have invested in a joint action. And crucially, the standard philosophical conception of commitment (i.e., commitment in the strict sense) does not explain why this should be the case, and it provides us with no basis for identifying factors (like prior history of repeated interaction and effort investment) that may have such effects.

As a sidenote: my group actually ran a series of experiments to test whether people in general share our intuitions about scenarios like the case of Polly and Pam. Specifically, we instructed participants to

imagine being in Pam's shoes in scenarios like this, and asked how annoyed they would be if Polly did not turn up, and to what extent they would think an apology to be in order. And indeed, the length of the shared history of repeated interaction, and the amount of effort invested, proved to make a difference – just as we expected (Bonalumi, Isella, & Michael, 2019).

2.4 Social commitment: a game-theoretic perspective

In the context of game theory, a commitment is a particular way of solving strategic problems where one agent would like to get a second agent to do X (or to refrain from doing X), but where that second agent is only willing to do (or refrain from doing) X if the first agent agrees to do Y (or to refrain from doing Y) (Frank, 1988; Luce & Raiffe, 1989; Nesse, 2001; Schelling, 1980). The challenge for the first agent, then, is to persuade the second agent that s/he (the first agent) really will do as s/he says at some point in time after the second agent has lost her leverage. In its purest form, commitment in the game-theoretic sense is a means of solving this kind of strategic problem by *removing the option of not acting as one has said one will act*. It is a

> device to leave the last clear chance to decide the outcome with the other party, in a manner that he fully appreciates; it is to relinquish further initiative, having rigged the incentives so that the other party must choose in one's favor.
>
> (Schelling, 1980: 37)

For example, one might attempt to win a game of chicken (a.k.a. hawk-dove) by removing the steering wheel and holding it out the window for the other driver to see, and thereby removing the option of swerving to avoid the other driver. This ploy effectively forces the other driver to make the final decision whether to collide or to swerve.

Commitments can also be successful without irrevocably removing options. It is often sufficient to alter one's incentive structure such that it would not be in one's interests to deviate from what one has said one will do. This is the rationale underlying contracts. By agreeing to sign a contract, the first agent ensures that if she does not do what s/he has promised to do, s/he faces a penalty – and the second agent, knowing that this is the case, is thereby assured that the first agent will, in fact, do as she says. But contracts are not always feasible. To borrow one of Schelling's (1980: 43–44) examples, a hostage would like to persuade his captor to release him, which the captor is in principle willing

to do (e.g., because it has become clear that the ransom will not be paid). The captor may hesitate out of fear that the hostage will testify against him. Of course, the hostage could promise not to testify, but why should the captor believe this promise? The hostage could offer to sign a contract to give teeth to the promise – but this would be of little use to the captor given that the contract would only be enforceable within the very legal system which he is eager to avoid. To solve this problem, what the hostage needs to do is to somehow change his own incentive structure such that testifying against the captor is not in his own interests. One way to do this is to confess to some other crime (or to commit some other crime), and to provide the captor with evidence of this which the captor could use against him.

A more commonplace strategy, when formal contracts are not feasible, is for one agent to put her reputation at stake by making her commitment to a second agent public (Heintz, Karabegovic, & Molnar, 2016; Michael & Pacherie, 2015). This way, if she doesn't perform the action that she committed to, she may suffer reputational costs. Down the line, of course, this is likely to have material implications insofar as she may find it more difficult to find partners for mutually beneficial cooperative endeavors, as illuminated by partner choice models of mutualistic cooperation (Barclay & Willer, 2007; Baumard, André, & Sperber, 2013). Thus, it may be in the first agent's material interest to act in accordance with her commitment even in the absence of a formal contract. In general, then, commitment in the game-theoretic sense can be thought of as 'an act or signal that gives up options in order to influence someone's behavior by changing incentives or expectations' (Nesse, 2001: 14).

This game-theoretic perspective provides a clear functional task description for commitment: a commitment is a deliberate and discrete act by which an agent changes the payoff structure of her own future options in order to convince some other agent that she will do one thing and not another at a future time point. It is worth highlighting four important features of this approach because these features will provide useful points of contrast with the framework that we will be developing here.

First, commitment in this sense is *strategic*: the first agent commits because she wants to convince the second agent that she will do something (or refrain from doing something) in order to get the second agent to do something else (or refrain from doing something).

Second, the game-theoretic approach takes social instances of commitment as its starting point – i.e., as opposed to instances of individual commitment, such as when one is committed to birdwatching or

to breaking the Coney Island hot dog eating record. Does this mean that it cannot be applied to instances of individual commitment? Not necessarily. One possibility is to consider that we can deliberately draw in other people in order to change the incentive structure of our own individual actions. For example, in order to enforce one's intention to maintain a diet, one can enter into a wager with a third party and thereby introduce an extra penalty for non-compliance (Luce & Raiffe, 1989). A further possibility is to think of cases of individual commitment as cases in which one makes a commitment to one's future self (Bryan, Karlan, & Nelson, 2010; Fudenberg & Levine, 2006; Thaler & Shefrin, 1981). Roger, the birdwatcher, might, for example, think that it is only worth resisting the temptation to sleep late this Sunday morning instead of birdwatching if he is also going to be able to resist the temptation on most future Sunday mornings (otherwise he should give up the ambition of being a serious birdwatcher anyway, and might just as well stay in bed). Having considered the matter in these terms, Roger may conclude that on future Sundays, he probably will be able to resist the temptation to sleep in often enough that it is worth investing the time and effort this Sunday, and this may contribute to his decision to get up and go birdwatching now. If so, then he may thereby put pressure on his own future self to conform to that expectation next Sunday. This would imply that the aversion to disappointing one's own expectations of oneself provides an additional motivation to act in accordance with one's commitment, lowering the net value of alternative options (i.e., skipping a day of birdwatching to sleep in). This conjecture gains face value from research showing that when people feel more strongly connected with their future selves, they tend to be more willing to forego current rewards to obtain larger rewards later in time (Bartels & Rips, 2010). But before attempting a thorough evaluation of this extension of the game-theoretic concept of commitment to individual commitment, it would be important to fill in further details. For example, we should ask whether there needs to be some equivalent of the deliberate act by which a committing agent changes her future payoff structure, and also whether there needs to be some equivalent of the strategic function of persuading some other agent to do something that she would not otherwise do. Often there seems not to be either of these things in cases of individual commitment.

A third feature I would like to highlight is that the game-theoretic account is tailored to cases in which a commitment is generated deliberately and explicitly. As a result, it does not illuminate the conditions under which commitments can arise unintentionally or gradually. To illustrate, consider the example of Polly and Pam discussed in the

previous subsection. Is it possible to map the game-theoretic conception of commitment onto these cases? There seems to have been no act by which Polly changed her payoff structure to prop up the value of the option of going to the balcony. One possibility is to think of it as the series of actions of going out onto the balcony rather than any single discrete action. But why should the repetition of the action (or Pam's investment of effort costs) increase Polly's valuation of the option of going to the balcony (or decrease the valuation of alternative options)? Some explanation would need to be provided of why these factors make a difference with respect to the reputational costs incurred by Polly if she does not show up. It is also worth highlighting that in this scenario Polly has not performed the action previously with the strategic intention of persuading Pam that she would do it in the future.

A fourth feature of this game-theoretic approach to commitment which I would like to highlight is that it conceptualizes commitment as an act with a particular pragmatic function, not as a psychological phenomenon. As a result, it is neutral with respect to the cognitive and motivational processes enabling individuals to commit or to remain committed – i.e., to resist short-term temptations and to act in accordance with commitments which optimize their long-term interests. From a normative point of view, it may be tempting to brush this psychological level aside. But it is well known that people are often tempted to make myopic decisions which fail to maximize their long-term benefits, and that they often succumb to such temptations (for an overview, see, e.g., Read, 2004). Given that it is it often tempting not to do what is in one's best interests in the long-term, how do people manage to resist such temptations?

I started out this subsection by characterizing commitment in game-theoretic terms. This provided us with a clear conception of commitment as a particular kind of problem. I also identified four distinctive features of this approach to commitment: it conceptualizes commitment as a strategic device, it takes social rather than individual commitment as its starting point, it is clearly tailored to cases of explicit rather than implicit commitment, and it is neutral with respect to the cognitive and motivational mechanisms underpinning commitment. None of these features is a problem per se, but it would be desirable to develop an account which is not restricted to instances of strategic thinking, which relates individual to social commitment, which specifies the circumstances giving rise to implicit commitment (as well as the factors which modulate the degree of commitment, as the example of Polly and Pam also illustrated), and which illuminates the cognitive and motivational mechanisms underpinning commitment.

2.5 Summing up so far

In this chapter, I canvassed three approaches to commitment, and considered what answers each might give to our key questions.

The first of these approaches was based on Bratman's (1984; 1987; 1992; 1999; 2013; 2018) theory of intentional agency. Bratman's theory conceptualizes commitment from the perspective of thinking about individual intentional agency, and accordingly about the norms of practical rationality. What I would like to take on board from this theory is the idea that it is rational to maintain some (unspecified) degree of commitment to one's intentions in general. This is simply a matter of bringing order into one's temporally extended agency – i.e., settling some questions (What goal to adopt?) in order to be able to move on to some other questions (How to achieve the goal? What other goals to adopt?). Of course, insofar as I am interested in the psychology of commitment, rationality per se is not directly relevant. However, an analysis of rationality provides benchmarks against which to compare evolved, cognitively sophisticated agents; we should expect human psychology to be equipped with mechanisms that ensure that we at least approximate rationality. I also noted that certain remarks made by Bratman provide a rough starting point for thinking about how social commitment may build on individual commitment. The idea here was that it is likely to be beneficial to maintain a reputation as someone who reliably sticks to the intentions which she adopts. Finally, I pointed out that Bratman does not address psychological aspects of commitment.

The second approach was what I take to be the mainstream conception of commitment among philosophers. It, or something quite similar, can be seen most clearly in the writings of speech act theorists and those influenced by them (Austin, 1975; Gilbert, 1990; Scanlon, 1998; Searle, 1965; Shpall, 2014), but very similar views also pre-date speech act theory (Reinach, 1913). This approach focuses on social commitments. Although I caution against thinking of social commitment as necessarily involving all the features picked out by this approach, I do think that it does a fine job of articulating paradigmatic cases of social commitment, and I believe that it is useful as a sort of a prototype concept, or as marking out one end of a spectrum of cases. Like Bratman's theory, this approach is cast in normative terms, but the norms in question are moral norms rather than norms of practical rationality. There are ways of applying it to individual commitment, but probably only in a narrow range of cases (in particular, thinking of individual commitments in terms of obligations that one has to

oneself seems like a bit of a stretch in many cases). It does not address the psychology of commitment.

The third, a game-theoretic approach, takes social commitment as the starting point (Frank, 1988; Schelling, 1980). It does not invoke norms directly, as least not as directly as the other two approaches, but it is certainly consistent with the idea that norms are at least typically at play in commitments. Moral norms may come into play in cases in which I tell someone that I will do X (and thereby put my reputation at stake), and so might norms of practical rationality insofar as maintaining a reputation for reliability is instrumental. As I have observed, this approach says nothing directly about the psychology of commitment. What I would like to take on board from the game-theoretic approach is the insight that the reward values of our action options both influence, and are influenced by, others' expectations about what we will do. In order to distill the strategic structure of social commitment as neatly as possible, the game-theoretic approach focuses on cases in which we deliberately perform discrete actions with a strategic intention. But many of the examples of social factors – discussed here in this chapter and also in later chapters – are intended to show that we do not always do this deliberately. Thus, we will have to look for a more general account to cash out this insight.

In the next chapter, I will draw upon components of these three approaches to begin to sketch a new framework for relating individual and social commitment, as well as psychological and normative aspects of commitment. The foundation of this framework is an analysis of the core function which unites many, though probably not all, instances of individual and social commitment. This functional analysis will also enable us to discern the underlying cognitive and motivational mechanisms which are common to instances of individual and social commitment, and to develop a comprehensive overview of individual and social factors that may trigger commitment. Once we have this foundation in place, we will home in on social commitment in Chapter 4.

Notes

1 I will use the expression 'new information' in a broad sense to include information about oneself or one's desires. For example, Frank may discover on this way to the cinema that he is surprisingly regretful about his decision and that he, in fact, has a deeper desire to go to the theater than he realized.

2 To motivate the notion that commitment comes in degrees, see the example of Polly and Pam in the next subsection, as well as the detailed discussion of resistance to reconsideration in Chapter 4.

3 For simplicity's sake, I will speak of one agent making a commitment. Thus, I will bracket out the interesting question whether there are any systematic differences between cases in which individuals enter into commitments and cases in which groups do so.
4 The concept of 'common knowledge' is a complex and contested one: according to more stringent analyses (e.g., Lewis, 1969; Schiffer, 1972), P is common knowledge for Susie and Jennifer if and only if Susie and Jennifer know that P, and both are in a position to know this. Thus, there is no common knowledge and accordingly no commitment in the strict sense if Susie mistakenly believes that Jennifer has not heard her assurance that she will pick up the kids, or if Jennifer mistakenly believes that Susie mistakenly believes this, etc.

3 Individual and social commitment

3.1 Introduction

In developing a new framework that illuminates the relationship between individual and social commitment, my starting point is the assumption of limited motivational integration:

> *Forming a goal does not automatically ensure that the steps which need to be taken to achieve the goal are themselves more rewarding than alternatives along the way; as a result, we are often tempted to act in ways that are inconsistent with our goals.*

In many cases, this is because our currently predominant motivation does not adequately reflect what is in our long-term interests. For example, one may be tempted to stay in bed and sleep for an extra hour rather than going for a jog or to smoke a cigarette, eat a second piece of cake or drink an extra glass of wine while out at a party. In many other cases, we simply find it difficult to stick to goals that we have adopted and to stop reconsidering alternatives which are more or less equally valuable. To illustrate, recall the example of Frank that we imagined in relation to Bratman's theory. One can imagine Frank deciding to go to the theater rather than the cinema and setting out in the appropriate direction, but then remembering how much he enjoyed the last film by the same director as tonight's film, hesitating, reversing his course, then noticing that it is too late to catch the metro to the cinema, again reversing his course, etc. What this example illustrates is that it can be psychologically difficult to shield one's goals from fluctuations in one's short-term interests and passing impulses – so difficult, in fact, that we often go to great lengths to devise means of removing alternative options before they tempt us (see the discussion of game theory in the previous chapter). In short, temporally extended agency is made difficult by the limits of our motivational integration.

DOI: 10.4324/9781315111308-3

The assumption of limited motivational integration is, in fact, supported by decades of research on reward processing and motivation. This research reveals that the regulation of motivation results from a complex interplay of distinct mechanisms which track and respond to different, imperfectly aligned indicators of value. One central distinction is that between mechanism for 'liking' and mechanisms for 'wanting'.

> "Liking" is essentially hedonic impact – the brain reaction underlying sensory pleasure-triggered by immediate receipt of reward such as a sweet taste…. "Wanting", or incentive salience, is the motivational incentive value of the same reward … "Wanting" is purely the incentive motivational value of a stimulus, not its hedonic impact.
>
> (Berridge, 2004: 194)

In the best case, liking and wanting normally go together: you want to eat when you are hungry (food has incentive value that motivates you to eat), and you like the experience of eating (you experience pleasure while eating). But they also come apart in many cases. For example, our limited capacity for affective forecasting sometimes leads to 'miswanting', or failing to accurately anticipate how much we will like or dislike something (Gilbert & Wilson, 2005). Or in addiction, people may want a drug but not take pleasure in it – i.e., not like the actual experience (Robinson & Berridge, 1993; 2003). Insofar as shorter-term goals are more driven by wants, and longer-term goals are more driven by likes, it is likely that they are particularly in need of shielding.

More generally, the assumption of limited motivational integration provides us with the core of a functional task description for commitment: to shield longer-term goals from fluctuations in our shorter-term goals and current impulses. Given this functional task description, it is apparent that there may be many reasons why a particular goal is valuable in the longer term – but these differences in the source of long-term value of goals do not entail that the mechanisms which shield goals from fluctuations in shorter-term goals or current impulses need to be different (Dreisbach & Haider, 2009; Hofmann, Friese, & Roefs, 2009; Shah, Friedman, & Kruglanski, 2002).

What are these mechanisms? Hofmann, Friese and Roefs (2009) distinguish three psychological mechanisms by which goals may be shielded: (i) attentional control is engaged to exclude information which is not relevant to the pursuit of the goal from being noticed; (ii) by exercising inhibitory control to avoid performing actions or

entertaining thoughts which are not conducive to the goal and (iii) affective control to enhance positive emotions arising from goal pursuit and to dampen negative emotions arising from resistance to temptation and distraction. These three mechanisms may work in concert or independently of each other – and indeed, if attentional and/or affective control are sufficiently effective, the demands on inhibitory control may be reduced. In addition, goals can sometimes be shielded with the help of what one might call 'situational strategies' – i.e., by avoiding situations in which one is likely to be tempted to deviate from the goal. Situational strategies recruit a different set of psychological processes, in particular prospection, forecasting and planning

Case 1: Goal selection

Roger is a birdwatcher and is out hiking in the woods. He catches a glimpse of the characteristic twinkle on the wing of a slender-billed curlew, and is inclined to chase after it. Just then, however, it occurs to him that he has lost track of where he is, and he notices that it is getting dark and chilly. He realizes that it is in his best interests to give up on birdwatching and to build a shelter to sleep in.

Case 1 illustrates something a bit different from commitment in the sense in which we have been discussing it so far – i.e. different from commitment in the sense of shielding goals. What it illustrates is the need to select goals that are in one's long-term interests in the first place. We can think of commitment coming into play in this case in the sense that Roger has a stronger commitment to the goal of staying alive than to the goal of observing this particular bird. More generally speaking, he *values* some goals more than others in the first place. There are many reasons why some goals are particularly valued over others, e.g., because of a pre-existing valuation of some broader goal, principle, type of activity, person, relationship or whatever. In such cases, it is not uncommon in everyday speech to use the term 'commitment' to refer to this pre-existing valuation (e.g., being committed to a principle, a value, an activity, a person or a relationship). We can think of commitment in these cases as a disposition to be committed to specific goals which are consistent with or serve the interests of the principle, activity, person, relationship, etc. to which one is committed.[1]

Case 2: Planning

As in Case 1, Roger has decided to build the shelter and formed a plan to collect some large branches to make a frame and a bunch of spruce boughs for the walls and roof. He is just about to set to work. But now it occurs to him that he has a bunch of clothing which could conceivably be tied together to form a makeshift tarp, which could perhaps somehow be used instead of the spruce boughs … or maybe he doesn't even need the frame if he just hangs everything from some trees … but will this really work? Instead of wasting time and energy evaluating this and the myriad other options that may occur to him, it may be wise just to stop thinking about it and get back to work. This recalls Hofmann, Friese and Roefs's (2009) notion that attentional control is engaged to exclude information which is not relevant to the pursuit of the goal.

In Case 2, in contrast to Case 1, it is not a question of selecting the appropriate goal in the first place but of resisting the temptation to reconsider what the most appropriate goal is. This is the kind of case which Bratman has in mind when he suggests that it is sometimes rational to resist reconsidering our options once we have settled on a plan. As noted earlier, in Chapter 2, Bratman does not offer any principles for determining when and to what extent it is rational or functional to resist reconsideration. One principle which at first blush seems compelling is that, when confronted with new information, one should consider whether one would have made a different decision if the new information had been available when one made the original decision in the first place. The problem with this principle is that to determine whether one would have decided differently, one needs to reconsider. In other words, the proposed principle does not really explain when one should reconsider.

To solve this problem, the aforementioned principle can be modified as follows:

> The Principle of Partial Reconsideration: *When confronted with new information, one should begin to reconsider, and then make a preliminary assessment as to how likely it seems that one would wind up with a different decision if one did re-hash the decision-making process – and continue to reconsider for as long as this likelihood exceeds a reasonable threshold.*

The principle of partial reconsideration immediately throws up two obvious questions: How high should one set the threshold for the likelihood that full reconsideration would lead to a different outcome? And: How much reconsideration should one engage in before making one's preliminary assessment? The answers to both questions depend on numerous contextual factors. A closer look at these contextual factors will enable us to formulate a series of sub-principles. These sub-principles – along with the principle of partial reconsideration – can be viewed as normative insofar as they provide guidance in determining the appropriate level of resistance to reconsideration. But they can also be viewed as hypotheses about human psychology: insofar as the normative analysis adequately captures the principles that would be most beneficial for human psychology to implement, we should expect that evolution and learning have shaped human practical reasoning to approximate those principles. Empirical research will be necessary to determine to what extent this is correct, and in what ways human psychology diverges from the normative analysis.

To begin with the first question, the appropriate threshold depends on how important it is that one make the best possible decision. Insofar as Roger's life is at stake, it is very important that he make the best possible decision. This means that even a slight possibility that a better plan might be available should be examined carefully. In contrast, if Roger were reconsidering whether to chase after the receding tune of the slender-billed curlew or to search for some other bird instead, it would not be quite as important which decision he makes. Thus, one sub-principle is that *resistance to reconsideration should be inversely proportional to the stakes*. If people operate with such a principle, we should predict that they would be less likely to consider alternatives, and would consider them less thoroughly, if stakes are low than if the stakes are high.

With respect to the second question, it is important to consider the opportunity costs of reconsideration. Most importantly, the *time* spent reconsidering could be spent implementing the current plan. This cost will be especially high in situations in which there is time pressure. This leads us to a further sub-principle: *resistance to reconsideration should be higher when there is time pressure*. If people, in fact, operate with such a principle, we should predict that they would be less likely to consider alternatives, and would consider them less thoroughly, to the extent that they are under time pressure.

Moreover, the (opportunity) costs of reconsideration will also depend on what one is doing already (to what extent the current course of action requires one's ongoing attention). Evaluation of new information may or not be possible without interfering with the current course of action. This leads us to the following sub-principle: *resistance to reconsideration should be higher to the extent that the current action requires attention*. If people operate with such a principle, we should predict that they would be less likely to consider alternatives, and would consider them less thoroughly, to the extent that the course of action they have chosen currently requires their attention.

It is also relevant that the costs and benefits of reconsidering options may be more or less certain. In Case 2, it seems that both the costs and the benefits of reconsidering are relatively uncertain. It is not obvious whether or not the alternative plan is better than the one Roger is already implementing. To determine this, Roger would need to look around a bit to evaluate the available resources and imagine going through the steps of the alternative plan. And since the alternative option begins as just a vague idea, Roger won't really know right away how much time and effort it would take to work out the details and thoroughly evaluate the idea. In other cases, however, an alternative option may appear clearly right away. For example, if Roger were to notice an apparently abandoned hut among the trees of a nearby hillside, it would not require time-consuming deliberation to determine whether to change plans: the advantages of sleeping in the hut rather than building a shelter from scratch are immediately obvious. This contrast illustrates the following sub-principle: *resistance to reconsideration should be higher when the costs and benefits of alternatives are uncertain*. If people operate with such a principle, we should predict that they would be less likely to consider alternatives, and would consider them less thoroughly, to the extent that they take the decision landscape to be uncertain or volatile.

A related point is that one may have been more or less confident in the original plan in the first place. If Roger was not really sure to begin with that the shelter he was constructing would be sufficiently warm and dry, he should be more open to considering alternatives that arise than if he had been fairly confident in his plan. This leads us to the following sub-principle: *resistance to reconsideration should be higher to the extent that one was confident*

that the original option. If people operate with such a sub-principle, we should predict that they would be less likely to consider alternatives, and would consider them less thoroughly, to the extent that they were confident when making the original decision.

Finally, it is also relevant to consider how many other plans are likely to depend on the current goal being achieved, and would therefore also need to be abandoned or revised. This leads us to a crucial sub-principle: *resistance to reconsideration should be higher to the extent that one has built on this goal in making further plans*. If people operate with such a principle, we should predict that they would be less likely to consider alternatives, and would consider them less thoroughly, to the extent that they have made further decisions or plans based upon the one that is currently a candidate for reconsideration.

This last sub-principle is particularly important for the current discussion insofar as it provides a platform for social commitment. This is because, if one has publicly selected a goal, other people may make decisions and plans based on the expectation that one will achieve that goal. In consequence of this, if one does not follow through and complete the goal, others may be disappointed and may have wasted time or other resources – an outcome which one may prefer to avoid in general in order to preserve one's relationships and one's reputation (Dana et al., 2006; Heintz et al., 2015; Székely & Michael, 2018).[2] This motivates the sub-principle that *one should resist reconsideration to the extent that others are aware of, and may be relying on, one's original decision*. If people operate with such a principle, we should predict that they would be less likely to consider alternatives, and would consider them less thoroughly, to the extent that others are aware of the original decision they have made, and all the more so to the extent that others are likely to have made decisions or plans based on this expectation.

In sum, there are a range of contextual factors that determine how much reconsideration is appropriate. In the foregoing discussion, I have not attempted to provide an exhaustive list but, rather, to identify factors which are likely to be particularly relevant in general: the stakes, time pressure, current attentional requirements, costs and benefits of alternatives, confidence regarding the original option, consequences concerning further plans, and the effects of other people's awareness of the goal-directed action.

Table 3.1 Factors influencing the appropriate level of resistance to reconsideration

Factor	*Greater resistance to reconsideration*	*Greater openness to reconsideration*
Stakes	Low	High
Time pressure	High	Low
Attentional demands of current task	High	Low
Confidence in original goal	High	Low
Certainty about costs and benefits of alternative goal	Low	High
Further goals built upon original goal	High	Low
Others relying on original goal being achieved	High	Low

Case 3: Goal pursuit

Roger has been working on the shelter for a while and gotten most of the frame up. Now, he notices some other branches that may work even better, and they are, in fact, shaped just right so that he could build a frame out of them fairly quickly and would be done just as quickly as if he continued with the frame he has been working on so far. If he had seen these in the first place, he surely would have selected them rather than the ones that he did select. But now he thinks that it would be a shame to waste the effort he has already invested in the current frame.

This is sunk cost reasoning – i.e., Roger is taking his past investment into account in deciding how to act in the future (Heath, 1995). There is some controversy as to whether sunk cost reasoning is ever rational (Kelly, 2004; Walton, 2002). Though we need not address this controversy here, one argument in defense of sunk cost reasoning is relevant for the current discussion. Specifically, a tendency to take sunk costs into account may be useful as a heuristic that functions to keep one on track when one should stay on track but might be at risk of deviating. When might this be the case?

i When one, in fact, should stay on track but is likely to form the mistaken belief that it is in one's interest to switch plans;

ii When one is tempted for the wrong reasons to switch, e.g., because it is boring or effortful to continue, and one therefore comes to experience the task as aversive (Botvinick & Braver, 2015);
iii When there is uncertainty about whether staying the course is the right choice (see Case 2 above), and one is likely to waste time and energy if one starts reconsidering.

Interestingly, Heath (1995) argues that people don't engage in sunk cost reasoning very often except under very special circumstances, namely when it is difficult to calculate or compare the costs and benefits (see (i) and (ii) above). This would be consistent with the principle, identified above, that we should avoid reconsideration to the extent that the costs and benefits of alternative options are uncertain.

Sunk cost reasoning can be distinguished from what has been called 'soft commitment' (Rachlin, 2016; Siegel & Rachlin, 1995). In soft commitment, merely beginning a behavioral pattern may increase the value of completing it, and as one progresses towards completion, the value of completion increases. To borrow an example from Rachlin (2016), a group of people playing baseball are going to be more willing to keep on playing despite a bit of rain if they have already reached the ninth inning than they would be if they had just started. Fictional examples aside, there has, in fact, been empirical research documenting soft commitment in human adults (Kivetz et al., 2006), as well as in rats (Hull, 1932) and pigeons (Siegel & Rachlin, 1995).[3] Soft commitment is different from sunk cost reasoning because the whole pattern may be rewarding rather than costly.[4] But like sunk cost reasoning, it implies a kind of 'mission creep' – i.e., the value of a goal is increased by acting towards that goal.

Why on earth would acting towards a goal increase one's valuation of the goal? Why does having played eight innings make it more attractive to keep going until the end? One hypothesis (which would apply to soft commitment as well as to sunk cost reasoning) is that one's prior actions (selecting a goal, planning and initiating goal pursuit) may indicate to oneself that one values the activity and/or the goal (Schrift & Parker, 2014). A further hypothesis (which would apply to soft commitment as well as to sunk cost reasoning) is that when one begins acting towards a goal, one tends to form other plans that presuppose

the completion of that goal. A third hypothesis is that it is taxing to maintain representations of our unfinished goals (and of the plans we have formed for achieving those goals), and we accordingly experience relief in completing goals so that we can forget about them. This may be what William James had in mind in remarking that: 'Nothing is so fatiguing as the eternal hanging on of an uncompleted task' (James, 2007). This latter hypothesis motivates the following sub-principle to the principle of partial reconsideration: *when confronted with alternatives to a current goal, one should resist reconsideration to the extent that other plans would also thereby be affected*. If people operate with such a principle, we should predict that they would be less likely to consider alternatives, and would consider them less thoroughly, to the extent that it is uncertain what other plans might be affected.

This last sub-principle reveals that commitment in this sense of 'mission creep' – i.e., valuing a goal more for having selected it, made a plan, initiated action, invested effort, made further plans that presuppose it – is useful as a default tendency for agents, like us humans, who make lots of interrelated plans that unfold over various timescales. Indeed, one might speculate that sunk cost reasoning may be a side effect of such a tendency to commit to goals more and more as a function of having already begun to progress towards completing them.

The idea of commitment as mission creep may be seen to complement Bratman's analysis by presenting a candidate psychological mechanism which implements resistance to reconsideration. This may sound strange at first blush, since I have characterized commitment as mission creep in such a way that it may be present in infants and non-human animals, whereas Bratman focuses on cases in which intentions results from conscious deliberation, and in which the decision whether or not to reopen deliberation is made consciously. But there is nothing in Bratman's (1987) analysis that is inconsistent with the idea that at least some resistance to reconsideration is implemented by basic psychological mechanisms which do not require consciousness, language or other sophisticated cognitive skills. Of course, sophisticated cognition should make it possible to calibrate resistance to calibration more precisely and flexibly, for example by considering the factors specified in the previous subsection (see Table 3.1). And indeed, this is crucially important for creatures such as us adult

humans, who have to juggle a wide range of novel goals in parallel, at various timescales. The present conjecture is that mission creep commitment is a basic mechanism implementing resistance to reconsideration which may be shared with other animals – not that it is the only such mechanism in adult humans.

Furthermore, commitment as mission creep provides a robust platform for social commitment. One reason for this is that one's investment of effort in pursuit of a goal may signal to others that the goal is worth achieving, leading them to adopt the goal as well. Indeed, this conjecture is supported by research on 'goal contagion' (Aarts, Gollwitzer, & Hassin, 2004), which suggests that when people hear or read about some other agent pursuing a particular goal, they are more likely to adopt that goal for themselves. It is also supported by research on 'goal slippage' in the developmental literature (Michael & Székely, 2017; Michael et al., under review; see also Kenward & Gredebäck, 2013; Paulus, 2014). For example, Michael et al., (under review) report evidence that toddlers' apparent 'helping behavior' is motivated by a preference to complete others' unfinished actions.

Moreover, by investing time, effort or other resources in persisting towards a goal, one indicates to others all the more clearly that one values the goal and will persist until one achieves it (just as one may do by publicly adopting a goal, as noted above). Insofar as this may strengthen their expectation that one will achieve the goal, it also constitutes an invitation to them to rely on the expectation and to plan accordingly, thereby further entrenching one's own commitment towards the goal, as well as theirs, etc. This conjecture gains credence from some recent research which we will discuss in later chapters.

Insofar as one is averse to disappointing others, then, one should adopt the sub-principle to *resist reconsideration to the extent that others have observed one acting towards a goal*. If people operate with such a principle, we should predict that they would be less likely to consider alternatives, and would consider them less thoroughly, to the extent that others have observed them performing actions in pursuit of a goal, or indeed any actions which are likely to be interpreted as indicating pursuit of a goal.

From this perspective, making a promise or simply stating one's intention to perform a particular action appears as a special case of a broader class of cases in which, by initiating

a sequence of actions which typically leads to a particular outcome, one invites others to form the expectation that one will bring about that outcome. In other words, the promise or the statement of an intention can be seen as a conventionalized first step in the sequence leading to a particular outcome. If we think in these terms, then commitment in the strict sense, as characterized by speech act theorists and philosophers such as Scanlon and Gilbert (discussed in Chapter 2), appears as a kind of limiting case on a continuum that also includes many of the fuzzier cases that we have discussed along the way. This does not imply that the analyses offered by these philosophers are invalid. Instead, it means that these analyses should be seen as characterizing a prototype rather than as providing necessary and sufficient conditions for membership in a sui generis category; in many instances, some but not all of the conditions will be met, and the resultant phenomenon will look and feel very much like a commitment.

Of course, the act of making a promise introduces norms and obligations that may otherwise be absent. The claim here is not that promising to do X is no different at all from simply starting to do X, but it can be seen as building upon this broader phenomenon of creating expectations by initiating and persisting in goal-directed action. And indeed, Bonalumi, Michael and Heintz (forthcoming) have recently shown that a sense of commitment can be elicited if one agent (the sender) has led a second agent (the recipient) to rely on her to do something, and if this is part of the two agents' common ground. Crucially, this situation can occur even if the sender has neither uttered a commissive speech act nor performed any action that would conventionally be interpreted as such.

(Duckworth, Glender, & Gross, 2016). For example, if one predicts that there will be cake at the reception after the lecture, and anticipates that one will be unable to resist the cake, one might make a point of scheduling an appointment right after the lecture so that one is not tempted to linger and indulge in the cake. Or there is what George Ainslie (2021) calls recursive self-prediction – related to his notion of choice bundling, Ainslie et al. (2003). This is the idea that I can resist a temptation now (e.g. to eat some cake) by seeing this case as a test case for a broader pattern – i.e. because I know that if I do eat the cake

now, it provides evidence to me that in the future, I will also fail to resist similar temptations. Insofar as such situational strategies solve the problem of goal shielding by removing temptations or making them less tempting, they can be regarded as analogous to contracts and to other devices described by game theorists.

Thus, cases of individual and social commitment can be seen to overlap with respect to the mechanisms engaged in stabilizing our motivation to act towards the goal. As we will see later, there are also specifically social mechanisms, which will be of special interest in this book. Moreover, individual and social commitment may be seen to differ with respect to the *source* of a goal's value. This already provides us with a partial answer to the question of how individual and social cases of commitment relate to each other: they are likely to overlap substantially with respect to the mechanisms, which they engage to stabilize motivation to act towards goals, and they differ with respect to the source of the value of the goals in question. We will return to this in Chapter 6.

For now, taking a step further, we also wish to understand how those aspects of individual and social commitment which differ relate to each other. To do this, we will need to examine how goals come to be selected, and in particular to home in on the individual and social factors which lead us to identify some goals as being valuable, and thus worth shielding from fluctuations in short-term interests and current impulses. As we shall see, individual and social factors interact with each other over the course of goal-directed action to progressively boost the valuation of goals.

3.2 Progressive goal valuation

In order to structure our examination of how some goals come to be valued and selected, it will be helpful to think in terms of a template of goal-directed action unfolding over several stages: selecting a goal, forming a plan, initiating action, persevering all the way through to completion. To be clear: this full template is not applicable in all cases. We may sometimes find ourselves acting in pursuit of a goal without having deliberately selected it or engaged in any conscious planning. Nevertheless, the full template will enable us to identify (individual and social) factors which can come into play to progressively boost goal valuation as goal-directed action unfolds – e.g., as a result of having selected, planned, initiated action towards the goal. To identify and characterize these factors, we will take one of the simple examples from the introduction and embellish it as we go along in order to

distinguish among a range of possible cases in which these different factors come into play at various stages.

3.3 Summing up so far

In this chapter, I started out from the assumption of limited motivational integration – i.e., the assumption that forming a goal does not automatically ensure that the steps which need to be taken to achieve the goal are themselves more rewarding than alternatives along the way; as a result, we are often tempted to act in ways that do not support our long-term goals. Insofar as our limited motivational integration presents an impediment to temporally extended agency, there is an important functional role for commitment as a device which shields longer-term goals from fluctuations in our short-term interests and passing impulses.

When viewed in this light, individual and social commitment can be seen to overlap substantially with respect to the basic mechanisms which they engage to stabilize motivation to act towards valuable goals but to differ in that social commitment introduces an array of additional mechanisms. Individual and social commitment can also be seen as involving distinct but complementary sources of goal valuation, i.e., as involving distinct but complementary reasons why some goals are valued and accordingly worth shielding.

To illuminate how these distinct but complementary mechanisms and sources of goal valuation relate to each other, I homed in on various stages along the way from goal selection to goal completion. This enabled me to identify various individual and social factors which can come into play to progressively boost goal valuation as a result of having selected, planned and initiated action towards the goal, etc.

Though I have been focusing on the ways in which social commitment builds upon individual commitment, this should not be taken to imply that individual commitment does not also build upon social commitment. Indeed, we did discuss some ideas along the way about how this may happen. For example, we considered the idea that I can make my individual commitments public and thereby leverage my desire for a good reputation to put pressure on myself to follow through on my commitment. More generally, I would speculate that our experiences with social commitments provide a kind of training which scaffolds the development of the skills we need to form and follow through on individual commitments. Specifically, the need to coordinate with others forces us to learn to form and stick to plans, and as we see how useful this is for coordinating with others, we import it

also into our individual planning, enabling us to achieve our desired outcomes more simply by forming plans that we can build upon. In other words, social context fosters the development of a sensitivity to the norms of practical reasoning.

This general way of conceptualizing the relationship between individual and social commitment laid out here provides the foundation for the framework for investigating social commitment. It is to this that I turn in the next chapter.

Notes

1 For more on commitment to principles, values, etc., see Sen (1977; 2002; 2005).
2 Indeed, it can be argued one has a moral obligation to avoid disappointing the expectations which one has led others to form about one's future actions expectations, in particular when others are likely to be relying on those expectations – at least when those expectations are reasonable (Scanlon, 1998).
3 One way to distinguish experimentally between soft commitment and sunk cost reasoning is to control for the amount (of money, effort, time, etc.) that has previously been invested. Soft commitment, in contrast to sunk cost reasoning, is sensitive to the distance *to the goal*, not the distance *from the starting point*.
4 It is also possible in cases of soft commitment to identify costs, such as time and energy, that have been invested during the course of goal-directed action. The point of the distinction is that in cases of commitment, the expenditure of those 'costs' is itself experienced as rewarding.

4 The sense of commitment

4.1 Introduction

In the last chapter, I introduced a way of thinking about how individual and social forms of commitment relate to each other. In particular, I proposed to conceptualize social commitment as building upon individual commitment, and introducing additional mechanisms. This also implies that it incorporates norms in the same sense as individual commitment – namely, the norms of practical rationality (being unreliable can be costly to your reputation and to your relationships with people, making it difficult to achieve some of your goals). In addition, the social dimension also brings moral norms into play: it is sometimes morally wrong to raise and then disappoint people's expectations, in particular when they are relying on one to fulfill those expectations.

Where does this leave us in terms of understanding how we identify and assess the degree of our social commitments? What situational factors play a role here, and what are the underlying cognitive and motivational mechanisms? In this chapter, I will introduce the framework that my research group has adopted in attempting to illuminate the psychology of social commitment. I will return to individual commitment later on, in Chapter 6.

4.2 The framework

Some of the components of this framework have already been developed in the previous chapter. In particular, I pointed out that by selecting and initiating pursuit towards a goal, one can sometimes lead others to form, and to rely upon, the expectation that one will complete the goal – and that this may lead one to feel committed to pursuing that goal in order to avoid disappointing others' expectations

DOI: 10.4324/9781315111308-4

and thereby undermining one's relationships and one's reputation. Building upon this starting point, the framework is structured by an analysis of the minimal structure that needs to be in place in order for a sense of commitment to emerge.

The core of this structure is illustrated in Figure 4.1. It consists of an outcome (O), two agents (ME and YOU), and at least one crucial contribution which YOU need to make in order to bring about the outcome. O may be a goal toward which ME is acting, in which case ME denotes ME's contribution to bringing about O. But O may also simply be an outcome which ME desires to be brought about, without ME having to do anything. For example, ME may desire that the grass be cut by virtue of YOU operating the lawnmower while ME does nothing.

In situations with this structure, YOU may have a sense of commitment to performing X. In the terminology I have adopted, YOU has a sense of being committed to performing X to the extent that YOU is motivated by an indication that ME expects her to contribute X and may be relying on that expectation.

A few remarks about this working definition are in order. First, the word 'indication' here is important. It is designed to encompass cases in which YOU is not sure that ME expects X – i.e., YOU does not quite have the belief that ME expects X. The reason for wanting to encompass such cases was initially a hunch that in some such cases, people do feel and act as if they believed that another agent was relying on them, even if they would not explicitly judge this to be the case. It is worth noting this leaves open the possibility that people may have a sense of commitment in interactions with robots, to whom they would not explicitly ascribe expectations (Michael & Salice, 2017; Salice & Michael, 2017). And, as we will see later on, people do, in fact, sometimes feel and act committed to robots. By defining the sense of commitment in broad terms, we can also capture these cases and study them to gain insight into the cognitive and motivational mechanisms underpinning the sense of commitment.

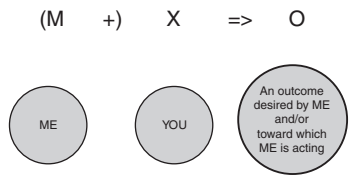

Figure 4.1 The Minimal Structure of Social Commitment.

Second, and relatedly, the definition is also designed to capture cases in which people feel and act committed even though they would not, if asked, judge that this were the case. Again, the reason for setting up the working definition in this way is that I believe there to be many such cases in everyday life. Moreover, I suspect that the underlying psychology is much the same in these cases as in cases in which one does judge oneself to be committed. To illustrate this, in addition to the example of human-robot interactions which seem to involve a sense of commitment, I sometimes like to use the following intuitive example of Sam and Woofer:

> Sam is cleaning up the living room and picks up a ball that had been lying on the floor. As it happens, his dog Woofer notices this and bounds over to him, apparently ready to play fetch. Sam was not intending to play fetch and does not particularly desire to, but may now feel obliged to, because he has generated an expectation on the part of Woofer that they will now play fetch together.
>
> (Michael, Sebanz, & Knoblich, 2016a: 5)

Third, it is worth noting that the definition makes reference both to expectations and to reliance. What role does each of these concepts play here? In a way, one could argue that it is superfluous to refer to reliance, since the minimal structure (i.e. the crucialness of X) already implies reliance. Alternatively, one could argue that it is superfluous to refer to expectations where there is reliance, since relying on YOU to do X implies an expectation that YOU will do X. However, I think that each of these two factors may make an independent contribution, and that it is therefore important to retain both of them as components in the definition. With respect to reliance, ME's reliance – in the sense of investing time, effort or other resources (including opportunity costs) on the basis of the expectation that YOU will do X – is likely to increase YOU's motivation even while keeping expectations constant. As for expectations, an indication of ME's expectation that YOU will do X may put pressure on YOU even without any further information about ME's current reliance – possibly because it indicates that ME may begin to rely on that expectation. It may well turn out that reliance only makes a difference insofar as it indicates expectations, or vice versa, or that both expectations and reliance make a difference because they indicate something else. If that turns out to be the case, the definition can be revised later. For now, in view of the prima facie independence of each factor from the other, I prefer a definition that incorporates both of these factors as independent components.

Fourth, this working definition may appear to present social commitment as something wholly distinct from individual commitment. After all, it makes explicit reference to social factors (other agents' expectations and reliance) which would not feature in cases of individual commitment. So, if we adopt this working definition of the sense of commitment to conceptualizing social commitment, where does this leave us with respect to the relationship between individual and social commitment? My working hypothesis, as intimated in previous chapters, is that they share many of the same underlying psychological mechanisms in common – namely, aspects of executive function and reward processing which will be discussed in Chapter 6 – but that social commitment involves an additional suite of mechanisms by which social factors can modulate those common underlying psychological mechanisms. The working definition of the sense of commitment is meant to capture that additional suite of social mechanisms.

Fifth, it must be highlighted that this is a functional definition: it allows us to speak of a sense of commitment whenever an agent's motivation to perform an action is increased by a cue indicating that some other agent is expecting her, and possibly relying on her, to perform that action, irrespective of what the psychological mechanism is that boosts motivation. This may sound strange, given that my stated aim is to illuminate the underlying psychological mechanisms. But the point is not for the definition itself to illuminate the underlying mechanisms. Rather, it is a working definition, the purpose of which is to structure empirical investigation that, in turn, illuminates the underlying psychology. The working definition allows us to formulate and test hypotheses about different situational factors that boost motivation in situations with this structure, and to formulate and test hypotheses about the underlying psychological mechanisms that boost motivation in situations with this structure. This procedure will allow us to develop a theory about the underlying psychological mechanisms, and potentially to replace our rough working definition later on with a more informed definition.

There may nevertheless be a concern that such a broad definition takes us *too* far away from the everyday concept of commitment (as codified, for example, in the concept of 'commitment in the strict sense', which we discussed in relation to philosophers such as Searle and Gilbert in Chapter 2). In other words, does this very broad definition lack the focus that we would need to learn anything specifically about the psychology underpinning paradigmatic cases of social commitment? In response to this concern, I propose to retain the concept of commitment in the strict sense as marking out one end

of a continuum. The cases I will be focusing on fall on a continuum ranging from the minimal structure (think Woofer, robots, etc.) all the way up to cases of commitment in the strict sense (e.g., promises and contracts). Many cases falling along this continuum will include some but not all of the features of commitment which are necessary for commitment in the strict sense. In adopting this approach, I am not denying that there are sensible boundaries to draw along this continuum but, rather, holding off on drawing them until the phenomenon is understood better.

In any event, the minimal framework is a much more useful guide to empirical investigation than the concept of commitment in the strict sense. It provides us with a basis for identifying situational factors which may give rise to or enhance the sense of commitment – i.e., any factor which raises MEs' expectation in situations with the minimal structure. Moreover, it enables us to measure the sense of commitment by measuring motivation. In the studies that my group has carried out – and which I will be discussing in the coming chapters – we have most commonly operationalized the sense of commitment in terms of agents' willingness to persist at boring and/or effortful tasks and in terms of their willingness to resist tempting alternative options. This approach enables us to discern to what extent a mechanism is in place that boosts or sustains motivation when motivation would otherwise wane – either because the costs of performing an action increase, because the rewards decrease or because the opportunity costs increase.

Since we are retaining the concept of commitment in the strict sense as a sort of limiting concept (i.e., one end of a continuum), we can also use it to identify additional features which are indicative of a sense of commitment. For example, in instances in which a commitment is violated, the agent who violated the commitment often feels guilty, and the agent to whom the commitment is owed often feels annoyed, and may either exact punishment or refrain from engaging in further interactions with the agent who violated the commitment (Barclay & Willer, 2007; Schino & Aureli, 2017). Indeed, this is often true even when the conditions necessary for commitment in the strict sense are not fulfilled. Thus, we can also use guilt and annoyance/avoidance to operationalize the sense of commitment.

4.3 Why commit?

This way of approaching commitment raises an obvious question: why would anyone be motivated to perform an action or to persist in performing it simply because some other agent seems to expect her to?

To address this question adequately, it is useful to take a step back and recall that we humans just do often help others to achieve their desired outcomes – irrespective of their expectations or reliance on us. When a stranger in front of you in line at the post office drops her pen and it rolls just out of her reach and towards you, you are likely to pick it up and hand it to her. When a fellow passenger boarding a flight is struggling to shove her suitcase into the overhead compartment, and you see that there is a smaller parcel blocking her, you will probably reach out and move that parcel to the side. The tendency to help in such situations is so deep-seated that it can be observed even in toddlers. In particular, it has been observed that infants and toddlers point to provide others with information (Liszkowski et al., 2006), and spontaneously help others to achieve their instrumental goals (Hepach et al., 2012; 2016; 2017; Svetlova, Nichols, & Brownell, 2010; Warneken & Tomasello, 2006).

Thus, the tendency to perform an action because someone else seems to expect one to, and possibly to be relying on one to do so, is, in fact, a special case of a general prosocial tendency: as a default, we tend to help others where we can (Tomasello, 2009). This does not mean that we *always* help others to achieve their desired outcomes (clearly we do not), but that, other things being equal, we at least prefer doing so over not doing so. Why would evolution have equipped us with such a general prosocial tendency, and with the more specific tendency to be responsive to others' expectations and their reliance on us?

From an evolutionary perspective, we can distinguish between two hypotheses about the evolutionary function of our default prosocial tendency. First, one may be motivated to perform actions for others in the present in order to increase one's likely future benefits – either through eliciting reciprocal prosocial behavior from them in the future (i.e., direct reciprocity; Trivers, 1971) or through a boost to one's reputation among potential interaction partners (i.e., indirect reciprocity; Nowak & Sigmund, 1998). We may call this the *strategic prosociality* hypothesis. Second, Roberts' (2005) *interdependence hypothesis* explains why one might be genuinely interested in the well-being of other group members. The interdependence hypothesis holds that humans' tendency to cooperate arose evolutionarily in a period in which our ancestors lived in small groups of individuals whose interests were largely interdependent, and for whom it was therefore not typically beneficial to act selfishly to the detriment of other group members. This implies that if an outcome is valuable to some agent with whom one is interdependent, one should value the outcome as well. As a result, any indication that one shares a valuable relationship with some

other agent (that we are interdependent) and that a particular outcome is valuable to them should lead one to value that outcome. Crucially, this line of reasoning does not depend on the expectation of reciprocity. This means that one may value O because O is in some other agent's interest even though that agent does not know this, and may mistakenly believe that some other goal would be better for her. The interdependence hypothesis therefore explains the phenomenon of paternalistic helping – i.e., that people sometimes help others by contributing to goals other than the goals currently desired by the recipient of the help (Martin & Olson, 2013; Sibicky, Schroeder, & Dovidio, 1995).

It is worth noting that the strategic prosociality hypothesis and the interdependence hypothesis are not mutually exclusive. Indeed, they are two among potentially many evolutionary mechanisms which may have jointly given rise to the sort of general default prosociality that we observe. Accordingly, these various evolutionary selection pressures are likely to have given rise to a plethora of proximate psychological mechanisms.[1]

In the literature on infant helping behavior, a number of potential proximate mechanisms have been identified. The hypothesis that has been most influential is that infants' helping behavior is motivated by an altruistic concern for the well-being of the recipient of help (Warneken & Tomasello, 2006; 2008; Warneken et al., 2007). But this is not the only hypothesis out there. A second hypothesis is that infants and toddlers who exhibit spontaneous instrumental helping behavior may do so at least in part because they like engaging in joint actions and are motivated to do so (Paulus & Moore, 2012; Rheingold et al., 1982; Svetlova et al., 2010), i.e., not because of any benefit that their contribution brings to anyone else. A further motivation for prosocial behavior is that seeing others nervous or upset (e.g., about not achieving a goal) can be aversive; thus, a third hypothesis is that infants and toddlers are motivated to help in order to avoid being exposed to an agent who is upset (Michael & Székely, 2019). This hypothesis would be consistent with the idea of a sense of commitment, i.e., it could pick out a mechanism underpinning the sense of commitment. Fourth, they may help to win praise or improve their reputation (but see Hepach, 2016). This hypothesis may also pick out a mechanism underpinning the sense of commitment. Fifth, a further class of models, which Paulus (2014) has dubbed 'goal-alignment models', is based on the core idea that the identification of an agent's goal leads infants to take up that goal as their own. This may occur because of the lack of self-other differentiation in young infants (cf. Barresi & Moore, 1996) – i.e., having identified the goal, the infant lacks the resources to quarantine it

from her own endogenous goals and simply treats it like any other goal that she has (Michael & Székely, 2019). Insofar as this latter hypothesis is correct, apparent 'helping' behavior would in fact be driven by individual commitment: infants would be helping because they themselves have the goal of completing the action and prefer to see it through.

This latter hypothesis gains credence from a study my group recently carried out (Michael et al., Under Review). We designed a paradigm in which two-year-olds could continue an adult's action when the adult no longer wanted to complete the action. The results showed that children continued the adult's actions more often when the goal had been abandoned (experimental condition) than when it had been reached (control condition), although in both conditions, it was equally feasible for the children to continue the action. This suggests that apparent helping behavior in two-year-olds is at least in part motivated by a preference for completing unfinished actions.

In the developmental literature, there is ongoing debate about which of these hypotheses best explains the earliest instances of helping behavior, and which hypothesized mechanisms arise when over the course of development. For my present purposes, it is not crucial to answer these (interesting and important) questions; what matters is that these various hypotheses probably all successfully pick out mechanisms that do arise at some point in development, and that are at work in adults.

With this in mind, we may add a further hypothesis into the mix: infants may infer that they are expected to help and/or that the helpee is relying on them, and then conform to the expectation (Bonalumi, Michael & Heintz, forthcoming; Dana et al., 2006; Heintz et al., 2015). This hypothesis, like the third (avoiding distress) and fourth (boosting one's reputation) ones mentioned above, would be consistent with the idea of a sense of commitment as we have defined it above. While there is as yet no data bearing on this hypothesis with respect to infant helping behavior, it provides a plausible explanation of the robust finding that adults tend to give away money in anonymous one-shot dictator games (i.e., when an experimenter seems to expect them to) but do not just go around handing out money in everyday life (Camerer, 2003). This explanation fits well with the findings from a classic study by Gaertner (1973), in which a confederate called people on the telephone asking for money to help him out of a difficult situation. Political liberals were more likely to help than political conservatives – but only if they stayed on the phone long enough to hear his request, and, in fact, liberals were more likely to hang up sooner. These findings support two important claims: first of all, that people have a tendency to feel

pressured into fulfilling others' expectations; and second, that they accordingly try to avoid learning of others' expectations in order to avoid being pressured into carrying out actions they do not want to carry out. More recently, Dana et al. (2006) designed a dictator game in which the participant playing the role of the dictator could pay $1 in order to exit from a situation in which they could choose either to keep $10 for themselves or to give away as much as they wanted to. Many of the participants did indeed choose the option of paying $1 to exit, but not in a condition in which they were told that the other person (the receiver) was unaware that she was a potential receiver in a dictator game. This suggests that making people aware of others' expectations makes them more likely to be cooperative.

The hypothesis that the sense of commitment is at work in these cases (at the level of proximate psychological mechanisms) is consistent both with the interdependence hypothesis and with the strategic prosociality hypothesis (at the level of evolutionary function). According to the interdependence hypothesis, an agent expecting or relying on one to do X would be a cue that one has a relationship with that agent (otherwise they would not likely form such an expectation) and that X is valuable to that agent. According to the strategic prosociality hypothesis, meeting expectations and doing what others rely on us to do is an effective way to maintain working relationships and to manage one's reputation – even if individuals do not think of it in these terms (i.e., consideration of reputation need not feature at the level of proximate psychological mechanisms).

Taken together, these reflections about the psychological mechanisms underpinning prosocial behavior, and about the evolutionary origins of those psychological mechanisms, reinforce and explain the simple everyday observation that we tend by default to contribute to bringing about other people's desired outcomes where we can – and all the more so to the extent that we already have a positive relationship with the potential recipient of help. Indeed, because we tend to behave prosocially towards each other, we also tend to *expect* it of each other. And, if the foregoing considerations are correct, these expectations introduce further pressure to behave prosocially. And, given that we also intuitively know each other to be responsive to expectations, we have a positive feedback loop: expectations of prosociality and motivations to behave prosocially mutually reinforce each other, thereby stabilizing the sense of commitment.

This last point raises an interesting possibility: a sense of commitment that detects and responds to cues of others' expectations will only be efficacious in coordinating agents' motivations and expectations

about each other's actions if it is calibrated in a sufficiently uniform manner within a social group. For example, if Polly and Pam diverge in their sense of what constitutes a good excuse for skipping the daily coffee break, or of what factors are relevant in assessing the level of commitment that is appropriate, then there is a risk that someone's expectations will be disappointed, which could threaten the harmony of their relationship. This implies that individuals whose intuitive sense of commitment is not well calibrated to their social group may find themselves frequently experiencing surprise and/or annoyance over others' failures to meet their expectations, and that their behavior may frequently be interpreted by others as evincing over- or under-commitment.

In one of the studies recently carried out in my group, Ooi et al. (2019) investigated the conjecture that personality traits characteristic of borderline personality disorder (BPD) may give rise to such disturbances of the sense of commitment. This conjecture is motivated by the observation that BPD is associated with difficulties in issues related to commitment – i.e., conflicted relationships, difficulty trusting others, fear of abandonment and patterns of over-involvement/withdrawal as well as idealization/devaluation of relationships (American Psychiatric Association, 2013). In more general terms, *impairment in interpersonal functioning* has been identified as one of the core features of psychopathology in BPD, alongside *affect dysregulation* and *behavioral dysregulation* (in particular impulsivity; Sanislow et al., 2002). We reasoned that if we could illuminate how BPD traits give rise to specific pathological disturbances of the sense of commitment, this may also help us to understand the cognitive and motivational processes leading to impairments of interpersonal functioning in BPD. And indeed, the results of our study confirmed that individuals in the general population who have high levels of the traits associated with BPD react more strongly to perceived violations of implicit and explicit commitments, and have less confidence in others to honor their commitments (Ooi et al., 2019). This provides preliminary evidence that BPD may indeed involve a disturbance in the calibration of expectations and motivations at the core of the sense of commitment.

4.4 Summing up so far

Let's take stock. In the previous chapter, I introduced a framework for conceptualizing the relationship between individual and social commitment. In particular, I suggested that we think of social commitment as building upon a more general tendency to boost our valuation

of goals as we select them and progress towards achieving them. In the current chapter, I explained how this could be applied to the investigation of social commitment. The key new concept here was that of a sense of commitment. The sense of commitment is the psychological apparatus that enables us to identify cues that some other agent is expecting and relying on us to carry out particular actions, and to respond by boosting or stabilizing our motivations to perform those actions. As such, it serves to track and respond to situations in which someone is likely to be disappointed and potentially annoyed with us if we do not perform an action which they are expecting and potentially relying on us to perform. It is important to note that the set of such situations is broader than the set of situations in which there is a commitment in the strict sense; this is why the sense of commitment can serve to explain how and when implicit commitment arises, and how and why commitments come in degrees. It is also important to emphasize that it is a psychological construct, not a normative one. When we have a sense of commitment, we may or may not judge normatively that a commitment is in place. In contrast, when a commitment in the strict sense is in place, it is by definition normative. This way of thinking leaves open the possibility that there might there be normative cases of commitment which are not captured by the concept of a commitment in the strict sense.

In the next chapter, I will review recent research that has been undertaken to test hypotheses generated by this framework, and to catalog situational factors giving rise to a sense of commitment. Once we have done that, we will return in Chapter 6 to the question of how best to characterize the psychological mechanisms underpinning the sense of commitment.

Note

1 I am appealing here to the general distinction between the evolutionary (ultimate) level of explanation and the psychological (proximate) level (Tinbergen, 1963).

5 Empirical research on the sense of commitment

5.1 Introduction

In the previous two chapters, I spelled out a theoretical framework which can be used for investigating the sense of commitment. Within this framework, the sense of commitment appears as a graded notion: the greater the sense of commitment to perform an action in a situation instantiating the minimal structure of commitment, the higher the motivation to perform that action. The framework provides us with a means of identifying situational factors that give rise to or enhance the sense of commitment: anything indicating that some other agent may expect and be relying on one to perform X should, other things being equal, increase one's motivation to perform that action.

Of course, the most straightforward way in which others may indicate their expectations and their reliance is to communicate it verbally. But, in the absence of verbal communication, many subtler features of people's behavior and of situations might also suffice to indicate this in different contexts. It would be useful to know whether there are any such features that are sufficiently frequent, and sufficiently reliable as indicators of another agent's expectations and reliance, that people may respond to them as generalized cues that the minimal structure of commitment is in place. If so, it would imply that the sense of commitment is evolutionarily and developmentally quite basic (and potentially present in other species, as well as in very young children), and foundational for communication, the understanding of norms and other aspects of social cognition. Moreover, by de-coupling the sense of commitment from language or other forms of explicit communication, it may be possible to design robots that elicit a sense of commitment on the part of human interactants. In this chapter, I will recount some of the experimental research that has been done to investigate this issue.

DOI: 10.4324/9781315111308-5

5.2 Effort

One potential feature, already flagged in the toy example of Polly and Pam discussed in earlier chapters, is effort. When some other agent invests effort in the pursuit of an outcome which requires your contribution, you can infer at least two relevant pieces of information. First, the outcome must be valuable to them. This means that they are likely to be grateful to you for helping them to achieve it, and disappointed or annoyed if you do not. Second, it means that they probably expect you to do your part – otherwise they would not waste their own effort.[1]

If effort is indeed a frequent and reliable indicator of expectations and reliance, then we should expect that when people perceive that a partner has invested considerable effort in a joint action, that they boost their motivation to reciprocate by investing effort as well, by persisting, and by resisting distractions and tempting alternatives. And indeed, arm-chair reflection seems to provide preliminary corroboration of this. Imagine, for example, that you have agreed to attend a cocktail party at your colleague's apartment but, on the occasion, find yourself tired or otherwise tempted to leave after only a short time. If your colleague has obviously invested a great deal of effort in preparing the hors d'oeuvres and decorations, you might find that a sense of commitment leads you to stick around for a few hours after all.

If this is correct, then we should expect people's persistence in a joint action to be modulated by the amount of effort which they perceive their partner(s) to have invested. In order to test this hypothesis, Marcell Székely and I (Székely & Michael, 2018) developed a two-player version of the classic 'snake game'. In the snake game, you have to navigate a snake around the screen, using the up/down and left/right arrows, to gather apples as they appear at unpredictable locations. In the classic version of the game, the task becomes increasingly difficult (and exciting) as the snake gets longer and longer; this is because you are not allowed to leave the screen and not allowed to cross over your own tail, which becomes longer and longer. In our version of the game, however, the task does not become increasingly difficult, as the snake is allowed to leave the screen (it then reappears at the opposite side) and to cross over its own tail. Instead, it becomes increasingly boring, as the apples appear at an ever-slower rate. This enabled us to measure people's sense of commitment by measuring how long they persisted (we told them that they could end each round whenever they saw fit to do so).

The other innovation we introduced was that we transformed the snake game into a two-player game: the participant controls the

left-right axis while their partner (an algorithm) controls the up-down axis. In our first experiment, participants were led to believe that their partner was a person whom they had met in the waiting area, and that, before each round of the snake game, the partner had to perform a cognitive task in order to 'unlock' the round. The cognitive task consisted in deciphering a captcha, which could be either difficult (High Effort condition) or easy (Low Effort condition). Then, the participant and the partner retrieved as many apples as possible by jointly controlling the snake – until the participant decided to end each round. We found that participants persisted longer before pressing the 'finish' button in the High Effort condition than in the Low Effort condition, implying that the apparent perception of their partner's effort boosted their sense of commitment to the task.

In one follow-up study, Matthew Chennells and I (Chennells & Michael, 2018) developed a slightly different task which not only became increasingly boring but also increasingly effortful, as participants had to repeatedly press the space bar to move a cursor from left to right. We also built in an additional performance measure so that we could assess how well they were paying attention and staying on task: at unpredictable timepoints, coins would appear, which they had to collect by pressing a separate key as quickly as possible before the coin disappeared. As in the snake study, we led participants to believe that they were playing together with a partner who had to solve either difficult or easy captchas before each round. And again, as in the snake study, we found that participants persisted longer on High Effort rounds than on Low Effort rounds before quitting. Moreover, we also found that they performed better, collecting more coins and thereby obtaining greater bonus payments for themselves and their partners.

In a different follow-up study using the snake game, we told participants that their partner was a humanoid robot, with whom they were linked via internet (Székely et al., 2019). To make it seem as real and concrete as possible, we also showed them videos of their robot partner practicing the snake game, and practicing solving captchas. Interestingly, the results showed the same pattern as we had observed when participants believed they were paired with another person: i.e., they persisted longer in the High Effort condition than in the Low Effort condition.

Our reason for doing this experiment in human-robot interaction was the following: if people exhibit the same sensitivity to a partner's apparent investment of effort when the partner is a robot as when the partner is a human, this would provide us with some insight into the mechanism underpinning this sensitivity. Specifically, it would tell us

that the mechanism in question is not particularly flexible or context-sensitive, and specifically that it may not require participants to actually believe that a partner has invested real effort, that the task really is important to a partner or that they are under any obligation to a partner. Instead, it is enough if it just seems that way. If so, this would suggest that it is a highly powerful and general heuristic, possibly a product of evolution or of extensive routinization.

Another way of trying to establish just how basic the mechanism is which mediates between the perception of a partner's effort and one's own motivation is to look at development. If such a mechanism seems to be in place in very young children, it could speak against the idea that it results from the internalization of social norms (although social norms themselves do begin to be internalized quite early – i.e., by as early as 3 or even 2, Rakoczy & Schmidt, 2013; Schmidt, Rakoczy & Tomasello, 2012). In order to begin to probe this, we (Siposova, Székely, & Michael, in prep.) implemented a first study testing whether seven- and eight-year-olds' commitment to a joint task would be greater if their partner had made a large investment than if s/he had made only a small investment. (Note: in view of the much earlier emergence of a sensitivity to social norms, finding such an effect in seven- or eight-year-olds would not tell us that the underlying psychological mechanism is independent of learned norms. Instead, this study was intended as a first step towards understanding the development of a sensitivity to partners' investment in joint actions). In this instance, the investment in question was not effort but stickers, which the partner had to pay in order to unlock the next round of a game to be played together: the stickers could be either colorful (High Cost Condition) or black-and-white (Low Cost Condition). Comparing these two conditions, we measured how vigorously children tapped the spacebar in each round (they had to tap the spacebar as quickly as possible to power a snake who navigated through a maze). The results revealed that girls, but not boys, were sensitive to our manipulation – although we may have been observing a ceiling effect with the boys, who simply enjoyed smashing away at the keys as vigorously as possible irrespective of experimental condition. Further research will be needed to probe this, also looking at different kinds of investment (e.g., effort) and at various ages.

5.3 Coordination

A second situational factor which has been investigated in the past few years is coordination. Why would coordination give rise to or enhance a sense of commitment? 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.

As a test of this idea, we ran an observational study in which we asked participants to view videos of a joint action with high and low degrees of coordination (Michael, Sebanz, & Knoblich, 2016b). In the videos, one individual was presented as having the task of cleaning up a large pile of sand, and a second individual passing by joined in because the pile was blocking his way. In the High Coordination condition, the two agents then formed a chain, with one of them scooping sand into a bucket and passing the bucket to the other agent, who emptied it into a container. In the Low Coordination condition, the two agents worked in parallel, each with his own bucket. The conditions were matched for actual effectiveness (number of overall steps taken and buckets of sand cleaned up).

Across three experiments with this general scenario, we varied a number of details, including the nature of the tempting outside option with which the helper agent was confronted: In the videos in Experiments 1 and 2 of this study, it was apparent that the pile of sand would soon be reduced sufficiently for the second agent to pass. The possibility of moving on thus presented the helper with a tempting outside option. In Experiment 3, the helper's phone rang as the video stopped, presenting a different tempting outside option (i.e., taking the call). We operationalized perceived commitment as observers' expectation that the helper would resist the option and remain engaged in the joint action. We asked for an estimate of the time the helper would remain engaged as the pile grew smaller and the way past became clear (Experiments 1 and 2) and how long the observers themselves would remain engaged in that situation (Experiment 2). In Experiment 3, we asked participants how likely they thought it was that the agent would resist the temptation to take the call, and also how likely it was that they themselves would do so if they were in that situation. As predicted, our participants judged that the helper would help longer and be more likely to resist the

temptation to take the phone call in the High Coordination condition than in the Low Coordination condition.

As with the line of experiments using the snake game to probe the effects of effort perception on commitment, we also ran a version of this study in the context of human-robot interaction (Vignolo et al., 2019). Specifically, we made another set of videos in which the helpee was a robot, and instead of a pile of sand, the robot had to clean up a bunch of toys that had been left on a desk. In this experiment, we did not find a significant effect of our coordination manipulation (chain versus no-chain). However, we also asked participants how coordinated they perceived the interaction to be, and we found a clear correlation: the more coordinated they perceived the interaction to be, the more likely they were to expect that the helper would keep helping until everything was cleaned up and to resist distractions such as a ringing phone. In sum, then, though the results from this study were not clear, I think it at least serves as a proof of concept – i.e., it encourages us to think that perceived coordination in human-robot interaction may also elicit a sense of commitment.

If this is correct, it corroborates the hypothesis that the mechanism underpinning the effects of coordination upon commitment is relatively low-level. In other words, when participants imagine being in the role of the helper, they aren't consciously inferring that there is another agent with an expectation that it would be in their interest to meet (i.e., in order to maintain a valuable relationship or for the sake of managing their reputation). Rather, they are responding in a relatively routinized manner to a general situational cue that triggers their sense of commitment.

As with the previous factor (investment of costs), we have also begun to examine children's sensitivity to coordination as a factor potentially modulating the sense of commitment – this time with four-year-olds. In this study (Reddy et al., in prep), we devised a scenario where a child plays multiple rounds of one game together with an adult experimenter, which required them to collect balls at one location and carry them over to a second location (to feed imaginary animals). During this time, a second experimenter tried to lure the child to bail out of the main game and to come and play an alternative game with her. This enabled us to measure the children's persistence in terms of how many rounds of the main game they played before succumbing to the temptation. We manipulated the degree of coordination within the main game, and also the presence or absence of ostensive eye contact with the first adult experimenter. The results of this are currently being analyzed, so it is too soon to say whether the children were sensitive

to our manipulation. Whatever the outcome, this is just one initial attempt to begin exploring the sense of commitment in young children.

5.4 Repetition

A third situational factor which we have investigated, also flagged in the example of Polly and Pam introduced in the previous chapter, is repetition – i.e., the longer the history of successful, beneficial interaction, the greater the sense of commitment to carry on with it.

Theoretically, this makes sense for several reasons. One reason is that the prior history of successful interaction with a particular agent indicates that the relationship with this other agent is valuable to you and to them (otherwise you wouldn't both have repeatedly interacted). This idea resonates with observations that have been made by economists and game theorists going back to Thomas Schelling. As Schelling put it:

> Trust is often achieved simply by continuity of the relation between parties and the recognition by each that what he might gain by cheating in a given instance is outweighed by the value of the tradition of trust that makes possible a long sequence of future agreement.
>
> (1980: 134–135)

This idea also provides a rationale for the use of so-called 'confidence-building measures' (CBMs) to facilitate cooperative interaction between parties who do not initially trust each other. For example, two countries may organize joint military training, or even more informal sport or cultural activities – such as when the United States and China began an exchange of table tennis players in the early 1970s. CBMs also work at the level of individuals in such contexts as hostage negotiation or couples therapy. One friend of my mine who used to work as a bouncer at a bar told me that he would typically implement a kind of CBM in tense situations, for example, politely introducing himself and making small talk before asking a drunk or rowdy patron to leave the premises. With the sense of commitment framework in mind, one reason why we might think that CBMs work is that they establish a pattern of positively experienced interaction that raise the expectation that future interactions will also be positive and beneficial – an expectation which the parties are then reluctant to disappoint.

In the case of CBMs, it is not essential that the pleasant pattern of activities be of the same nature as the future activities or situations

for which one is aiming to build confidence. After all, having a pleasant tradition of table tennis competitions was supposed to make it easier for China and the United States to build confidence in other areas, such as military and trade. If the activities do happen to be of the same nature, though, the establishment of a pattern can be even more effective. For example, in the context of exchange relations on an open market, it has been argued that it makes sense for agents to persist in repeating the same transactions with the same partners rather than constantly shopping around for better deals. This is because it costs time and other resources to shop around, and also because the active maintenance of longer-term relationships generates mutual commitment to ensure the further stability of the ongoing relationship (Bowles, 2016: 177–182). This 'stickiness' of exchange relations is further enhanced if the relative value of alternative deals is uncertain, as demonstrated by an economics study back in the 1990s (Kollock, 1994), and also by earlier research comparing the markets for goods the quality of which is easily to determine on the spot (rice) and for goods the quality of which becomes apparent only much later (rubber) – the upshot of which is that agents remain more committed to repeating specific transactions under more uncertain conditions (Siamwalla, 1978).

Against the background provided by the sense of commitment framework, my conjecture is that these findings generalize beyond market exchange relations to interpersonal interactions and relationships generally. Just by repeating an activity over and over again, you reduce the planning costs of continuing with the activity, and also make it possible to make other plans that build upon the activity. When it is a joint activity, you also boost the other agent's expectation that you will continue, and invite them to make plans based on this expectation. Thus, the mere repetition of a joint activity should boost the sense of commitment to that activity.

We have run a few studies to test this. The first, already mentioned in the previous chapter, was a simple vignette-based study probing people's intuitions about the relevance of repetition in the example of Polly and Pam (the two colleagues who are in the habit of having a cigarette and chatting every day during their coffee break). Sure enough, participants indicated that, if they were in Polly's shoes and Pam simply did not show up, they would be more annoyed if the activity had been repeated for a few years than if it had only been repeated for a few days, and that an apology would be more appropriate (Bonalumi, Isella, & Michael, 2019).

Following on this, in a lab-based study (Chennells et al., Under Review), we probed the effects of repeated coordination upon cooperation. To this end, we implemented a sequential joint decision-making task in which participants could choose whether or not to coordinate with a partner. We varied whether and to what degree the option not to coordinate constituted a temptation and measured the frequency with which participants chose to coordinate despite this temptation (cooperation rates). In a within-subjects design, we manipulated the partner's relationship: in one experimental block, participants played with the same partner on every trial (Fixed Partner Condition), whereas in a separate experimental block, they played with different partners on each trial (Variable Partners Condition). This made it possible to probe whether the shared history of repeated coordination in the Fixed Partner condition would lead to a higher degree of commitment (i.e., whether it would boost people's willingness to resist tempting outside offers). Crucially, the choices made by their partners could not affect them negatively, and they were informed that their partners would receive no feedback about their choices. This ensured that participants' willingness to cooperate could only be explained by commitment, not by trust or by any expectation of reciprocity. As predicted, participants' commitment (their resistance to tempting outside offers) was higher in the Fixed Partner Condition.

This provides support for the idea that, just by repeatedly engaging in a particular activity, we can bolster others' expectations that we will continue to do so, thereby inviting them to rely on this expectation – which, in turn, triggers our sense of commitment to continue or repeat the activity.

5.5 Commitment and cue integration

So far in this chapter, I have reviewed evidence that various situational factors – effort, coordination, repetition – seem to serve as cues that trigger a sense of commitment. This is all well and good as far as it goes, but it would be desirable to have a systematic understanding of how these factors relate to each other, and of what other factors there may be.

One possibility which we have begun exploring in my research group is that the various factors all boil down to evidence of a partner's effort investment (see Figure 5.1). This would be consistent with the analysis developed in Chapters 3 and 4 insofar as evidence of a partner's effort investment is evidence of their reliance – and probably of an

expectation on which that reliance is based. Coordination involves the investment of effort in the sense of adaptation: if I coordinate with you, then I invest the cognitive and physical effort required to adapt to you, presumably on the basis of an expectation about what you will do. Repetition involves the investment of effort simply because it involves repeatedly investing whatever effort is required to make one's contribution to the joint action.

Though no direct attempt has yet been made to probe this conjecture experimentally, there is some existing research that bears upon it. For example, Luke McEllin, Annalena Felber and I (under review) designed a study looking at effort and coordination. Specifically, we aimed to test what we dubbed the '*effort investment hypothesis*'. It states that successful coordination reflects a partner's willingness to invest effort into the interaction, boosting an agent's sense of commitment towards that partner. This is based on the observation that

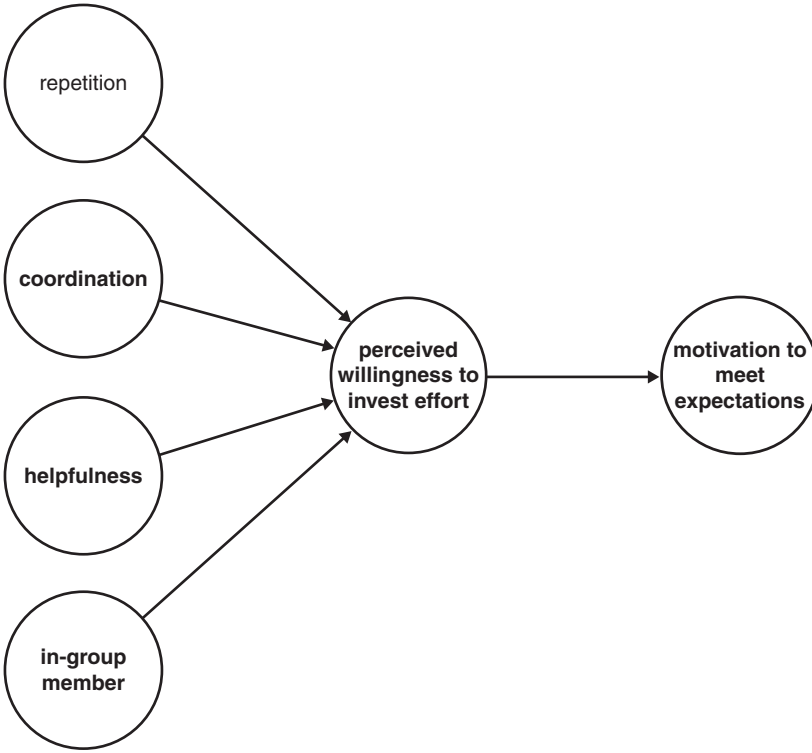

Figure 5.1 A Model of Commitment Cue Integration.

successful coordination requires agents to invest effort in order to make the interaction work. By adapting their movements or their decisions, agents facilitate alignment with their partners, and can even adapt in ways that make their actions or decisions easier for their partners to align with (Bacharach, 2006; Bardsley, Starmer, & Sugden, 2010; Keller, Novembre, & Hove, 2014; Pezzulo, Donnarumma, & Dindo, 2013; Schelling, 1960). This means that adaptation reflects an agent's willingness to invest effort into an interaction, insofar as it requires an agent to incur an individual cost in order to reduce the (e.g., planning) costs for their partner and/or to increase the chances of jointly succeeding (Green, McEllin, & Michael, 2019; Török et al., 2019).

In testing this hypothesis, one central challenge was to tease it apart from what might be called the *'similarity hypothesis*' – a hypothesis which, explicitly or implicitly, pervades much of the research on the effects of coordination upon people's prosocial attitudes and motivations towards their coordination partners (Michael, Felber, & McEllin, 2020). The similarity hypothesis states that coordination provides a cue to similarity, leading an agent to feel more committed towards the partner with whom one is coordinating. This is because those engaging in joint actions or in joint decision-making often exhibit similar movements or choices. For example, it has been suggested that cueing similarity through coordination leads us to project our own positive traits onto the agent with whom we are coordinating (Miles, Nind, & Macrae, 2009), it creates a merging between self and other (Cirelli, 2018) or that it increases awareness of our relationship as interdependent units of that group (Cross, Turgeon, & Atherton, 2019).

In most instances of coordination, adaptation as effort investment is confounded with similarity: by adapting to one another, two agents increase the similarity between their actions or decisions. In order to disentangle similarity and willingness to invest effort, we manipulated two factors separately. First, we manipulated whether the participant interacted with a partner who was adaptive, and who therefore exhibited similar actions and decisions to the participant, or with a partner who was unadaptive, and who therefore exhibited dissimilar actions and decisions to the participant. Second, we manipulated whether participants were led to believe that their partner was able or unable to adapt to them, and consequently what inferences they were likely to draw from the interaction about their partner's willingness to invest the effort required to adapt. We reasoned that by leading participants to believe that their partner was unable to adapt, we would

lead them to attribute the unadaptive partner's lack of adaptivity to an *inability to adapt*. In contrast, we expected that by leading participants to believe that their partner was able to adapt, we would lead them to attribute the unadaptive partner's lack of adaptivity to an *unwillingness to invest the effort required to adapt*. To probe the effects of these two manipulations on participants' sense of commitment, we instructed them after each trial to exert as much time and effort as they wanted to earn bonus points for their partners by pressing the space bar repeatedly.

To examine whether any effects we may find would generalize across different coordination problems, we devised two separate experiments implementing two distinct forms of coordination: action coordination and decision-making coordination. In the first experiment (looking at action coordination), participants were instructed to beat a drum in synchrony with a partner who was in a different room (they could hear the partner through headphones). They were informed in one condition that the partner could hear them, and in a separate condition that the partner could not hear them. Independently of this manipulation, we also manipulated whether or not the partner really could hear them. The idea was that if the partner could not hear them, then the partner would not be particularly adaptive – but they should only hold this against the partner if they had been (incorrectly) informed that the partner could hear them, whereas they should not hold it against the partner if they had been (correctly) informed that the partner could not hear them. The second experiment implemented an analogous setup but with a decision-making task rather than an action coordination task.

In both experiments, the results showed that participants exerted more effort and spent more time pressing the space bar for the adaptive partner than for the unadaptive partner – but only when they believed that the partner was *able* to adapt. These findings clearly support the hypothesis that the partner's investment of effort to adapt movements or decisions in order to ensure successful and smooth coordination fostered a sense of commitment towards that agent (though they are of course consistent with the possibility that perceived interpersonal similarity *also* matters).

The results of this study are particularly exciting insofar as they move us closer towards being able to tie together the various strands of research discussed in this chapter into a unifying account of the situational factors which can cue the sense of commitment. Specifically, they provide reason to be confident in the idea that what these various cues have in common is that they indicate that a partner has

expectations about what one will do, and is investing resources (in particular effort but possibly also time or other resources) on the basis of those expectations. Further research will be needed to see, for example, whether the effects of repetition can also be linked back to the investment of effort (or other resources).

It would also be useful for future research to investigate what happens when different cues are present. If there have been many repeated instances of a joint action and there is now also a high degree of effort investment from the partner, do these cues simply add up to boost the sense of commitment? What about conflicting cues – e.g., when there have been many repetitions of a particular joint action but there is a low degree of coordination? How might different cues be integrated into modulating the sense of commitment in cases like these?

5.6 Summing up so far

In this chapter, we reviewed recent research investigating situational factors that give rise to or enhance the sense of commitment. Our starting point was the idea that anything indicating that some other agent may expect and be relying on one to perform X should, other things being equal, increase one's motivation to perform that action. The findings we discussed confirmed that the perception of a partner's effort, a high degree of coordination and mere repetition may serve as cues to this effect.

We also looked at some evidence that this is even true in interactions where the partner is a robot. I suggested that this is interesting insofar as it may indicate that sensitivity to the cues of commitment is so routinized that it is relatively impervious to background knowledge. In particular, it can lead you to persist in doing something even if you do not really believe that you are committed to doing so – i.e., because after all your partner is just a robot. We also considered some ongoing research with kids which has the potential to provide converging evidence: the more we find that very young kids are sensitive to these same cues, the more we should think that the sense of commitment has been shaped by evolution as opposed to being merely a product of enculturation.

These last two strands of research – on human-robot interaction and on development – already provide a segue to the next big topic to which we now want to turn out attention: what are the cognitive and motivational mechanisms underpinning the sense of commitment? This will be the topic of the next chapter.

Note

1 This reasoning is consistent with recent theorizing about anger: According to the recalibrational theory of anger (Sell et al., 2017), people feel angrier towards individuals who have inflicted a higher cost on them (here, failed to honor a commitment in a way that had costly consequences). This would map onto the effort dimension: if I have wasted more effort, this is a higher cost for me, and I will feel angrier towards you, which should motivate you more to respect your commitment.

6 Mechanisms of commitment

6.1 Introduction

While the studies reviewed in the previous chapter provide evidence to support the hypothesis of a sense of commitment which tracks cues to other agents' expectation and reliance on one to contribute to their goals or to shared goals, the cognitive and motivational mechanisms underpinning the sense of commitment remain unclear. Indeed, the definition of the sense of commitment offered in Chapter 4, and which was assumed for the purposes of the empirical research on social commitment discussed in Chapter 5, is a functional definition: it applies to any process or mechanism which registers a cue that another agent may be expecting and relying on one to perform an action X, and responds by stabilizing the motivation to perform X. But what processes or mechanisms may actually fill this functional role? In the current chapter, I address this question by introducing a novel distinction between two forms of commitment, each associated with a different set of cognitive and motivational mechanisms: *engaged commitment* and *gritted teeth commitment*. It is worth noting at this point that, though I have been focusing on social commitment in the last two chapters, the distinction between engaged commitment and gritted teeth commitment also applies to individual commitment. This is because it pertains to underlying mechanisms which may be triggered by social or non/social factors, and which shield goals from fluctuations in short-term interests and distractions. In other words, these two distinctions are orthogonal to each other.

I begin by characterizing these two forms of commitment in intuitive terms, using everyday examples to illustrate each of them and to highlight the phenomenological differences between them (Section 6.2). I then consider the cognitive and motivational mechanisms which may underpin each, propose several ways in which the

DOI: 10.4324/9781315111308-6

two forms of commitment may be teased apart experimentally, and assess findings from previous studies which may provide evidence of either or both forms of commitment (Section 6.3). I conclude (Section 6.4) by identifying key questions for further research probing the two forms of commitment, and investigating how they may relate to each other in different contexts.

6.2 Two forms of commitment

Broadly, we may distinguish between two forms which commitment might take. The first form of commitment may be dubbed *gritted teeth commitment*. This is the form of commitment you experience when you find yourself bored or distracted, or otherwise tempted to abandon a goal, but nevertheless force yourself to persevere, and to resist temptations and distractions. For example, you may be highly committed to the task of helping your friend paint the walls of her apartment in the sense that you turn off the volume on your phone so as not to be distracted by messages, and close the windows so that you are not tempted to listen to the conversations on the street outside. The second form of commitment may be dubbed *engaged commitment*. This is the form of commitment you experience when you are so immersed in pursuing a goal that you do not notice temptations or distractions in the first place, and therefore do not need to force yourself to ignore or resist them. For example, you may be highly committed to the task of painting the apartment in the sense that you find yourself so immersed in painting that you do not even notice the messages arriving on your phone or on the conversations taking place outside the open window.

We may also apply the same distinction to commitment in the sense of commitment to people or to relationships. For example, you may be committed to your romantic partner in the (engaged) sense that you do not even notice other attractive potential partners, and therefore do not find yourself tempted to be disloyal. Alternatively, you may be committed in the (gritted teeth) sense that you do very well find yourself tempted by other attractive individuals you notice, but you overcome these temptations and remain true to your partner despite them.

Taking these everyday examples as our starting point, we can identify phenomenological differences between the two forms of commitment. If you are committed to helping your friend paint the room in the sense of gritting your teeth resisting temptations and distractions, then you are likely to find the activity unpleasant. The time will go slowly, you will catch yourself thinking ahead to what you will do afterwards, repeatedly reminding yourself to remain focused in order

to avoid mistakes (e.g., dripping paint on the floor), and you will experience negative emotions associated with the activity itself. You may find that you can best maintain your motivation by thinking of rewards and costs extrinsic to the activity – e.g., your friend will be disappointed if you abandon the task or grateful if you complete it adequately. Similarly, gritting your teeth and remaining true to your romantic partner despite tempting alternative options may give rise to a sense of conflict and stress. You may find yourself thinking of other attractive people or trying to avoid doing so.

In contrast, engaged commitment implies a positive experience of the task. The time goes by quickly, you remain focused on the task itself rather than on things you will do afterwards, and you will experience positive emotions associated with the task itself. You are able to maintain your motivation without focusing on extrinsic rewards.[1] Similarly, if you find yourself committed to your romantic partner in the engaged sense, then you are likely to find yourself thinking about your partner, and you are not likely to experience conflict at all when choosing to spend time with your partner rather than with others.

6.2.1 Is engaged commitment really a form of commitment at all?

One possible response to this intuitive distinction is to acknowledge that it captures a distinction between two ways in which an activity or a relationship can be experienced (or two extremes of a spectrum), but to wonder whether engaged commitment should really be labeled commitment at all. After all, an important function of commitment is to stabilize one's motivation to perform actions that one might not otherwise be inclined to perform (and thereby also to assure others that they can confidently rely on one to do so). The key phrase here is 'that one might not otherwise be inclined to perform'. It seems that cases of what I am calling 'engaged commitment' are cases in which one *is* inclined to perform the action, and therefore cases in which commitment would be superfluous.

This would be a serious objection if the aim of the present book were to provide an analysis of the everyday concept of commitment, or a phenomenological analysis of the experience of forcing oneself to do something despite being disinclined to do it. But my aim is to illuminate the mechanisms which shield goals from fluctuations in short-term interests. So, we should not rule out the possibility that there are instances where one experiences oneself as being inclined to perform a particular action, and where one is oblivious to distraction – and

where one's inclination to perform an action and one's obliviousness to distraction while performing the action are due to the operation of a mechanism which serves to stabilize one's motivation and to shield it from temptations and distractions. If there are such instances, their discovery would be important.

Before moving on to consider the cognitive and motivational mechanisms which may underpin each form of commitment, it is worth pausing to note that this discussion of engaged commitment (and of whether it is a form of commitment at all) resonates with a longstanding debate about commitment in behavioral economics. The debate turns on the question as to whether and how commitment can be reconciled with the theory of revealed preferences. In a nutshell, the theory of revealed preferences (Samuelson, 1938) states that people's choices are determined by their preferences, and that we accordingly can infer people's preferences from the choices that they make. With this basic idea in mind, commitment appears to present a puzzle. This is because commitment implies a willingness to perform particular actions or make particular choices *irrespective of one's current preferences*. When someone makes a commitment to do something at a later point in time, they typically mean that they will do it even if something else comes up or they just don't feel like it at that moment. In the same vein, as Amartya Sen points out:

> If knowledge of torture of others makes you sick, it is a case of sympathy; if it does not make you feel personally worse off, but you think it is wrong and you are ready to do something to stop it, it is a case of commitment.
>
> (1977: 326)

The force of Sen's observation here is that commitment falls outside of the net of the theory of revealed preferences. This is because, in instances where an agent has acted out of commitment, the method of inferring preferences from choice behavior leads to the inference that the agent must have had a preference for that action. In other words, it rules out the possibility of acting out of commitment in spite of one's preferences. This means that the theory of revealed preferences can only ever describe commitment as one preference among others. In response to this, one may conclude that the theory of revealed preferences obscures important distinctions among different motivations which can under some circumstances lead to the same choice behavior. This is Sen's conclusion; he decries "…the common tendency to make 'preference' (or a general-purpose 'utility function')

an all-embracing depository of a person's feelings, values, priorities, choices, and a great many other essentially diverse objects" (1991: 589).

Alternatively, one may respond to Sen's criticisms by affirming that, for the purposes of decision theory, commitment can be regarded as a kind of preference after all – as long as one conceptualizes 'preference' in broad terms. This is Daniel Hausman's (2005) strategy. Hausman adopts a broad concept of 'all-things-considered-preferences', which allows for decisions and actions to be driven by all manner of preferences (preferences for principles, values, others' well-being, etc., not just for what is in one's narrow interests). The project of characterizing the cognitive and motivational underpinnings of these heterogeneous preferences is an important one, but it is nevertheless useful for decision theorists to work with a single broad concept of preferences. On this view, the theory of revealed preferences does not need to be revised to accommodate the psychological complexity underpinning our decisions and actions. However, in order to develop economic theory that accurately reflects people's actual decision-making, the theory of revealed preferences should be supplemented by insights from psychology about the factors and mechanisms that drive people's actual decision-making – and the current discussion may be understood as a contribution to this project.

6.3 Teasing apart the two forms of commitment

The two forms of commitment distinguished here imply two different sets of underlying mechanisms. *Gritted teeth commitment* involves the deployment of executive control mechanisms (inhibitory control and supervisory attentional control) to maintain task focus and to avoid temptations and distractions. *Engaged commitment* boosts the relative salience and attractiveness of task-relevant information, making task-irrelevant stimuli in the environment and task-irrelevant thoughts less tempting or distracting than they otherwise would be.

Now, in terms of differentiating between these two forms of commitment, the key idea is that it should be possible to do so by probing to what extent an individual is performing a task despite being distracted and/or tempted, and to what extent they are just not distracted or tempted at all. To the extent that they are doing a task despite temptation or distraction, we can infer that she is persisting through gritted teeth. This means that if some factor (for example, the perception of a partner's effortful investment in a task) boosts an individual's persistence on a task, it should be possible to ascertain whether or not that factor boosted her persistence by triggering gritted teeth commitment.

If her increased persistence was accompanied by an increase (or at least not a decrease) in her processing of distractions or temptations, then she is in gritted teeth mode. If, on the other hand, her increase in persistence is accompanied by a decrease in her processing of distractions or temptations, then we can infer that the factor in question (e.g., the perception of a partner's effort) has elicited engaged commitment.

At present, there are unfortunately no published studies that would directly test this out. There are, however, a couple of studies in progress in my research group, and I hope that by describing them, I can at least illustrate the kind of experimental logic that may be used to differentiate between engagement and gritted teeth.

6.3.1 Probing movement trajectories

First, in a study mentioned in Chapter 5 (Reddy et al., in prep), we devised a scenario where a child plays multiple rounds of one game together with an adult experimenter, which required them to collect balls at one location and carry them over to a second location (to feed imaginary animals). During this time, a second experimenter tried to lure the child to bail out of the main game and to come and play an alternative game with her. This enabled us to measure the children's persistence in terms of how many rounds of the main game they played before succumbing to the temptation. We manipulated the degree of coordination within the main game, and also the presence or absence of ostensive eye contact with the first adult experimenter – both factors which we expected to increase the children's sense of commitment and thus also their persistence in the main game. The results of this are currently being analyzed.

But now comes the crucial part. During the pilot phase of this experiment, we noticed that the children would often not just walk directly back and forth between the two locations of the main game but, instead, would veer off towards the tempting distraction before curving back towards the relevant location in the main game. This gave us the idea that we could measure the extent to which they were distracted by measuring the curvature of their walking trajectory: the more distracted they were by the tempting alternative game, the more curved their walking trajectory should be. In principle, this should enable us to test whether the children who were more committed (i.e., who persisted longer), were more committed despite being equally or more distracted (i.e., exhibiting equal or greater curvature in their walking trajectory) or whether they were just less distracted. If the former turns out to be the case, we should infer that the children who

persisted longer did so out of gritted teeth commitment; if the latter, we should infer that they did so out of engagement.

The upshot is that measuring walking trajectory may enable us to probe whether an individual is persisting despite being distracted, or persisting because they are not distracted in the first place. It may be possible to utilize the same rationale in combination with other measures. For example, it may also be possible to get at this distinction by measuring the trajectories of people's mouse movements as they perform a computer-based task. If they have to use a mouse to drag a cursor to one of two response options (e.g., persist versus stop), we may be able to detect that in some instances participants veer towards the 'stop' option but then veer back before selecting 'persist'. By measuring this, we should be able to ascertain whether those who persist longer do so because they are less tempted to stop (straighter mouse trajectories) or despite being tempted (more curved mouse trajectories).

6.3.2 Probing attention

In principle, a study looking at attention could implement the same logic as in the aforementioned research looking at movement trajectories. That is, it could measure people's eye movements to ascertain to what extent they are attending to task-relevant information and to what extent they are exploring the environment with their attention. If the latter is the case, we could infer that they are distracted or tempted to bail on the current task. With this as a basis, we could check whether people who persist longer on some boring task do so because they are not distracted (engaged commitment) or despite being distracted (gritted teeth commitment). We have not yet tried doing this in my group, but it does seem like one option to explore moving forward.

What we have done in relation to attention in one ongoing study rests on a different rationale (Székely et al., in preparation). The study was designed to probe whether the perception of a partner's effort in a joint action boosts the deployment of executive control mechanisms (inhibitory control and supervisory attentional control) to maintain task focus and to avoid temptations and distractions. To this end, we developed a social version of the sustained attention to response task (SART) developed by Robertson and colleagues (1997; see also Manly et al., 1999; Smallwood et al., 2004). The SART is a go/no-go task: participants must respond with a spacebar press whenever a digit appears on the screen, unless that digit is a '3', in which case they must withhold their response. Since the '3' is displayed relatively infrequently (the frequency can be varied), participants are inclined to forget about

this component of the task and to fall into a routine of pressing the spacebar whenever any digit appears. Indeed, the longer the period of time/number of trials without a '3', the more likely do such false alarms become (Manly et al., 1999; Smallwood et al., 2004). In order to avoid such mistakes, participants must actively hold the task instructions in mind (indeed, it is in this sense that the SART is a test of top-down supervisory attentional control). And when participants do incorrectly respond on no-go trials, they tend to decelerate on subsequent trials and to register fewer false alarms, perhaps because they engage executive control to re-establish the task set in working memory (Manly et al., 1999; Robertson et al., 1997; Smallwood et al., 2004).

In our study, we capitalized on this phenomenon of post-error deceleration to test the gritted teeth commitment hypothesis. We reasoned that if the perception of a partner's effort boosted executive control to maintain focus on the task, then we may expect this boost to be particularly visible at those moments when it is relevant to performance on the task. If so, then we should see a more pronounced post-error deceleration in the High Effort condition than in the Low Effort condition. To manipulate the perception of a partner's effort, we used the same stimuli as in Székely and Michael's (2018) study using the snake game experiment discussed in Chapter 5 – i.e., we created the impression that the partner had to solve either easy or difficult captchas at the beginning of each round.

We elected (Székely et al., in prep) to focus on post-error deceleration as our dependent variable of interest. We reasoned that if the gritted teeth commitment hypothesis is correct, we should expect a more pronounced deceleration after errors in the High Effort condition than in the Low Effort condition, as the perception of a partner's effort investment gives a boost to executive function, helping participants to maintain or re-establish focus on a joint task. The results corroborate this prediction: participants indeed exhibited greater deceleration after errors in the High Effort condition than in the Low Effort condition, providing support for the hypothesis that the perception of a partner's effort leads participants to grit their teeth and utilize their executive resources to stay on task.

In particular, it is worth specifying that the SART is likely to target two of the three components of executive function which have been distinguished in the literature: *inhibition* of dominant or prepotent responses and *updating* working memory contents in response to changing situational demands – i.e., it is not designed to probe the capacity to *shift* flexibly between tasks (Miyake et al., 2000). To stabilize performance on boring tasks, inhibition may be especially important for

resisting distractions and temptations, and updating may be important for maintaining the task goals and strategies in working memory. These two components enable one to actively maintain task goals and task-related information when in gritted teeth mode (Dreisbach & Haider, 2009; Shah, Friedman & Kruglanski, 2002), and potentially to use this information to effectively bias lower-level processing in working memory (Baddeley, 1986; Barkley, 1997; Hofmann, Schmeichel & Baddeley, 2012).

6.4 Summing up so far

In this chapter, I introduced a distinction between two forms of commitment, each of which is associated with a different set of underlying cognitive and motivational mechanisms. One limitation of the studies I have discussed here is that they make the simple assumption that these two forms of commitment are mutually exclusive: either the intrinsic reward value of the task is enhanced (engaged commitment), in which case the need for executive control is reduced, or executive control is enhanced (gritted teeth commitment), stabilizing task focus and performance despite a reduction in the intrinsic reward value of a task. But there are also many ways in which the two forms of commitment may support each other. For example, gritted teeth commitment may work by focusing attention on aspects of a task which are rewarding, leading to an increase in the intrinsic reward value of a task. A similar but distinct idea, already mentioned in Chapter 3, is what George Ainslie (2021) calls recursive self-prediction. This is the idea that I can resist a temptation now (e.g. to eat some cake) by seeing this case as a test case for a broader pattern –i.e. because I know that if I do eat the cake now, it provides evidence to me that in the future I will also fail to resist similar temptations. This can be seen as a way of mobilizing my positive motivation for a future reward (to have a nice figure) so that I do not need to exercise as much inhibitory control to resist a temptation in the present. In this sense, it also constitutes a way of using executive function in order to get oneself into a mindset in which one needs less executive function to resist temptation. Things can work the other way around as well: by enhancing the intrinsic reward value of a task, engaged commitment may also lead to the recruitment of executive resources to stabilize or boost performance. Future theoretical and empirical research will be needed to go beyond our simple starting assumption of mutual exclusivity and to explore different possibilities concerning how the two forms of commitment may relate to each other.

To conclude this chapter, it is worth mentioning one other idea concerning how to test for gritted teeth commitment which Natalie Sebanz spontaneously suggested to me in conversation – an idea which I think was intended in jest but which actually may even work. As I was holding forth about various convoluted plans for measuring walking trajectories and obscure attentional markers, she asked flatly: Why not just put a force plate in their mouths and measure whether they are gritting their teeth?

Note

1 This is the kind of experience described by Csikszentmihalyi, M. (2014) in his seminal work on flow theory.

7 The developmental origins of commitment[1]

7.1 Introduction

In the previous chapter, I examined recent attempts to begin formulating and testing hypotheses about the mechanisms underpinning the sense of commitment. In order to illuminate these mechanisms, one further, complementary strategy is to investigate the emergence of an understanding of commitment in ontogeny, i.e., to isolate distinct components of this proficiency as they emerge, and to learn how they relate to each other, which are the most basic, etc. And indeed, there has been some research conducted in this vein, which I will survey in the current chapter. The key question for this chapter, then, is: *How do children attain a mature proficiency at identifying, keeping track of and responding appropriately to their own and others' commitments?*

If one thinks in terms of commitment in the strict sense, it may seem that there is a simple answer to this question: children acquire the concept of commitment sometime during development, and it is the mastery of this concept which underpins adults' proficiency in generating commitments, and in identifying, keeping track of and responding appropriately to one's own and others' commitments. In the following section (Section 7.2), I will evaluate this simple answer and identify theoretical and empirical reasons for finding it unsatisfactory. In Section 7.3, the main body of the chapter, I articulate and defend the hypothesis that the aforementioned proficiency rests upon an intuitive sense of commitment, which is more basic than a conceptual understanding of commitment the strict sense, and which the latter builds upon and extends. In Section 7.4, I offer some speculations about the relationship between the sense of commitment and a conceptual understanding of commitment in the strict sense. In Section 7.5, I conclude by returning to our question about the origins of characteristically human proficiency in managing commitments.

DOI: 10.4324/9781315111308-7

7.2 A simple conjecture

A simple conjecture about how children acquire proficiency with commitments is that they acquire the concept of commitment in the strict sense, as articulated in Chapter 2. According to this conjecture, possession of the concept of commitment in the strict sense leads children to act in accordance with their commitments and to otherwise acknowledge the appropriateness of censure, and to believe that they themselves are entitled to censure others who do not act in accordance with their commitments. But, though acquiring the concept of commitment in the strict sense is surely very important, there are compelling reasons to be unsatisfied with this simple conjecture. I will first identify theoretical reasons, and then turn to empirical considerations which also compel us to look beyond the simple conjecture.

7.2.1 Theoretical reasons for being unsatisfied with the simple conjecture

There are numerous features of our mature human proficiency in managing commitments that are not yet explained by appealing to the concept of commitment in the strict sense. Specifically, the concept does not clarify (a) how people determine when commitments are in place in the absence of an explicit agreement or promise, (b) how they determine what the precise content of an explicit or implicit commitment is, (c) how they assess the appropriate degree of commitment and (d) how they determine what grounds are acceptable for abandoning a commitment.

Consider the following example: Roger often volunteers as an assistant at a local retirement community. One of the residents, Patricia, is celebrating her birthday today. Roger was not explicitly invited, but he knows that Patricia would be delighted if he dropped by, and that the other people involved could use his help setting up for the party, ensuring that it runs smoothly, and cleaning up afterward. He may not have made any explicit commitment to anyone, but he may nevertheless have a sense that he is implicitly committed, either to Patricia, or to the other people involved, and this may motivate him to attend the party and to help out anyway (see (a) above). Or, he may have agreed to drop by but be surprised to discover that he is expected to help out by parking cars for the guests (see (b) above). Or, he may even have agreed to help park the cars but be surprised to discover that he is, in fact, expected to persist at this cheerless task for several hours in the hot sun (see (c) above). Or, if we tweak the example slightly, we might

also imagine that he did agree to go and help park the cars, but that he would now like to get out of this commitment because his friend has invited him to go to the pub for drinks. On the face of it, this excuse does not seem to be very compelling. But what if he suspects that his friend has been depressed and that it is important to him that they discuss something together?

Such cases are very common in everyday life. But the concept of commitment in the strict sense will not on its own be sufficient to make an appropriate judgment in such cases. This is because the concept of commitment in the strict sense provides no reason for Roger to show up to the party at all if he has not expressed his willingness to do so to any relevant party under conditions of common knowledge, and even less reason to park cars for the guests – and yet, a mature adult would often feel committed and act accordingly in such cases, and expect others to do so as well. Nor does the concept of commitment in the strict sense help in deciding which grounds for abandoning a commitment are appropriate and which are not. This means that in developing a mature proficiency in managing commitments, it is not sufficient for children to acquire the concept of commitment in the strict sense.

7.2.2 Empirical reasons for being unsatisfied with the simple conjecture

In order to evaluate the empirical credentials of the simple conjecture, it will be necessary to begin by specifying the predictions that it generates. Of course, the simple conjecture is very broad as I have formulated it. As a result, it does not entail very many specific predictions about issues for which we would, in fact, like to have specific predictions. For example, it does not entail any specific positive predictions about when children will acquire the concept of commitment in the strict sense, although it does of course predict that they will not acquire this concept before acquiring the other concepts of which it is composed, such as the concepts of 'obligation' and 'common knowledge', and possibly also the concepts of 'intention', 'belief' and 'desire,' which feature indirectly in the definition. There is evidence that one-year-olds are able to identify intentions (Behne et al., 2005). At the moment, however, it is unclear when children are able to understand the concepts of belief (Butterfill & Apperly, 2013; Carruthers, 2013; Christensen & Michael, 2016), desire (Rakoczy, 2007; Steglich-Petersen & Michael, 2015), obligation (Astington, 1988; Rakoczy et al., 2008; Vaish et al., 2011) and common knowledge (Carpenter & Liebal, 2011). In view of this uncertainty, I will not evaluate this prediction here.

The simple view also entails that once children have acquired the concept, they will exhibit a suite of behavioral tendencies which are licensed by the concept. They should, for example, be inclined to wait for a partner to whom they are committed and who is slower than they are in the context of a activity, to check on the progress of their partner(s), to offer help where appropriate, to refrain from abandoning the activity until all parties are satisfied that the goal has been achieved or until all have agreed to abandon it (Gilbert, 1990; Tuomela, 2007). They should also be inclined to censure others who violate explicit verbal agreements to perform actions, and acknowledge others' rights to censure *them* if they themselves do so (Gilbert, 1990). This may take the form of explicitly censuring and explicitly acknowledging others' right to censure, or it might take a more implicit form. For example, they may be inclined to cry and/or to protest if agreements are violated but without explicitly stating the reason why. Similarly, they may exhibit signs of guilt or of fearing punishment when they themselves violate agreements. The crucial point is that the simple conjecture predicts that once children acquire the concept of commitment in the strict sense, there should be an uptick in these behaviors because these behaviors are licensed by commitments, as one would understand by grasping the concept. What do the data show?

Gräfenhain and colleagues (2009) implemented a paradigm in which an experimenter and a child play various games together. In experiment 1 of their study, they were interested in how children would react when, at some point, the experimenter abruptly stopped playing. Specifically, they compared a condition in which the experimenter had made an explicit commitment to the joint action and a condition in which she had simply entered into the action without making any commitment. Interestingly, three-year-olds, but not two-year-olds, protested significantly more when a commitment had been violated than when there had been no commitment. In experiment 2 of the same study, the tables were turned and the children were presented with an enticing outside option that tempted *them* to abandon the joint action. The children were less likely to succumb to the temptation if a commitment had been made. In cases in which they did succumb, they were more likely to 'take leave', to look back at the experimenter nervously or to return after a brief absence.

The interpretation of these findings suggested by the simple conjecture is that children acquire the concept of commitment in the strict sense by around three. But consider a study conducted by Mant and 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 behavioral tendencies may emerge first (waiting for a partner, checking on her, helping her, persisting until all parties are satisfied that the goal has been reached, protesting if a partner abandons a joint action, etc.). In other words, the simple conjecture presents a complex concept and a suite of behaviors licensed by the concept as a single package. But these components may come apart, and some may be more basic than others. The simple conjecture does not tell us in what order these components should emerge, which components are most basic, or how the developmental process should unfold.

Moreover, there is a further detail in the findings reported by Gräfenhain and colleagues which should give us pause. Specifically, it is not the case that the two-year-olds do not protest at all, and only the three-year-olds understand the situation well enough to feel entitled to protest. In fact, there is no increase in appropriate normative protest from two to three. On the contrary, the two-year-olds protest just as much in both conditions as the three-year-olds do in the commitment condition. This suggests that the sense of entitlement that inspires protest over an unfulfilled expectation is not the product of developmental changes over the third year but, rather, it is the default

that is already in place by two or earlier. What changes in the third year is that children learn that they are not always entitled to expect contributions to their goals.

There is also a further detail in Mant and Perner's (1988) study which bears emphasizing: 22 of the 46 six-year-olds actually rated the protagonist as being naughty in both conditions (while 11 rated him as neutral in both conditions), i.e., when Peter had violated a commitment and thereby caused Fiona to be disappointed and sad, and when he had not made any commitment in the first place and Fiona had been disappointed and sad. It is as though, whenever a goal is not achieved and somebody is left disappointed, the default is to assign blame, and to work out the details later. This is not the pattern that one would expect on the basis of the simple conjecture. This is because the simple conjecture predicts that normative protest emerges as a result of the understanding that one is entitled to protest because a commitment in the strict sense is in place.

I propose to develop a different approach to explaining the developmental trajectory of children's proficiency in identifying, keeping track of and responding appropriately to our own and others' commitments. Rather than taking the concept of commitment in the strict sense as a starting point, and interpreting the findings of Gräfenhain and colleagues (2009; 2013; cf. Hamann et al., 2012; Kachel & Tomasello, 2019; Kachel, Svetlova, & Tomasello, 2018) as evidence that three-year-olds understand and respond to commitments in the strict sense, I will attempt to identify a broader, less complex phenomenon that young children may understand and respond to even in the absence of a sophisticated understanding of common knowledge, obligations and the speech act of promising. My aim will be to explain how an understanding of commitments emerges through engagement in joint actions, as several distinct cognitive and affective mechanisms are integrated and calibrated through social experience. This more psychological approach (i.e., in contrast to an approach based on normative notions) resonates with the view of many theorists that a simplified conception of joint action is needed in order to account for young children's engagement in joint actions (Brownell, 2006; Butterfill, 2012; Tollefsen, 2005).

7.3 The development of the sense of commitment

In theorizing about the 'broader, less complex phenomenon' that children are progressively able to identify and respond to, I will draw

upon Michael, Sebanz, and Knoblich's (2016a) analysis. As already discussed in Chapter 4, this analysis provides a characterization of the minimal structure of situations in which a *sense of commitment* can arise:

i There is an outcome (O) which an agent (ME) either desires to come about, or which is the goal of an action which ME is currently performing or intends to perform.
ii The external contribution (X) of a second agent (YOU) is crucial to bringing about G.

Clearly, conditions (i) and (ii) specify a broader category than that of commitment in the strict sense. Nevertheless, situations with this structure may elicit a sense of commitment on the part of YOU, and it may lead ME to expect commitment from YOU. My proposal now is that children first acquire a sense of commitment, and that this sense of commitment is gradually calibrated through social experience to give rise to a mature proficiency in managing commitments. In order to spell out this proposal, I will first need to explain how a sense of commitment would arise in the first place. To answer this, it will be useful to consider these kinds of expectation and these kinds of motivation as separable components, and to ask: Why would children, or indeed anyone at all, have such expectations and/or motivations? Next, I will need to explain how the sense of commitment could develop into a mature proficiency in managing commitments.

My attempt to meet these challenges will consist of three steps, which I will discuss in the next three subsections:

1 There are numerous mechanisms leading humans (and possibly in some cases other species as well) *to be motivated* to contribute X in situations in which the minimal structure is instantiated (i.e., (i) and (ii) obtain), and some of these mechanisms are present already in infancy.
2 There are numerous mechanisms leading humans (and possibly other species as well) *to expect* X to occur because (i) and (ii) obtain.
3 These expectations and motivations reinforce each other over time, and are calibrated through joint actions and other social experiences, leading children ultimately to a mature proficiency in identifying, keeping track of and responding appropriately to their own and others' commitments.

7.3.1 How would YOU come to be motivated to do X because the minimal structure is instantiated?

This question already came up in Chapter 4 (in the subsection 'Why commit?'). There, I referred to six distinct hypotheses about why agents, including even infants and toddlers, may have such motivations.

First, they may be motivated by an altruistic concern for the well-being of the recipient of help (Warneken & Tomasello, 2006; 2008; Warneken et al., 2007). Second, infants and toddlers may help because they like engaging in joint actions and are motivated to do so (Paulus & Moore, 2012; Rheingold et al., 1982; Svetlova et al., 2010), i.e., not because of any benefit that their contribution brings to anyone else. Third, seeing others nervous or upset (e.g., about not achieving a goal) can be aversive; thus, infants and toddlers may be motivated to help in order to avoid being exposed to an agent who is upset (Michael & Székely, 2019). Fourth, they may help to win praise or improve their reputation (but see Hepach, 2016). Fifth, they may help because they think they are expected to, and are motivated by a preference to avoid disappointing expectations (Heintz et al., 2015). Sixth, a further class of models, which Paulus (2014) has dubbed 'goal-alignment models', are based on the core idea that the identification of an agent's goal leads infants to take up that goal as their own. This may occur because of the lack of self-other differentiation in young infants (cf. Barresi & Moore, 1996) – i.e., having identified the goal, the infant lacks the resources to quarantine it from her own endogenous goals and simply treats it like any other goal that she has (Michael & Székely, 2019).

Developing a version of this sixth hypothesis, Michael and Szekely (2019) introduced the term 'goal slippage'. On their account, goals that are identified in instances instantiating the minimal structure are sometimes represented as motor representations within the observer's motor system – namely, when the observed action is in their own motor repertoire. When this occurs, the identified goal becomes the observer's own goal, and the observer will automatically act to bring about the identified goals unless some other mechanisms inhibit their automatic action. For example, YOU may observe as ME attempts to toss a pillow onto a seat in the row in front of her on an airplane, and notice that the pillow, unbeknownst to ME, has rolled onto the floor. In such as case, YOU may pick up the pillow and place it on the seat in order to facilitate the achievement of the goal. Although an agent's motivation to bring about such goals may generally be lower than her motivation to bring about endogenously generated goals, goal slippage could nevertheless increase the likelihood of YOU doing X. As noted already in the discussion in Chapter 4, one recent study in our group

(Michael et al., under review), provides support for this hypothesis as an explanation of helping behavior in two-year-olds.

To be clear, though there is lively debate in the developmental literature as to which of these hypotheses is correct, I do not aim at present to adjudicate among them. On the contrary, I would tend to think that they are all true. Certainly they are all true as at least partial explanations of some instances in which the minimal structure is in place and *adults* do X. What we don't know yet with much confidence is at what age which motivations become operative. In any case, I regard these as separate, complementary factors which may conspire to sustain a default preference on the part of YOU to contribute X when she detects that a situation with the minimal structure is in place.

7.3.2 How would ME come to expect YOU to do X because the minimal structure is instantiated?

I believe that there are numerous reasons why infants in the role of ME tend to expect YOU to do X in cases instantiating the minimal structure. At the most basic level, the expectation that the desired outcome O will occur when desired may have the status of a default in infants. This is because an infant may not entertain the possibility that O is only her own desired outcome (Piaget, 1950). A default expectation that O will occur when desired would be consistent with many experiences that infants and young children have in their first years of life. Indeed, as soon as infants begin pursuing goals, there is usually at least one parent who is motivated to support them in their goals. Moreover, infants experience distress or conflict when their goals are not met.

Our hypothesis is that this default expectation of O is progressively qualified over the course of development – i.e., it becomes increasingly context-specific as children develop more sophisticated abilities to understand the instrumental structure of action, to evaluate agents and to identify and integrate more and more relevant factors which are relevant to predicting whether X will occur. A first step beyond the very basic default expectation which I have proposed is to identify specific agents that are associated with successful outcomes. For example, an infant may come to associate mommy with good outcomes, and thus expect O to occur specifically when Mommy is present. A further important step is to be able to identify the specific external contributions (X) which are required for their desired outcomes. For example, Billy may come to notice that in order to bring about the goal of feeding him (O), Mommy needs to grasp the bottle and present it to his mouth (X).

When Billy has attained this level of sophistication, his default assumption will be that those contributions (X) will be made. And instances in which he does not meet a goal because X is not contributed may also elicit signs of distress and/or conflict. Moreover, as Billy gets older, he will also need to learn to evaluate many more factors, such as whether Mommy is aware of his desire to eat, whether it is reasonable to expect her to feed him now (which would, for example, not be the case if Mommy is currently driving in heavy traffic), whether she has made a promise to feed him at this moment in particular, etc.

One possibility raised by this view of development is that this bedrock sense of entitlement remains into adulthood, usually below the surface of behavior. Indeed, I suspect that this is the case, and that this default attitude can be glimpsed in those moments when one is stressed or tired and, struggling to tie one's shoe or to close a drawer, catches oneself cursing at the shoe or the drawer and feeling inclined to mete out punishment to whatever objects or agents happen to be around. My conjecture is that, psychologically, this sense of outrage and frustration is the very same sense of outrage and frustration as what one experiences when there really is an agent who is to blame for some normative violation.

Be that as it may, such a default expectation – suitably qualified on the basis of knowledge gained through social experience – could generate or reinforce specific expectations that ME would not otherwise have about contributions (X) to be made to ME's goals or to outcomes which ME desires to be brought about.

But on top of this basic default expectation, there are many further reasons for ME to expect YOU to do X – namely, the very same reasons why YOU is in fact often motivated to do X (as I set out in the previous section). Specifically, YOU is motivated by such mechanisms as altruism, a collectivity preference, an aversion to others' distress and an aversion to disappointing others' expectations. Of course, very young children will not be aware of these reasons, but people's tendency to act on the basis of such motivations will buttress their expectation that YOU will typically perform X. And, as they do become aware of such reasons over the course of development, their expectations will become increasingly accurate.

7.3.3 Expectations and motivations reinforce each other over development

In the previous two subsections, I gave reasons why some agents, in particular infants and young children, may expect X to occur because

(i) and (ii) obtain. I also gave reasons why some agents, in particular infants and young children, may sometimes be motivated to contribute X because (i) and (ii) obtain, and also sometimes because they believe that they are expected to. In this section, I will explain how these expectations and motivations can reinforce each other over the course of development, and how the sense of commitment can thereby become calibrated to the norms within a culture.

On the one hand, ME's default expectation that others (such as YOU) will contribute to ME's goals will be likely to be met and reinforced if other agents (such as YOU) are indeed likely to contribute because of the processes referred to in the previous two subsections. On the other hand, YOU will be more likely to contribute X if YOU believes that ME expects this.

This does not imply that children (or, for that matter, adult humans) always expect others to contribute X in situations instantiating the minimal structure, nor that they always contribute X when they think they are expected to. In many such instances in which an agent expects X, X simply does not occur. Indeed, even infants' and young children's parents don't always support their goals or fulfill their desires. So, as noted already above, in order to differentiate among various degrees of likelihood that X will occur, children must develop a more nuanced sensitivity to features of interactions that carry information about the reliability of various kinds of cues to X in various situations.

Is YOU aware of ME's expectation of X? Did YOU do anything to cause ME to have this expectation? If so, was this intentional? Is there any precedent for this expectation? That is, has YOU made the contribution of X in previous similar situations? If, for example, Daddy has played catch with Leonardo every Saturday in the garden for many months, it is more reasonable to expect this to occur this Saturday than if Daddy had only done it once or twice. Similarly, it is also important to assess to what extent ME is relying on X for the achievement of the outcome O. If X is something that can really only be achieved with YOU's contribution, and if it is very important, then it is less appropriate for YOU to refuse unless there is a good reason. Leonardo, for example, can play with some of his toys alone if Mommy is busy, but his new wiffle ball bat is only fun to play with if someone pitches the wiffle ball for him to swing the bat at – so it is more reasonable to expect Mommy to play together with him, and all the more so if he needs to practice for a wiffle ball game at his friend's birthday party the following day.

Moreover, through social experience over many years, children also learn when it is appropriate to abandon or postpone commitments

(Bonalumi et al., Under Review; Chennells & Michael, Under Review; Kachel, Svetlova, & Tomasello, 2018; Michael & Székely, 2018). For example, if daddy promises to take Leonardo to the zoo but then has to rush off to work to deal with an urgent matter, Leonardo will need to understand that daddy's urgent matter provides a good reason to postpone the zoo trip until the following day.

By the same token, it would be inefficient for an agent *always* to contribute to others' goals or desired outcomes whenever she believed that she were expected to. Hence, children must also learn to apply the same criteria in determining whether to make crucial contributions to others' goals or desired outcomes as they apply in determining whether to expect others to make those contributions. And, more generally speaking, the processes which I have postulated as underpinning a sense of commitment are likely to become calibrated through experience to match those of other people in their culture, and to conform to cultural norms concerning when it is considered appropriate to make contributions to others' goals and to expect contributions from others. As a result, people's expectations about the extent to which others will be motivated by such processes will roughly match the extent to which others really are so motivated – unless they happen to suffer from an impairment in their sense of commitment, as may be the case for individuals with borderline personality disorder, as discussed in Chapter 4 (Ooi et al., 2019).

7.4 What about the simple conjecture?

So far, I have given an account of how various sources of motivation and of expectations reinforce each other over the course of development. Through this long process of mutual reinforcement, expectations are calibrated such that children come to have increasingly accurate expectations about when others will perform actions which are contributions to outcomes which they desire or towards which they are acting. Similarly, motivations are calibrated such that children come to be motivated to make contributions when they are expected to – and particularly when it is important to others that they do so, and particularly when the other person in question is one with whom it is important to maintain a good relationship. The upshot of this account is that proficiency in generating commitments, and in identifying and tracking the degree of one's own and others' commitments, crucially involves managing expectations about contributions to goals and desired outcomes.

Judging by the research reviewed in Chapter 5, this process of calibrating expectations and motivations crucially involves becoming sensitive to cues such as a partner's investment of effort, coordination and repetition. As I explained in Chapter 5, we have begun to chart the development of children's sensitivity to these cues (Reddy et al., In Prep; Siposova, Székely, & Michael, In Prep), but the findings from this research so far are just too scant to really support any conclusions.

Where does all this leave the concept of commitment in the strict sense and the simple conjecture? Mastering the concept of commitment in the strict sense does not appear to be necessary in order to identify and respond to such expectations on the part of others, or to have such expectations about others. Nor does it appear to be sufficient in order to (a) determine when commitments are in place in the absence of an explicit agreement or promise, (b) determine what the precise content of an explicit or implicit commitment is, (c) assess the appropriate degree of commitment or (d) distinguish between good and bad reasons for abandoning commitments. However, this does not make the concept of commitment in the strict sense irrelevant. On the contrary, once the concept of commitment in the strict sense is bootstrapped out of the more basic sense of commitment, there are several important functions which it can serve.

For example, mastery of this concept makes it possible to quickly and efficiently engage the machinery of expectations and motivations that I have been attempting to illuminate here. Doing this proficiently, however, also requires that one's expectations and motivations are properly calibrated to begin with. For example, if Orsi gives Vanda an assurance that she will clean up every mote of dust that ever falls onto his car, he is unlikely to form the expectation that she will actually do this, because it is simply not a realistic suggestion. Similarly, if she requests after their first date that Vanda promise to be forever true, it might well have the opposite effect, because it is an unreasonable request, and indeed one which exhibits an alarming lack of social skill.

Moreover, the concept may help in various ways to facilitate the calibration of motivations and expectations that I have been discussing. For example, the concept of commitment in the strict sense highlights some features of situations that are relevant to determining whether ME can reasonably expect YOU to do X, such as whether YOU did something to generate this expectation in ME, whether this was intentional and whether it is common knowledge that this is the case. These are not the only relevant factors but they are among the relevant factors. So, if daddy promises to give Leonardo some ice cream after

dinner and then only gives him a single scoop, and Leonardo begins to cry and protest about this, daddy may point out to him that he promised to give him only a bit and was never intending to suggest that he would give him any more – Leonardo will have to calibrate his expectations downward about what 'some ice cream' means.

7.5 Summing up so far

We humans are quite proficient in generating commitments and in identifying, keeping track of and responding appropriately to our own and others' commitments. In the current chapter, I have attempted to shed light on the cognitive processes underpinning this proficiency by examining the emergence of a proficiency in managing commitments in ontogeny.

One unsurprising general conclusion to draw is that humans, armed with the concept of commitment and with the language skills to make verbal agreements and otherwise to form and communicate detailed plans about future behavior, are highly adept at generating expectations, which others can rely on. It would also be unsurprising if some of the source of the motivation to fulfill those expectations are uniquely human.

One perhaps surprising consequence of my account is that a very powerful source of motivation to fulfill those expectations, and basis for expecting others to do so as well, is, in fact, the product of a very basic tendency that is present early in ontogeny and likely shared with other species –namely, a tendency to become frustrated and angry if our goals are not met and the outcomes we desire not achieved. Specifically, my account generates a novel claim about the origins of the sense of entitlement that inspires protest over unfulfilled expectations, i.e., unfulfilled expectations about one's goals being met and about the outcomes one desires coming about. In contrast to the hypothesis suggested by the simple conjecture, my account suggests that this sense of entitlement to protest is not the product of developmental changes by which one acquires the concept of commitment in the strict sense but, rather, it is the default that is already in place by two years or earlier. What changes over the course of childhood is that children learn that they are not always entitled to expect their goals to be met or all contributions to their goals to be made.

In one sense, this means that the developmental process chips away from, rather than adding to, the cognitive architecture that underlies normative protest. In a different sense, however, the developmental process of course also involves the addition of further cognitive

machinery. For example, an increasingly sophisticated understanding of the instrumental structure of action will make it possible to better identify when there is a contribution (X) which some other agent could make to bringing about one's goal. As a result, older children (and adults) will also protest or otherwise be annoyed in situations in which younger children (and cognitively less sophisticated animals) would not even notice that another agent has failed to help them. To probe the development of the capacity to identify crucial external contributions to one's goals, and to compare humans and non-humans with a view to illuminating the phylogeny of this capacity, one useful starting point would be a paradigm implemented by Plotnik et al. (2011). In this paradigm, the target agent (an elephant) needs to pull on a rope in order to retrieve a platform with food on it – but in order to be successful, must wait for a second agent to arrive on the scene and pull on the other end of the rope (the crucial external contribution). In Plotnik and colleagues' study, it was shown that elephants can understand the crucial role of a partner in a task with this structure, and accordingly wait before pulling the rope until the partner can do so as well. Building upon this, it would be interesting to test whether elephants (and other species) but be annoyed at a partner who fails to pull on the rope or refrain from cooperating with her/him on future occasions.

Note

1 Some of the material in this chapter is adapted from Michael and Székely (2018).

8 Further directions

8.1 The three key questions

The overarching aim of this book has been to illuminate three key questions about commitment: How does social commitment relate to individual commitment? How do normative and psychological aspects of commitment relate to each other? What are the cognitive and motivational mechanisms that underpin commitment? By providing answers to these three key questions, I hope to have gone some way to explaining how commitment can function as a glue holding together characteristically human forms of sociality.

I started out by acknowledging the fact that 'commitment' is not only a vague and complex concept but also a heterogeneous one – we can distinguish between individual and social forms of commitment, and we can speak about commitment in normative or in psychological terms. I have tried to be careful not to assume that, just because we use the same word – 'commitment' – to these heterogeneous phenomena, they really do have anything deep and interesting in common. And yet, I have also tried to show that it is, in fact, fruitful to look for a core function which unites many, though probably not all, instances of commitment: namely, to shield long-term goals from fluctuations in short-term interests and impulses.

In spelling out this way of thinking, I have drawn upon insights gleaned from existing accounts which target either individual or social commitment specifically. For example, Bratman's account provided us with the important thought that there is some value in settling matters so that we can get on with planning and acting. So, even when confronted with two possible goals which are otherwise equally valuable (e.g., going to the cinema or to the theatre), it is rational to make a choice and stick to it, ignoring further information and resisting the impulse to reconsider. In this sense, *the selection and pursuit of a goal*

DOI: 10.4324/9781315111308-8

itself boosts the subjective valuation of the goal. This is a crucial feature of temporally extended agency: it enables us to form plans (since we would otherwise not know which goals to plan towards) and to adopt further goals that presuppose the fulfillment of the first goal (e.g., meeting a friend for a drink after the movie at the bar near the cinema).

Bratman's account also provided us with a basis for understanding how social commitment can build on individual commitment. Bratman himself only notes that social context may bolster the case for resisting reconsideration in cases in which we have stated our intentions publicly, because we may want to maintain our reputation as predictable, reliable agents so that others will be willing to interact with us in the future. But, as revealed by the research presented here, this is just one of many ways in which social context can build upon and enhance individual commitments.

To illuminate these multifarious influences of social context on individual commitment, I drew upon a game-theoretic conception of commitment. Game theory explains how we can persuade others to do what we want them to do by changing our own incentive structures (in extreme cases even removing options for ourselves altogether which would otherwise be too tempting to resist). In order to distill the strategic structure of social commitment as neatly as possible, the game-theoretic approach focuses on cases in which we deliberately perform discrete actions with a strategic intention. But many of the examples of social factors discussed in this book show that this does not always occur deliberately. Sometimes, just by selecting a goal, making a plan, initiating an action, investing effort, etc., we raise others' expectations about what we will do and lead them to rely on those expectations – *whether or not we intended to influence them.* In many such cases, our realization of the expectations we have caused others to form and to rely on gives rise to a sense of commitment to carry on pursuing the goal whether we want to or not. Indeed, as we select particular goals, construct plans to achieve them, initiate action and persist towards goal completion, various social factors (reviewed in Chapters 4 and 5) pile up and progressively buttress our commitment.

It may initially seem mysterious why evolution should have designed us to progressively boost our valuation of goals as we pursue them. But this seems less mysterious if we consider the possibility that it is a consequence of the evolution of dispositions that make it possible for us to simultaneously pursue innumerable long-term (individual and shared) goals that build upon and complement each other, unfolding over variable timescales. After all, we cannot always use enforceable

contracts to stick to our goals and to sustain cooperation (and in deep evolutionary history this would have been all the less viable). So, instead we evolved proximate mechanisms which increasingly boost our valuation of goals as we select them, make plans for attaining them and act towards them. Moreover, it also makes good sense that we should have developed a sense of commitment which makes us responsive to other people's expectations about and reliance upon our future actions: this is because the sense of commitment enables us to rely on each other to plan and carry out complex joint actions unfolding over various timescales. In this regard, the sense of commitment can be seen as an amalgam of proximate mechanisms for managing one's reputation and one's relationships, and thereby stabilizing cooperation over time.

Although I have been focusing on how social commitment builds upon individual commitment, this is not intended to deny that social commitment can also structure and transform individual commitment. Indeed, we did discuss some ideas along the way about how this may happen. For example, we considered the idea that I can make my individual commitments public and thereby leverage my desire for a good reputation to put pressure on myself to follow through on my commitment. In addition to this, I would speculate that our experiences with social commitments provide a kind of training which scaffolds the development of the skills we need to form and follow through on individual commitments. Specifically, the need to coordinate with others forces us to learn to form and stick to plans, and as we see how useful this is for coordinating with others, we import it also into our individual planning, enabling us to achieve our desired outcomes more simply by forming plans that we can build upon. In this sense, social context may help to scaffold the development of practical reasoning.

In addition, social commitments also introduce normative aspects, which do not apply to individual agency. Most obviously, the voluntary creation of commitments typically also generates entitlements and obligations (Gilbert, 1990; 2006a; 2006b). Two points are in order about this. First, according to the perspective I have been presenting here, however, cases in which it makes sense to speak of entitlements and obligations are just the tip of the iceberg. Beneath the surface of normative discourse, there is a much broader set of cases constituted by the more general practice of managing and meeting expectations. And in many such cases, talk of obligations and entitlements is just too heavy-handed (Michael & Butterfill, Under Review). Rather, people's actions and decisions in many such cases are better described in subtler terms like 'thoughtfulness/thoughtlessness' or '(in-)considerateness'.

Second, the normative dimension of social commitment does not just fall down from the sky; it grows out of and is continuous with practical rationality. It makes sense from the perspective of practical rationality to be able to promiscuously create and rely on expectations because this enables us to act much more efficiently together than we otherwise could. And once we do this, it also makes sense to avoid disappointing expectations as a proximate mechanism for managing one's reputation and one's relationships. Moreover, this introduces a normative perspective insofar as it can be morally wrong to disappoint others' expectations, particularly insofar as they are relying on those expectations (Scanlon, 1998).

8.2 Outlook

At the outset, I began by lamenting that so little progress has been on understanding the psychology of commitment because the concept itself is so inchoate and heterogeneous. This motivated the project of engaging in some conceptual house-cleaning, coming up with a way of thinking that encompasses most cases of commitment while also doing justice to the diversity of phenomena picked out by the term. I believe that the framework resulting from that conceptual house-cleaning has proven fruitful so far, insofar as it has structured and constrained research into the factors giving rise to commitment, and on the mechanisms underpinning commitment. Moreover, it has opened up new avenues for research on the development of commitment and of a conceptual understanding of commitment, as well as on potential pathologies of commitment such as Borderline Personality Disorder. However, it is important to emphasize that the theoretical framework I have developed here is little more than a set of provisional working definitions and a way of relating them to each other – not intended as a complete theory but as a platform on which to base research which, inevitably, will also lead to theoretical refinement. Ultimately, it should be judged not on how true it is but how productive it has been.

Bibliography

Aarts, H., Gollwitzer, P. M., & Hassin, R. R. (2004). Goal contagion: Perceiving is for pursuing. *Journal of Personality and Social Psychology*, *87*, 23–37.

Ainslie, G. (2021). Willpower with and without effort. *Behavioral and Brain Sciences*, *44*, 1–57.

Ainslie, G., & Monterosso, J. R. (2003). Building blocks of self-control: Increased tolerance for delay with bundled rewards. *Journal of the Experimental Analysis of Behavior*, *79*(1), 37–48.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders*, 5th Edn. American Psychiatric Publishing.

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

Austin, J. L. (1975). *How to do things with words* (Vol. 88). Oxford University Press.

Bacharach, M. (2006). *Beyond individual choice: Teams and frames in game theory*. Princeton University Press.

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1610), 749–753.

Bardsley, N., Mehta, J., Starmer, C., & Sugden, R. (2010). Explaining focal points: Cognitive hierarchy theory versus team reasoning. *The Economic Journal*, *120*(543), 40–79.

Barresi, J., & Moore, C. (1996). Intentional relations and social understanding. *Behavioral and Brain Sciences*, *19*(1), 107–122.

Bartels, D. M., & Rips, L. J. (2010). Psychological connectedness and intertemporal choice. *Journal of Experimental Psychology: General*, *139*(1), 49.

Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(1), 59–78.

Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology*, *41*(2), 328.

Berridge, K. C. (2004). Motivation concepts in behavioral neuroscience. *Physiology & Behavior*, *81*(2), 179–209.

Bonalumi, F., Isella, M., & Michael, J. (2019). Cueing implicit commitment. *Review of Philosophy and Psychology*, *10*(4), 669–688.

Bonalumi, F., Michael, J., & Heintz, C. (2021). Perceiving commitments: When we both know that you're counting on me. *Mind and Language*, *2021*, 1–23. http://dx.doi.org/10.1111/mila.12333

Bonalumi, F., Siposova, B., Christensen, W., & Michael, J. (under review). Should I stay or should I go? Three-years olds' sensitivity to appropriate motives to break a commitment.

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: From behavior to neural mechanism. *Annual Review of Psychology*, *66*, 83–113.

Bowles, S. (2016). *The moral economy: Why good incentives are no substitute for good citizens*. Yale University Press.

Bratman, M. E. (2018). *Planning, time, and self-governance: Essays in practical rationality*. Oxford University Press.

Bratman, M. E. (2013). *Shared agency: A planning theory of acting together*. Oxford University Press.

Bratman, M. E. (1999). *Faces of intention: Selected essays on intention and agency*. Cambridge University Press.

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

Bratman, M. (1987). *Intention, plans, and practical reason* (Vol. 10). Harvard University Press.

Bratman, M. (1984). Two faces of intention. *The Philosophical Review*, *93*(3), 375–405.

Brownell, C., Ramani, G., & Zerwas, S. (2006). Becoming a social partner with peers: Cooperation and social understanding in one- and two-year-olds. *Child Development*, *77*(4), 803–821.

Bryan, G., Karlan, D., & Nelson, S. (2010). Commitment devices. *Annual Review of Economics*, *2*(1), 671–698.

Butterfill, S. (2012). Joint action and development. *The Philosophical Quarterly*, *62*(246), 23–47.

Butterfill, S., & Apperly, I. (2013). How to construct a minimal theory of mind. *Mind and Language*, *28*(5), 606–637.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Carpenter, M., & Liebal, K. (2011). Joint attention, communication, and knowing together in infancy. *Joint Attention: New Developments in Psychology, Philosophy of Mind, and Social Neuroscience*, 159–181.

Castro, V. F., & Pacherie, E. (2020). Joint actions, commitments and the need to belong. *Synthese*, 1–30.

Chennells, M., & Michael, J. (forthcoming). Breaking the right way: A closer look at how we dissolve commitments. *Phenomenology and the Cognitive Sciences*.

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.

Chennells, M., Wozniak, M., Lindeløv, J., Butterfill, S., & Michael, J. (under review). Coordinated Decision-Making Boosts Altruistic Motivation – But Not Trust.

Cirelli, L. (2018). How interpersonal synchrony facilitates early prosocial behavior. *Current Opinion in Psychology*, *20*, 35–39.

Cross, L., Turgeon, M., & Atherton, G. (2019). How moving together binds us together: The social consequences of interpersonal entrainment and group processes. *Open Psychology*, *1*(1), 273–302.

Csikszentmihalyi, M. (2014). Play and intrinsic rewards. In *Flow and the Foundations of Positive Psychology* (pp. 135–153). Springer, Dordrecht.

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.

Dreisbach, G., & Haider, H. (2009). How task representations guide attention: Further evidence for the shielding function of task sets. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 477.

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2016). Situational strategies for self-control. *Perspectives on Psychological Science*, *11*(1), 35–55.

Elliot, A. J., & Harackiewicz, J. M. (1996). Approach and avoidance achievement goals and intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology*, *70*, 461–475. http://dx.doi.org/10.1037/0022-3514.70.3.461

Fessler, D. M., & Quintelier, K. (2013). Suicide bombers, weddings, and prison tattoos: An evolutionary perspective on subjective commitment and objective commitment. *Cooperation and Its Evolution*, 459–484.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. WW Norton & Co.

Fudenberg, D., & Levine, D. K. (2006). A dual-self model of impulse control. *American Economic Review*, *96*(5), 1449–1476.

Gaertner, S. L. (1973). Helping behavior and racial discrimination among liberals and conservatives. *Journal of Personality and Social Psychology*, *25*, 335–341. https://doi.org/10.1037/h00 34221

Gilbert, M. (2009). Shared intention and personal intentions. *Philosophical Studies*, *144*(1), 167–187.

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

Gilbert, M. (2006b). *A theory of political obligation*. Oxford University Press.

Gilbert, M. (1990). Walking together: A paradigmatic social phenomenon. *MidWest Studies in Philosophy*, *15*, 1–14.

Gräfenhain, M., Behne, T., Carpenter, M., & Tomasello, M. (2009). Young children's understanding of joint commitments. *Developmental Psychology*, *45*(5), 1430–1443.

Gräfenhain, M., Carpenter, M., & Tomasello, M. (2013). Three-year-olds' understanding of the consequences of joint commitments. *Public Library of Science One*, *8*(9): e73039. https://doi.org/10.1371/journal.pone.0073039

Green, A., McEllin, L., & Michael, J. (2019). Does Sensorimotor Communication Stabilize Commitment in Joint Action?: Comment on "The body talks: Sensorimotor communication and its brain and kinematic signatures" by G. Pezzulo et al. *Physics of Life Reviews*. https://doi.org/10.1016/j.plrev.2019.01.009

Hamann, K., Warneken, F., & Tomasello, M. (2012). Children's developing commitments to joint goals. *Child Development*, *83*(1), 137–145.

Hausman, D. M. (2005). Sympathy, commitment, and preference. *Economics & Philosophy*, *21*(1), 33–50.

Heath, C. (1995). Escalation and de-escalation of commitment in response to sunk costs: The role of budgeting in mental accounting. *Organizational Behavior and Human Decision Processes*, *62*(1), 38–54.

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), 442–455.

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

Hepach, R. (2017). Prosocial arousal in children. *Child Development Perspectives*, *11*(1), 50–55.

Hepach, R., Vaish, A., Grossmann, T., & Tomasello, M. (2016). Young children want to see others get the help they need. *Child Development*, *87*(6), 1703–1714.

Hepach, R., Vaish, A., & Tomasello, M. (2012). Young children are intrinsically motivated to see others helped. *Psychological Science*, *23*(9), 967–972.

Hofmann, W., Friese, M., & Roefs, A. (2009). Three ways to resist temptation: The independent contributions of executive attention, inhibitory control, and affect regulation to the impulse control of eating behavior. *Journal of Experimental Social Psychology*, *45*(2), 431–435.

Hull, C. L. (1932). The goal-gradient hypothesis and maze learning. *Psychological Review*, *39*(1), 25.

James, W. (2007). *The principles of psychology* (Vol. 1). Cosimo, Inc.

Kachel, U., Svetlova, M., & Tomasello, M. (2018). Three-year-olds' reactions to a partner's failure to perform her role in a joint commitment. *Child Development*, *89*(5), 1691–1703.

Kachel, U., & Tomasello, M. (2019). 3-and 5-year-old children's adherence to explicit and implicit joint commitments. *Developmental Psychology*, *55*(1), 80.

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

Kelly, T. (2004). Sunk costs, rationality, and acting for the sake of the past. *Noûs*, *38*(1), 60–85.

Kivetz, R., Urminsky, O., & Zheng, Y. (2006). The goal-gradient hypothesis resurrected: Purchase acceleration, illusionary goal progress, and customer retention. *Journal of Marketing Research*, *43*(1), 39–58.

Knoblich, G., Butterfill, S., & Sebanz, N. (2011). Psychological research on joint action: Theory and data. In *Psychology of Learning and Motivation* (Vol. 54, pp. 59–101). Academic Press.

Kollock, P. (1994). The emergence of exchange structures: An experimental study of uncertainty, commitment, and trust. *American Journal of Sociology*, *100*(2), 313–345.

Liszkowski, U., Carpenter, M., Striano, T., & Tomasello, M. (2006). 12-and 18-month-olds point to provide information for others. *Journal of Cognition and Development*, *7*(2), 173–187.

Löhr, G. (In prep). Normative and non-normative uses of commitment.

Luce, R. D., & Raiffa, H. (1989). *Games and decisions: Introduction and critical survey*. Courier Corporation.

Manly, T., Robertson, I. H., Galloway, M., & Hawkins, K. (1999). The absent mind: Further investigations of sustained attention to response. *Neuropsychologia*, *37*(6), 661–670.

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

Martin, A., & Olson, K. R. (2013). When kids know better: Paternalistic helping in 3-year-old children. *Developmental Psychology*, *49*(11), 2071.

McEllin, L., Felber, A., & Michael, J. (under review). The fruits of our labour: Coordination generates commitment by signaling a willingness to adapt.

Michael, J., & Butterfill, S. (under review). Intuitions about joint action. https://doi.org/10.17605/OSF.IO/W2JDV

Michael, J., & Christensen, W. (2016). Flexible goal attribution in early mindreading. *Psychological Review*, *123*(2), 219–227.

Michael, J., Felber, A., & McEllin, L. (2020). Prosocial effects of coordination: What, why and how? *Acta Psychologica*, *207*, 103083.

Michael, J., Green, A., Siposova, B., Jensen, K., & Kita, S. (under review). Finish what you started: Instrumental helping in two-year-olds motivated by a preference to finish uncompleted actions. Pre-registration available at: http://aspredicted.org/blind.php?x=qz8dy6.

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

Michael, J., & Salice, A. (2017). The sense of commitment in human–robot interaction. *International Journal of Social Robotics*, *9*(5), 755–763.

Michael, J., Sebanz, N., & Knoblich, G. (2016a). The sense of commitment: A minimal approach. *Frontiers in Psychology*, *6*, 1968. 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.

Michael, J., & Székely, M. (2019). Goal slippage: A mechanism for spontaneous instrumental helping in infancy? *Topoi*, *38*(1), 173–183.

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

Miles, L. K., Nind, C., & Macrae, N. (2009). The rhythm of rapport: Interpersonal synchrony and social perception. *Journal of Experimental Social Psychology*, *45*(3), 585–589.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Nesse, R. (Ed.). (2001). *Evolution and the capacity for commitment*. Russell Sage Foundation.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291.

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*(6685), 573–577.

Ooi, J., Francóva, A., Székely, M., & Michael, J. (2019). The sense of commitment in individuals with borderline personality disorder traits in a non-clinical population. *Frontiers in Psychiatry*, *9*, 519. https://doi.org/10.3389/fpsyt.2018.00519

Paulus, M. (2014). The early origins of human charity: Developmental changes in preschoolers' sharing with poor and wealthy individuals. *Frontiers in Psychology*, *5*, 344.

Paulus, M., & Moore, C. (2012). Producing and understanding prosocial actions in early childhood. In *Advances in child development and behavior* (Vol. 42, pp. 271–305). JAI.

Pezzulo, G., Donnarumma, F., & Dindo, H. (2013). Human sensorimotor communication: A theory of signaling in online social interactions. *PLoS One*, *8*(11), e79876.

Piaget, J. (1950). *The psychology of intelligence*. Harcourt, Brace.

Plotnik, J. M., Lair, R., Suphachoksahakun, W., & De Waal, F. B. (2011). Elephants know when they need a helping trunk in a cooperative task. *Proceedings of the National Academy of Sciences*, *108*(12), 5116–5121.

Rachlin, H. (2016). Self-control based on soft commitment. *The Behavior Analyst*, *39*(2), 259–268.

Rakoczy, H., & Schmidt, M. F. (2013). The early ontogeny of social norms. *Child Development Perspectives*, *7*(1), 17–21.

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

Read, D. (2004). Intertemporal choice. *Blackwell Handbook of Judgment and Decision Making*, 424–443.

Reddy, M. D., Kita, S., Michael, J., & Siposova, B. (in prep). Does increased coordination in joint action increase young children's commitment or decrease their need to social reference?

Reinach, A. (1913). Die apriorischen Grundlagen des bürgerlichen Rechtes. Now In: Schuhmann K, Smith, B. (1989) Adolf Reinach. Sämtliche Werke. Textkritische Ausgabe, 2 vols., Munich: Philosophia Verlag, 141–278. Engl.

trans. by J. Crosby (1983). The a priori Foundations of the Civil Law, Aletheia. *An International Journal of Philosophy III*: 2–142.

Rheingold, H. L. (1982). Little children's participation in the work of adults, a nascent prosocial behavior. *Child Development*, 114–125.

Roberts, G. (2005). Cooperation through interdependence. *Animal Behaviour*, *70*(4), 901–908.

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). Oops!': Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, *35*(6), 747–758.

Robinson, T. E., & Berridge, K. C. (2003). Addiction. *Annual Review Psychology*, *54*, 25–53.

Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research Review*, *18*, 247–291.

Roth, A. (2018). Interpersonal obligation and joint action. In M. Jankovic & K. Ludwig (Eds.), *Routledge Handbook of Collective Intentionality* (pp. 45–57). Routledge, Milton Park.

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

Salice, A., & Michael, J. (2017). Joint commitments and group identification in human-robot interaction. In *Sociality and Normativity for Robots* (pp. 179–199). Springer, Cham.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71.

Sanislow, C. A., Morey, L. C., Grilo, C. M., Gunderson, J. G., Shea, M. T., Skodol, A. E., et al. (2002). Confirmatory factor analysis of DSM-IV borderline, schizotypal, avoidant and obsessive-compulsive personality disorders: Findings from the collaborative longitudinal personality disorders study. *Acta Psychiatrica Scandinavica*, *105*, 28–36. https://doi.org/10.1034/j.1600-0447.2002. 0_479.x

Scanlon, T. M. (1998). *What we owe to each other*. Harvard University Press.

Schelling, T. C. (1980). *The strategy of conflict*. Harvard University Press.

Schino, G., & Aureli, F. (2017). Reciprocity in group-living animals: Partner control versus partner choice. *Biological Reviews*, *92*(2), 665–672.

Schmidt, M. F., Rakoczy, H., & Tomasello, M. (2012). Young children enforce social norms selectively depending on the violator's group affiliation. *Cognition*, *124*(3), 325–333.

Schrift, R. Y., & Parker, J. R. (2014). Staying the course: The option of doing nothing and its impact on postchoice persistence. *Psychological Science*, *25*(3), 772–780.

Searle, J. R. (1965). What is a speech act. *Perspectives in the Philosophy of Language: A Concise Anthology*, *2000*, 253–268.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76.

Sell, A., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., & Tooby, J. (2017). The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition*, *168*, 110–128.

Sen, A. (2005). Why exactly is commitment important for rationality? *Economics & Philosophy*, *21*(1), 5–14.

Sen, A. (2002). *Rationality and freedom*. Harvard University Press.

Sen, A. (1991). Opportunities and freedoms (from the Arrow Lectures) in Sen 2002 (583–622).

Sen, A. (1977). Rational fools: A critique of the behavioural foundations of economic theory. *Philosophy and Public Affairs*, *6*, 317–344.

Shah, J. Y., Friedman, R., & Kruglanski, A. W. (2002). Forgetting all else: On the antecedents and consequences of goal shielding. *Journal of Personality and Social Psychology*, *83*(6), 1261.

Shpall, S. (2014). Moral and rational commitment. *Philosophy and Phenomenological Research*, *88*(1), 146–172.

Siamwalla, A. (1978). Farmers and middlemen: Aspects of agricultural marketing in Thailand. *Economic Bulletin for Asia and the Pacific*, *39*(1), 38–50.

Sibicky, M. E., Schroeder, D. A., & Dovidio, J. F. (1995). Empathy and helping: Considering the consequences of intervention. *Basic and Applied Social Psychology*, *16*(4), 435–453.

Siegel, E., & Rachlin, H. (1995). Soft commitment: Self-control achieved by response persistence. *Journal of the Experimental Analysis of Behavior*, *64*(2), 117–128.

Siposova, B., Székely, M., & Michael, J. (in preparation). Children's sense of commitment to a partner who has invested in a joint action.

Smallwood, J., Davies, J. B., Heim, D., Finnigan, F., Sudberry, M., O'Connor, R., & Obonsawin, M. (2004). Subjective experience and the attentional lapse: Task engagement and disengagement during sustained attention. *Consciousness and Cognition*, *13*(4), 657–690.

Sperber, D., & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind & Language*, *27*(5), 495–518.

Sterelny, K. (2012). *The evolved apprentice*. MIT Press.

Svetlova, M., Nichols, S. R., & Brownell, C. A. (2010). Toddlers' prosocial behavior: From instrumental to empathic to altruistic helping. *Child Development*, *81*(6), 1814–1827.

Székely, M., McEllin, L., Butterfill, S., & Michael, J. (in preparation). The perception of a partner's effort boosts cognitive control to sustain commitment in joint action.

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. ISO 690.

Székely, M., Powell, H., Vannucci, F., Rea, F., Sciutti, A., & Michael, J. (2019). The perception of a robot partner's effort elicits a sense of commitment to human-robot interaction. *Interaction Studies*, *20*(2), 234–255.

Thaler, R. H., & Shefrin, H. M. (1981). An economic theory of self-control. *Journal of Political Economy*, *89*, 392–406.

Theriault, J. E., Young, L., & Barrett, L. F. (2020). The sense of should: A biologically-based framework for modeling social pressure. *Physics of Life Reviews, 36*, 100–136.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für tierpsychologie*, *20*(4), 410–433.

Tollefsen, D. (2005). Let's pretend: Children and joint action. *Philosophy of the Social Sciences*, *35*(75), 74–97.

Tomasello, M. (2009). *Why we cooperate*. MIT Press.

Török, G., Pomiechowska, B., Csibra, G., & Sebanz, N. (2019). Rationality in joint action: Maximizing coefficiency in coordination. *Psychological Science*, *30*(6), 930–941.

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

Tuomela, R. (2007). *The philosophy of sociality: The shared point of view*. Oxford University Press.

Vaish, A., Missana, M., & Tomasello, M. (2011). Three-year-old children intervene in third-party moral transgressions. *British Journal of Developmental Psychology*, *29*(1), 124–130.

Vignolo, A., Sciutti, A., Rea, F., & Michael, J. (2019). Spatiotemporal Coordination Supports a Sense of Commitment in Human-Robot Interaction. In: Salichs, M. et al. (eds) Social Robotics. ICSR 2019. Lecture Notes in Computer Science, vol 11876. Springer, Cham.

Walton, D. (2002). The sunk costs fallacy or argument from waste. *Argumentation*, *16*(4), 473–503.

Warneken, F., Hare, B., Melis, A. P., Hanus, D., & Tomasello, M. (2007). Spontaneous altruism by chimpanzees and young children. *PLoS Biology*, *5*(7), e184.

Warneken, F., & Tomasello, M. (2008). Extrinsic rewards undermine altruistic tendencies in 20-month-olds. *Developmental Psychology*, *44*(6), 1785.

Warneken, F., & Tomasello, M. (2007). Helping and cooperation at 14 months of age. *Infancy*, *11*(3), 271–294.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301–1303.

Wilson, T. D., & Gilbert, D. T. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science*, *14*(3), 131–134.

Index

A Note: **Bold** page numbers refer to tables, *Italic* page numbers refer to figures and page number followed by "n" refer to end notes.