COLLECTIVE GOALS, COLLECTIVE REASONING, COLLECTIVE ACTION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF PHILOSOPHY,
THE GRADUATE PROGRAM IN HUMANITIES,
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
PHILOSOPHY AND HUMANITIES

Randall Edward Harp
September 2009
I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Michael Bratman) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Debra Satz)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Joshua Landy
Department of French & Italian)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(John Ferejohn
Department of Political Science)

Approved for the University Committee on Graduate Studies.
Abstract

In this dissertation I provide a philosophical account of deliberative collective action. Deliberative collective action is characterized by the fact that each of the agents in the collective action must judge her or his own contribution to the collective action to be rationalizable by the collective action—that is, each agent’s action must be judged to be best supported by the available reasons.

I argue that the best way to account for deliberative collective action is through collective goals. Agents who participate in a collective action can deliberate about what to do, or justify their own contribution to the collective action, through appeal to the collective goal. Existing theories of collective action are unsatisfactory in accounting for deliberative collective action, I argue, because they cannot, for various reasons, adequately account for collective goals.

My own account of collective goals holds that collective goals are functionally characterized by their imposing rational constraints upon the agent’s deliberation. Collective goals are distinct from individual goals, I argue, in that collective goals are understood to impose rational deliberative constraints on every member of the collective.

I conclude that deliberative collective action is best analyzed as being the successful production of an outcome by a set of agents who each share a collective goal, who each accept that collective goal as imposing rational constraints upon their deliberation and decision making, and who each hold beliefs to the effect that every member of the collective also holds the goal and constrains her or his deliberation accordingly.
Preface

The philosophical argument that I present in this dissertation makes every effort to minimize the use of formal language and models. Nevertheless, the arguments I present are strongly influenced by my thinking through these issues in a formal way. Some of the details of the formal model that I adopt are presented in the appendix; my hope, however, is that none of the specifics of the formal model are necessary for the philosophical arguments I present in the body of the dissertation.

Many of the examples use two agents for simplicity, although they can be modified to include more than two without loss of generality.

A note about my terminology: specific agents are referred to throughout as X or Y. I arbitrarily refer to X consistently as ‘she’ and to Y consistently as ‘he’. X’s action is generally referred to as x, and Y’s action is generally referred to as y. Intentions (of either X or Y, depending on the example) are referred to with the greek letters φ and ψ. Outcomes are generally referred to as o (or o₁, o₂, et cetera). When exceptions are made to these terminological rules, I try to be as clear as possible.
Acknowledgements

My development as a scholar reflects the generosity and patience and assistance and wisdom of very many people, and I regret that I cannot mention them all here. These acknowledgements are poor thanks for the help that others have given me. I apologize for any unintentional omissions.

I am grateful to my principal advisor, Michael Bratman, for his patient shepherding of this project, and for his willingness to humor my well-meaning attempts to find something valuable in a field that he has already cultivated so well. He has been extremely generous in his time and in his insightful comments, and I thank him for it.

I am equally grateful to Debra Satz, one of my readers, for her unflagging encouragement and patience; that I was able to finish is testament in no small part to her supportive words. My arguments are greatly improved because of her comments and suggestions.

I had many very interesting conversations with Joshua Landy, the second reader on my committee; I regret that the most interesting and fruitful of our discussions did not find place in the final project here. In an academy that is increasingly insular, I am always grateful for discussions that breach disciplinary walls.

John Ferejohn was gracious enough to become a reader of my dissertation at a very late stage, and I am extremely thankful for the time and energy that he put into doing so. That he would lend me his considerable expertise, on such short notice, is humbling.

Work that developed into parts of this dissertation has been presented at Stanford University, at the University of California-Santa Cruz, at the Philosophy of Social Science Roundtable at the University of Washington, and at the University of Vermont; I am grateful for the many helpful comments I received from the attentive and astute audiences at each location.

Stanford was a wonderful place to be a graduate student, and I am grateful for all who contributed to the outstanding intellectual atmosphere. I would like to thank my one-time
office mates, Facundo Alonso and Laurel Scotland-Stewart, for each knowing more about my project than I do. Audrey Yap was a great and wise friend. I would like to thank Simon May, and Laura Maguire, and Manuel Vargas. I am grateful to my cohort: Dan Corbett, Robert Jones, Daniel Lafave, Ken Reisman, Ken Stalzer, and Michael Weisberg, for creating an environment in which we could all grow as scholars. I am also thankful for valuable discussions with Lael Weis, Tobey Scharding, Darko Sarenac, Damon Horowitz, and Jennifer Morton. I found my discussions with Olivier Roy to be extremely helpful, and I thank him sincerely.

The Graduate Program in Humanities gave me an intellectual grounding that I have found invaluable; I would like to thank the directors of the program, Paul Robinson and Helen Brooks. I would also like in particular to thank one of my colleagues in the program, Holly Case, for many valuable discussions and advice.

Michael Eichmair helped me immensely both academically and personally, and he has my thanks.

I must also thank my friends who kept me sane and kept me on track: Mathias Wintzer, Richard O’Grady, and Tricia Lee.

The Introduction to Humanities program at Stanford was my academic home for two years during the writing of this project, and I would like to thank my colleagues there: Jennifer Barker, Magdalena Barrera, Phaedra Bell, Rashida Braggs, Sarah Jane Cervenak, Galen Davis, Jehangir Malegam, Kathryn Mathers, Carol Pal, Carrie Schroeder, and Gabriel K. Wolfenstein.

During my last two years in Burlington I have been a George Washington Henderson scholar at the University of Vermont; I thank the University and Wanda Heading-Grant for the opportunity.

The community of scholars at the University of Vermont has been as welcoming and supportive as I could ever ask. The members of the department of philosophy have given freely of their time and their wisdom, and I would like to thank them: Bill Mann, Sin Yee Chan, Terrence Cuneo, Louis DeRosset, Tyler Doggett, Mark Moyer, Adam Wager, and Matthew Weiner. I would like to thank Don Loeb and Arthur Kuflik in particular for being incredibly generous of their time and advice; they each read substantial parts of the dissertation and offered suggestions and advice throughout. This project would be substantially worse without their help.

I would also like to thank other scholars in Vermont for their being so welcoming and
helpful: Pablo Bose, Kareem Khalifa, Dorian McCoy, Travis Nelson, Rashad Shabazz, Harvey Amani Whitfield, and Alex Zakaras. Folashade Alao and Rachael Neal helped keep me on track with my writing and offered excellent advice.

I must also thank the people who have administrated the various programs I have been fortunate to have been a part of; theirs is often a thankless task, and I cannot express how much I value their friendship, their advice, and above all their incredible abilities to get things done: Joanie Berry, Jill Johnston Covington, Lori Jean, Monica Moore, and Kiki Pritchard enabled this project to be done and provided me with incomparable support along the way. I am also grateful to Rosemary Leute, in memoriam.

I appreciate as well the numerous stimulating discussions that I have had with my students, both at Stanford and at the University of Vermont.

Burlington has been a wonderful place to finish my writing, and I am thankful to all who have made it so. I also thank the ladies of the Sapa for creating a wonderful tea shop in which I could write.

And Kae.

Finally, I must thank my family. All of my cousins have been supportive; I thank Ashley and Mercedes for being especially so. My brothers have been the guiding forces in my life, and I cannot thank them enough. Phillip has been the perfect older brother: he has always taught me the right ways to do things, and the best ways to get things done. Without his example, this project would never have been completed. Reynaud is my idol; he is better than me in all the things I think worth doing, and I have always striven to live up to him. And Eric, who has always helped me keep my perspective, and has been the best friend I could ask for.

This dissertation is dedicated to my father, Reynaud Lemoine Harp Sr., and to my mother, Sandra Lorraine Hedrick Harp. Without which not.
## Contents

Abstract \hspace{2cm} v

Preface \hspace{2cm} vi

Acknowledgements \hspace{2cm} vii

1 Introduction

1.1 The problem of collective action \hspace{2cm} 1

1.2 A catalog of possible theories of collective action \hspace{2cm} 5

1.2.1 Reductive theories: fully reductive theories \hspace{2cm} 5

1.2.2 Reductive theories: team preference theories \hspace{2cm} 6

1.2.3 Reductive theories: we-action theories \hspace{2cm} 6

1.2.4 Reductive theories: we-intention theories \hspace{2cm} 7

1.2.5 Other possible reductive theories \hspace{2cm} 8

1.2.6 Plural subject theories \hspace{2cm} 8

1.3 Desiderata for an adequate theory \hspace{2cm} 10

1.4 My proposed solution \hspace{2cm} 11

1.5 The plan of this work \hspace{2cm} 13

2 Fully reductive theories and Searle’s challenge

2.1 Introduction \hspace{2cm} 16

2.2 Individual action \hspace{2cm} 18

2.2.1 Rationalizations \hspace{2cm} 22

2.2.2 Constraints on value functions \hspace{2cm} 25

2.3 Is collective action reducible to individual action? \hspace{2cm} 27

2.4 Searle’s challenge \hspace{2cm} 29
2.5 Answering Searle’s challenge ........................................... 34

3 Theories of collective action .............................................. 36
  3.1 Introduction ................................................................... 36
  3.2 The team preference approach ........................................ 38
    3.2.1 Team preferences .................................................... 39
    3.2.2 Team reasoning ...................................................... 42
    3.2.3 Problems .............................................................. 45
  3.3 We-action theories ......................................................... 50
    3.3.1 Shared intentions and team goals ............................... 50
    3.3.2 Bratman’s intention-based theory of collective action .... 53
    3.3.3 Shared intentions and circularity ............................... 54
  3.4 We-intention theories ..................................................... 60
    3.4.1 Searle ................................................................. 60
    3.4.2 Tuomela ............................................................... 63
    3.4.3 On the conceptual links between we-intentions and I-intentions . 69
    3.4.4 Phenomenological properties of collective and individual intentions . 75
  3.5 Plural subject theories .................................................... 79
    3.5.1 Gilbert and Roth against reductive theories ............... 81
    3.5.2 Positive plural subject theories of collective action ....... 88
    3.5.3 Against Gilbert and Roth ......................................... 91

4 Deliberation under constraints ............................................ 102
  4.1 Introduction ................................................................... 102
  4.2 Minimal individualism .................................................... 103
  4.3 A proposed model of deliberation ..................................... 106
    4.3.1 A normative model of decision theory ......................... 107
    4.3.2 Subjective normativity versus objective normativity ..... 109
    4.3.3 The rational resolution of subjective decision problems . 112
    4.3.4 Why not objective decision problems? ....................... 115
  4.4 The possibility of constrained rational deliberation ........... 117
    4.4.1 Preferences .......................................................... 118
    4.4.2 Intentions and values .............................................. 121
    4.4.3 Intentions and rational constraints ........................... 132
4.4.4 Objections to intention rationality considered and refuted . . . . . . . 137
4.5 Rational deliberation over intentions .................................. 141

5 Collective goals and deliberation ......................................... 148
  5.1 Introduction ........................................................................ 148
  5.2 Intentions and goals .......................................................... 151
  5.3 Goals and rational decisions ............................................. 158
    5.3.1 Rational resolution of SADPs .................................... 159
    5.3.2 Rational resolution of MADPs .................................... 160
  5.4 Action intentions and outcome intentions ......................... 162
  5.5 Action intentions fail to explain collective goals ................. 167
    5.5.1 Conditional intentions ............................................ 171
  5.6 Outcome intentions and collective goals .......................... 174
  5.7 Outcome intentions and the own action condition ............... 176
    5.7.1 Negative and positive constraints ........................... 179
  5.8 Rationally holding and acting upon collective goals ............ 180
    5.8.1 Individual rationality and collective goals ................ 181
    5.8.2 Collective rationality and collective goals ................ 184
  5.9 Collective goals and coordination .................................. 185
  5.10 Collective goals versus common goals .......................... 188

6 Searle’s challenge answered ............................................. 190
  6.1 Introduction ........................................................................ 190
  6.2 A theory of collective action summarized ......................... 191
  6.3 Collective action applied .................................................. 192
    6.3.1 The footballers’ problem and Hi-Lo games ................ 192
    6.3.2 Going to New York together .................................. 197
    6.3.3 Walking and driving together ................................. 201
    6.3.4 Searle’s cooperative businessmen .......................... 205
  6.4 Contrastive examples ..................................................... 207

A A formal model of intentions in rational choice .................. 213
  A.1 Games with and without intentions ................................ 213
  A.2 Outcome intentions and action intentions ...................... 215
Chapter 1

Introduction

1.1 The problem of collective action

When we speak of the ‘problem of collective action’, what is it that we are speaking of? We can start by stating the things that are not the problem of collective action—or, at least, are not the problem of collective action that is our focus here. We are not interested in determining ways to make people more cooperative than they already are. This might be a laudable goal in general, although we would be remiss not to point out that many instances of cooperation are towards irrational or otherwise undesirable goals. Groups can collectively engage in irrational actions: collective action can be directed towards promoting actions or traits which are destructive for the group in question, just as collective action can be directed towards outcomes which are unjustified by the group’s reasons for acting (as when a group collectively persecutes an individual on the basis of false beliefs). Likewise, groups can collectively engage in undesirable or immoral actions: banks can be robbed through collective actions, and genocides can be enacted. We cannot assume that collective projects are always beneficial. Whether collective actions tend to be beneficial or not is irrelevant for our current project, however. Our project here is not to make people more cooperative.

Nor is our project to explain what would count as a solution to the narrow class of collective action puzzles which have an inordinately privileged place in decision theory and literature on rational and social choice—namely, prisoner’s dilemma (PD) games. The prisoner’s dilemma, a general class of games first theorized about by Merrill Flood and Melvin Drescher and popularized by Albert Tucker, is not without interest. But while some

\footnote{See Kuhn 2009 Spring.}
CHAPTER 1. INTRODUCTION

Theorists see PD games as synonymous with games of collective action, this interpretation is plainly too strong. Even within the game theoretic tradition, there are many other notable cooperation games than just PD games. As just one example, Skyrms has pointed out the importance of stag hunt games for understanding the rise of cooperation in evolutionary game theoretic contexts. But, more importantly, once we leave the confines of game theory it is harder to see why we should equate cooperation with behavior in a prisoner’s dilemma. Prisoner’s dilemma games all share a specific form, in which defecting from the cooperative outcome is a dominant strategy because of the costs associated with cooperation. There are many instances of cooperation, however, which do not require that cooperation be costly relative to defection, or that defecting be a dominant strategy. It is sometimes in people’s best interest to cooperate. Indeed, it can at times be the clearly rational thing to do, even according to the standards of rationality presupposed by rational choice theories. We limit our attention to the prisoner’s dilemma at the cost of having nothing to say at all about other types of cooperation. And when we leave the context of game theory, other types of cooperation are abundant.

This is not to say that we should ignore the prisoner’s dilemma completely. On the contrary, in the discussion to come we will frequently have occasion to analyze PD games. We should, however, put the prisoner’s dilemma in its proper context. A consideration of the prisoner’s dilemma reminds us of two things: it reminds us that there are genuine instance of collective action—that collective action does happen, even if it does not seem to be in the best interest of the individual participants—and it shows us that agents can justify contributing to collective actions—that agents view their actions in pursuit of collective goals as rational. One outcome in PD games has traditionally been labeled the ‘cooperative’ or ‘collective’ outcome (namely, the outcome in which both agents play the action typically labeled ‘cooperate’). The PD game is supposed to model real-world interactions that agents have; in particular, it is supposed to model instances of real-world collective action that occur even though agents have individual reason to not engage in collective action at all.

---

2Examples include Gauthier 1986 and Medina 2007.
4Some, of course, would object to the claim that cooperation in prisoner’s dilemmas runs counter to the best interest of the individual participants. One way that theorists attempt to ‘solve’ the prisoner’s dilemma is by redefining prisoner’s dilemma games so as to eliminate the seeming irrationality of cooperating, either by redefining the preferences involved or by taking into consideration reputation effects in future games; see e.g. McMahon 2001 or Hollis 1998. We will discuss these proposals in greater detail below, but it suffices here to say that we should be skeptical about grounding a satisfactory general theory of cooperation in such approaches.
It is puzzling that such kinds of collective action occur, but they are no less instances of collective action for that reason. We can thus conclude that collective action does, in fact, occur. As for the second thing, collective action (such as that which arises out of some PD games) is justifiable by agents in two ways: agents can justify the overall collective action, and they can also justify performing some specific action as their contribution to the collective action. Focusing on PD games is thus too narrow, but there are important lessons to draw from PD games that we will continue to focus on later.

Let us leave the prisoner’s dilemma for now, then, and concentrate on the claims that collective action sometimes exists, and that there is something rational in agents’ participation in collective action. Although some collective action is not deliberative—the skilled tango dancers do not deliberate about how to move, they simply do—some collective action is deliberative. Each of the participants in the collective action must deliberate, either individually or collectively, about what to do so as to satisfy the group goal. Two people might intend to take a trip together, for instance, without yet knowing where they will go, or how they will get there, or what they will do.

It is this phenomenon that we are interested in—namely, deliberative collective action, in which a group of agents face a decision problem about what to do and must resolve the decision problem one way or another. Now, why should we restrict our attention just to deliberative collective actions? Our doing so should not be taken to imply that non-deliberative collective behaviors cannot be perfectly respectable instances of collective action. Jazz musicians who collectively improvise over a song are engaged in a collective action; nevertheless, it is a stretch to suggest that their choices of which notes to play are deliberative choices. Rather, the jazz musicians possess non-deliberative ways of selecting notes to play—ways which are admittedly honed through practice and careful study and deliberate reflection, but which are non-deliberative nonetheless. The same is true of skilled athletes who are playing together against a common opponent; they are engaged in a collective action, though their choices are not deliberative at all.\[^5\]

The answer is that we focus on deliberative collective actions here because focusing on deliberation is the easiest way to ensure that we are looking at collective actions and not merely collective behaviors. Actions, as we will see in greater detail below, are behaviors performed for reasons—behaviors which are rationalizable given the reasons that the agent possesses at the time of action. Deliberation is an active consideration of the reasons that

\[^5\text{But cf. Sugden’s presentation of his footballers’ problem below.}\]
are relevant for any given decision. Consequently, when through a process of deliberation an agent \( X \) judges the performance of action \( x \) to be most in accordance with her reasons at the time, then \( X \)’s doing \( x \) is likely to be her action.\(^6\) That \( X \) decides to do \( x \) through a process of deliberation, in other words, is a good indicator that \( X \)’s doing \( x \) is an action and not mere behavior.

Moreover, though we do not develop it here, the account of collective action that we give will be adaptable to circumstances which do not feature deliberative behaviors. The core feature of my theory is that agents in collective actions possess collective goals. Collective goals, like individual goals, can be pursued deliberatively or non-deliberatively. (In the collective case, it is also possible for there to be mixed cases, in which some agents deliberate towards a goal while others do not.) A more complete account of collective action, then, will provide an account of both deliberative cases of collective action and of non-deliberative cases of collective action. Giving an account of non-deliberative cases of collective action, however, requires that we have a theory of non-deliberative action for reasons. While I do not suggest that such an account cannot be given, I merely state here that such an account will take us beyond the scope of the current work.

Accordingly, let us return to deliberative resolutions of decision problems. Some resolutions of a decision problem are reflective of collective actions, as when agents in a prisoner’s dilemma act cooperatively, or when agents facing a decision about how to travel together decide to take the train together to New York. Other resolutions of the decision problem are not reflective of collective action, as when agents in a prisoner’s dilemma selfishly defect or when agents facing a decision about traveling choose to travel to distinct places.

We face, then, the puzzle of how to account for deliberative collective action. This puzzle has two facets. The first facet is the \textit{metaphysical question}: what is the difference between a collective action and an individual action? What does it mean for agents to resolve a decision problem in a way consistent with their being engaged in a collective action? What is the difference, in other words, between resolving a decision problem \textit{collectively} and resolving the decision problem \textit{individually}? Given some set of actions, how do we demarcate it as either a set of collective actions or a set of individual actions?

\(^6\)A common challenge to this claim is that deviant causal chains might exist: an agent might recognize that she has most reason to do \( x \), and then proceed to do \( x \) through a process which is causally unrelated to her decision to do \( x \), or which is related in inappropriate ways to her decision. We can bracket for the moment the issue of deviant causal chains.
The metaphysical question is closely connected with the descriptive properties of collective action. The metaphysical question speaks to what collective action is, as opposed to what ought to be given the existence of a collective goal.

To address the normative side of collective action, we should turn to the second facet, namely the deliberative question. Suppose we have a rational agent, facing a decision problem, who intends to engage in a collective action with other agents. How should this agent’s deliberation be affected by the agent’s intention to engage in a collective action? What, in other words, ought the agent to do, assuming that the agent intends to act collectively?

1.2 A catalog of possible theories of collective action

Any successful theory of collective action must provide an answer for the questions posed by these two facets of the collective action problem. At the most general level, theories of collective action fall into one of two categories: they either appeal to a set of individual agents, each of whom deliberates and acts on the basis of individually-held mental states, or they do not. We can call the first category of theories reductive theories of collective action, and we can call the second category of theories non-reductive, where both reductive theories and non-reductive theories admit of many variations.

1.2.1 Reductive theories: fully reductive theories

All reductive theories, at bottom, appeal to individual agents each acting on the basis of individually-held mental states. Therein lies the reduction: collective actions can be reduced to a set of individual actions performed by individual agents acting upon individually-held mental states. We face the further question of what kind of mental states the agents possess, and what the content of those mental states are. Some theories of collective action, which we will call fully reductive theories of collective action, hold that agents engaged in a collective action differ from agents engaged in a set of individual actions neither in the kind of their mental states, nor (in principle) in the content of their mental states. If our theory of individual action holds that individual actions can be accounted for by appeal to intentions, beliefs, and desires, for instance, then the fully reductive theorist argues that so too can collective actions. Likewise, whatever the content may be of those mental states which account for individual action, so too would be the content of the mental states which account for collective action. Some examples of a fully reductive theory of collective action
are those proposed by traditional rational choice theories of action, including most accounts of collective action that are grounded in game-theoretic analyses of rational action.

1.2.2 Reductive theories: team preference theories

Reductive theories of collective action need not be fully reductive. When a reductive theory is not fully reductive, then the theory will have some element which makes a non-reductive appeal to the collective or to the collective action. As one example, a theory of collective action might appeal to mental states which are the same kind as those mental states appealed to in accounts of collective action, but where the contents of the mental states are different—in particular, the contents of the mental states might make non-reductive reference to the collective action itself. Along these lines, one way of accounting for the collectivity of collective action while still appealing only to individual agents and individually-held mental states is by allowing for a unique and non-reductive kind of preference. Preferences have a privileged position in traditional rational choice theories: preferences rationalize actions and goals, in the sense that they provide rational justification for the performance of some action or the adoption of some goal. If, then, we want a theory of collective action in which actions taken in support of collective goals are rationalized, it makes sense to try to account for such rationalization through collective preferences. This is the approach taken by Michael Bacharach, by Robert Sugden, and by Natalie Gold. On this theory of collective action, which we will call a team preference theory of collective action, collective actions occur when each of a set of agents rationalizes her or his action according to a common collective (or team) preference rather than according to her or his individual preferences.

1.2.3 Reductive theories: we-action theories

The team preference approach is somewhat unique, in that it attempts to explain collective actions by appealing to distinctive preferences, and to a distinctive process of reasoning that agents in a collective action engage in. Most accounts of individual action are based upon identifying the intentions of the agent (rather than the preferences of the agent) and whether those intentions are properly responsive to the reasons that the agent possesses. Accordingly, if we want a theory of collective action in which the agents have ‘ordinary’ mental states but with ‘special’ content, then the mental states with special content might plausibly be intentions, rather than preferences. Our account of collective action, then, would consist of a set of agents who each intend a collective action, rather than an individual
action. The intentions themselves are on par with intentions that would be held by agents in individual actions—there is no difference in the kind of intention employed—but their content is distinct, insofar as it is the collective action and not an individual action. We will call all such theories *we-action theories*. We-action theories of collective action differ from fully reductive theories primarily in the contents of the intentions. Whereas fully reductive theories restrict themselves to the same type of content as that used to explain individual action, we-action theories differentiate collective actions from individual actions according to the content of the intentions: collective actions feature intentions which have as their content the collective action, while in individual actions agents’ intentions have as their content the individual action. We-action theories are still reductive theories, however, because collective action still reduces to a set of individual actions appropriately interrelated. Michael Bratman’s theory of collective action is an instance of a we-action theory.

### 1.2.4 Reductive theories: we-intention theories

There is a fourth reductive option for accounting for collective action; we will call this class of theories *we-intention theories*. For we-intention theories, as with all reductive theories, collective actions are still performed by individual agents acting on the basis of individually held mental states. The difference here is that the *kind* of mental state that the agents are acting upon is different than it is in the case of individual action. We still focus on intentions as the action-relevant mental state, just as we do with we-action theories; whether agents are engaged in a collective action depends crucially on what kinds of intentions they possess. With we-intention theories, however, the kind of intention that agents hold and act upon in collective actions is different from the kind of intention that agents hold and act upon in individual actions. We can call this distinctive kind of intention a *we-intention*, which contrasts with the usual *I-intentions*. When agents are engaged in individual actions, then, they act on the basis of I-intentions; when agents are engaged in collective actions, they act on the basis of we-intentions.

We-intention theories of collective action are, as mentioned above, reductive. They are not fully reductive, however, which means that they make some non-reductive appeal to collectivity at some place in the theory. For we-intention theories, the non-reducible collectivity is introduced at the level of the intention: collective actions are performed by agents who hold we-intentions, where we-intentions are not reducible at all to I-intentions. Thus, while we-intention theories are reductive theories, they are not fully reductive theories.
We will examine two non-reductive collective intention theories: one by John Searle, and one by Raimo Tuomela.

1.2.5 Other possible reductive theories

It is theoretically possible for there to be a fifth kind of reductive theory: namely, a reductive theory which makes use of both irreducibly collective kinds of intentions and also irreducibly collective content to the intentions. This would seem to be overkill, however; it is not clear why we would need a theory with both collective intentions and also collective content to the intentions. Presumably, either one is sufficient to account for the collectivity of the resulting action, and so we can disregard reductive theories that require both distinct types of intentions and also distinct contents of the intentions. And, of course, there might be theories of collective action which involve some combination of intentions and preferences, with varying degrees of irreducible collectivity contained in these mental states. These theories do not have any champions in the existing literature on collective action, however, and so we will disregard these theories as well.

1.2.6 Plural subject theories

So much for the reductive theories. We can take it to be uncontroversial that, when it comes to explanatory theories, theories which are more reductive are to be preferred over equally explanatory theories which are less reductive. Non-reductive theories of collective action, then, face the pressure of providing some reason as to why reductive theories are not likely to be successful. The argument might be contingent: we might argue that no reductive theory to this point has been successful, and so in the absence of any plausible reductive theory we will favor a non-reductive theory. Likewise, the argument might be more principled; we might argue that there is some feature of collective actions that reductive theories of whatever stripe cannot provide any satisfactory account of. Either way, we might despair of the possibility of a reductive theory successfully accounting for collective action, and so favor a non-reductive theory.

To my knowledge, no one who argues for a non-reductive theory of collective action gives the contingency argument. Rather, proponents of non-reductive theories argue that collective actions necessarily entail directed normative commitments that hold between the participants, and that reductive theories cannot account for these directed normative
commitments. This being so, non-reductive theorists who take principled stands against reductive theories have to do two things: they have to first explain why reductive theories cannot account for these directed normative commitments, and they have to second explain what a plausible non-reductive theory of collective action looks like.

We will consider two ways in which a non-reductive theory might be constructed. The first way of building a non-reductive theory accepts that collective actions are performed by individual agents with individual mental states (in particular, intentions). What is denied here is that the action-relevant mental states must be held by the agent. Although it is difficult to see what the other option is—what sorts of mental states could move an agent to act, if not the agent’s own?—there is, in fact, another option. Collective action might occur, in other words, when one agent acts directly upon the mental states held by another. This is the approach taken by Abraham Sesshu Roth.

A second approach is just to deny that collective actions are performed by individual agents. Instead, collective actions might be performed by a special kind of agent that Gilbert terms a plural subject. A plural subject is an agent, and so can engage in rational action, but plural subjects are agents that are not reducible to a set of individual agents with individual mental states. Plural subjects are comprised of individual agents, of course, but plural subjects are not reducible to individual agents; plural subjects are independent, autonomous agents who act just as individual agents do. Accordingly, collective actions are performed by plural subjects and not, strictly speaking, by individual agents. Margaret Gilbert is the main proponent of this kind of theory of collective action.

Though there are important differences between Roth’s view (which holds that collective actions are performed by individual agents) and Gilbert’s view (which does not), they are both non-reductive theories of collective action, because they both deny that collective action can be explained by individual agents acting on the basis of individually held mental states. By allowing agents to act directly upon mental states held by other agents, Roth blurs the boundaries of individual agency enough for us to consider his view a plural subject view. Thus, we will subsume both Roth’s view and Gilbert’s view under the moniker plural subject theories, although we should acknowledge that it is not the best fit for Roth, since he does not speak directly of plural subjects at all. More important than the name, however, is the claim that the two make—namely, that we cannot account for collective action with a reductive theory and that we need something else. It is their similarity in this respect which warrants their being grouped into the same general category.
1.3 Desiderata for an adequate theory

Any successful theory of collective action, regardless of how reductive, must be compatible with a set of plausible assumptions. The first is an assumption about the composition of collective behaviors: we assume that there is no remainder left over once we subtract the behavior of the agents from the behavior of the group. We can call this the behavior compositionality condition. Note that we are talking about behaviors and not actions so as to not beg any questions about collective action. Actions are necessarily connected with agents and with reasons, and so we might seem to beg the question against non-reductivists were we to claim that collective actions are reducible to individual actions. If collective actions can only be performed by plural subjects, then the action of the group is not reducible to the actions of the individuals. Talk about behavior is less controversial, however. The behavior compositionality condition only claims that the behavior of the group is reducible to the behavior of the constituent individuals, which is akin to saying that groups act through the behaviors of their members.

We also make an assumption regarding the deliberation of the individuals who constitute a group: we assume that collective goals constrain the deliberation of the individuals who constitute the collective. Individual agents exhibit behavior in pursuit of a collective goal, and individual agents sometimes need to deliberate about what to do in order to further the collective goal. Call this the distributive deliberation condition: collective goals must be able to be distributed over the deliberation of the individuals who constitute the collective. (It is possible that deliberation upon collective goals can only occur in a collective—that is, that deliberation upon a collective goal is always a collective action; I do not believe this plausible, but even in such cases, the deliberation must manifest itself in the psychology of each individual agent.)

A third assumption speaks to the individual attitudes that are held by the agents in a collective action. The assumption is that collective goals must be instantiated in individual agents’ mental states. Whatever a collective goal is, that collective goal must be reflected in the mental states that each agent in the collective holds. This third assumption we will call the individual possession condition.

We can call these three assumptions, taken together, minimal individualism. Minimal individualism is prima facie compatible with both reductive theories of collective action and with non-reductive theories of collective action. Minimal individualism is a plausible
position to hold because, at bottom, collectives are just collectives of individuals. Even if collective action is not fully reducible to individual action, it is still plausible to assume that collectives are made up of individuals, and that collective goals can be represented in some way in the psychologies of the individual agents. This is not full-blown methodological individualism; methodological individualism would require that we make no appeal to irreducibly collective actions. Gilbert’s account of collective action, for instance, violates methodological individualism. Gilbert’s account does not necessarily violate minimal individualism, however; her claims might be compatible with the assumptions of minimal individualism, so long as she can provide an account of plural subjects which explains how belonging to a plural subject affects an individual agent’s mental states and deliberation. I argue in Chapter 4 that our assumptions of minimal individualism are reasonable assumptions to make. I believe that all of the theories of collective action that we consider here—including Gilbert’s—are compatible with minimal individualism. If not, however, then so much the worse for those theories that cannot account for the assumptions.

1.4 My proposed solution

The theory of collective action that I set out to defend in this work is a reductive theory of collective action; it is a version of a we-action theory. The account of collective action that I defend has the following features. First, agents in a collective action each deliberate according to a collective goal. This collective goal is held by individual agents, not by plural subjects, and so is represented in each agent’s individual psychology. We explain what a collective goal is through appeal to the mental states that each agent in a collective action individually possesses.

Collective goals serve to justify the behavior of deliberating agents. If an action is believed to promote the production of the collective goal, then the agent is justified in performing that action. Further, if a set of agents each acts in justified fashion towards the pursuit of a common goal, then they are engaged in a collective action.

The account presented here, then, depends crucially on an analysis of collective goals. Collective goals, I argue, are analogous to individual goals. Individual goals have as their

---

It is less clear whether methodological individualism would rule out a theory like Roth’s, where the unit of agency is still the individual, yet where agents act upon mental attitudes that they do not hold. The point is ultimately a moot one, as we do not need a precise characterization of the boundaries of methodological individualism.
content an outcome which can be (or is believed to be) produced by an individual, through some action or set of actions performable by the agent. Collective goals are similar, but they have as their content outcomes which can be produced by the collective, through some set of behaviors performable by the collective.

A worry becomes salient here: namely, a worry about the own action condition. The own action condition is a plausible necessary condition on theories of action; it holds that agents can only intend those actions which are theirs, or those outcomes which they can produce through actions of their own. If a collective goal has as its content outcomes which are produced through a set of actions by a collective, then how can an agent deliberate according to a collective goal without thereby violating the own action condition? The answer to this depends on the account we give of goals (both individual and collective).

Individual goals, I argue, are intentions. Intentions, functionally, are rational constraints upon an agent’s deliberation. They are, in other words, filters of admissibility: holding an intention makes it irrational to choose any action which produces an outcome incompatible with the intention. Individual goals can either be represented as intentions to produce an outcome, or intentions to produce an action. When individual goals are represented as outcome intentions, the outcome must be of a specified form; the intended outcomes must be such that the production of the outcome is a function of the individual agent’s agency, and not of the agency of other agents in the decision problem.

Collective goals, I argue, are also intentions: collective goals are a set of permissible outcomes for the collective to bring about. However, whereas individual goals can be either action intentions or outcome intentions, collective goals must be outcome intentions. Collective goals are distinct from individual goals qua outcome intentions, because whereas individual goals must be translatable into action intentions, collective goals cannot be.

Just like individual intentions, collective goals constrain deliberation. But where individual intentions constrain the deliberation of only the individual deliberating agent, collective goals constrain the deliberation of every agent in the collective. For an agent to rationally hold and deliberate according to a collective goal, then, the agent must allow the collective goal to constrain her or his deliberation. Further, the agent must also believe the other agents in the collective to hold the same collective goal, and thus to constrain their deliberation according to it.

It is necessary for collective action, then, that each of a set of agents rationally holds and deliberates according to a common collective goal. It is further necessary that agents
perform those actions which are justified by their deliberation, and that their actions jointly produce an intended outcome. When those agents are successful in producing one of the outcomes which is compatible with their collective goal, then the agents’ collective action is, in that respect, a success. If, however, the relevant mental states are not shared—if the agents do not hold a collective goal in common—or if the agents are not successful in coordinating on one of the collective goal-compatible outcomes, then the agents are not engaged in successful collective action, although they still might be rationally justified in performing their actions in virtue of their being directed towards a (perceived) collective goal.

1.5 The plan of this work

My argument in favor of this theory proceeds as follows. In Chapter 2, I consider the difference between individual and collective action. One intuitive approach to analyzing collective action, I argue, is the fully reductive approach. This approach faces a serious worry, which was emphasized by, among others, Searle, and which I call Searle’s challenge. Searle’s challenge states that every collective action can be described as a set of actions by a set of agents, and that the challenge facing a theory of collective action is to explain the difference between the set of actions conceived as a set of individual actions, and the set of actions conceived as a collective action. Fully reductive theories of collective action are, I suggest, collective outcome theories of collective action: they identify an outcome as a collective outcome in virtue of some properties that it possesses, and a set of agents are cooperative so long as the agents succeed in bringing about a collective outcome. But such theories, I argue, cannot answer Searle’s challenge. Collective outcomes can be produced either strategically or cooperatively; the fact that a collective outcome has been produced is, by itself, not an indication of whether the collective outcome has been produced as a result of collective action or as a result of a set of individual actions. Thus, I argue, collective outcome theories should be rejected as adequate theories of collective action.

In Chapter 3, then, I look at theories of collective action which are not fully reductive, and which purport to have answers to Searle’s challenge. I look first at the team preference theories of Bacharach, Sugden, and Gold. Next, I look at the we-action theory of Bratman. Then, I look at the we-intention theories of Searle and Tuomela. Finally, I look at plural subject theories of collective action, in particular those of Gilbert and of Roth. I argue that
none of these theories is an adequate theory of collective action as presented.

I go on in Chapter 4 to begin building a positive theory of collective action. In this chapter we hew mostly to the individual level; it is in the following chapter that we broaden our focus to include the collective. Here, I first argue that, if we are to account for minimal individualism, we need to have a model of deliberation which can accommodate goals (whether those goals are individual or collective). I propose such a model. I argue further in the third chapter that constrained deliberation is sometimes a rational thing—that it is sometimes rational for an agent to impose constraints on her or his deliberation. I argue that unconstrained deliberation attempts to maximize one’s preferences, whereas constrained deliberation imposes constraints on the maximization of preferences. When we allow for constrained deliberation, we can understand how rational counter-preferential choice is possible. The constraining attitudes, I argue, are intentions or values; I argue that we can conflate intentions and values for the purposes of a model of deliberation. If this argument is correct, then intentions and values allow for the rational constraint of an agent’s maximization of her or his preferences. This rational constraint is a particular kind of rational constraint which I suggest we can label intention rationality. Intention rational agents constrain their deliberation in accordance with their intentions.

In Chapter 5, I argue that we can use our notion of constrained deliberation to account for collective goals. I first argue for the previously mentioned functional identity between intentions and goals. I then look at the ways in which agents can rationally resolve subjective decision problems when those agents hold intentions or goals. In the first case, I look at single agent decision problems, or SADPs; I argue that agents with goals rationally resolve SADPs insofar as they maximize their preferences over outcomes subject to the constraints imposed on them by their goals. In the second case, I look at multi-agent decision problems, or MADPs. Agents rationally resolve MADPs, I argue, when they choose so as to bring about an equilibrium compatible with the goals of all the agents in the decision problem. (These goals might be a set of individual goals, or these goals might be a collective goal.) I then draw a distinction between action intentions and outcome intentions; I argue that agents in SADPs can rationally hold either outcome intentions or action intentions, but that any outcome intention that an agent in a SADP holds must be capable of being translated into an action intention. Similarly, I argue that agents in MADPs must hold action intentions if they are not engaged in collective reasoning, but that they can hold outcome intentions if they are engaged in collective reasoning.
To this end, I argue that action intentions cannot account for collective goals. Nor, I argue, can conditional intentions get us collective goals. Instead, if we want to model collective goals, we must appeal to outcome intentions. I go on to suggest that holding and reasoning according to a collective goal does not violate the own action condition, because in the model of deliberation that I propose goals impose strictly negative constraints and not positive constraints. I end the fourth chapter talking about the limits of my account of collective goals. Collective goals, I argue, do not guarantee coordination. Moreover, we should distinguish collective goals from what I call common goals, which are outcomes that are commonly believed to be desired or preferred by each in a group.

In Chapter 6, then, I recapitulate our theory of collective action and I give examples of how my theory can accommodate the examples of collective action that have been proposed in the literature: Sugden’s footballers’ problem (which is, in essence, a Hi-Lo game of rational choice); Bratman’s example of agents going on a trip together (as, for instance, going to New York together); Gilbert’s example of walking together; Roth’s example of agents taking a trip together (as, for instance, driving to Las Vegas together); and, finally, Searle’s example of cooperative and non-cooperative businessmen. I show how we can make sense of these examples according to the theory I have proposed. Finally, I compare my theory more directly with the theories we consider in Chapter 2 by introducing examples which my theory of collective action can account for, but which the other theories cannot.
Chapter 2

Fully reductive theories and Searle’s challenge

2.1 Introduction

In this chapter I distinguish individual action from collective action. The term ‘individual action’ is a bit misleading, as it might suggest a self-directed or self-regarding motive. We can clarify, then, what we mean by individual action. Individual action is action which can be rationalized by an agent in virtue of its maximizing some value function which the agent *endorses*. Without giving a complete account of what it means for an agent to endorse a value function, we can say here that an agent who endorses some particular value function takes the value function to be reason-providing for the purposes of deliberation. We speak of ‘value functions’ here, so as to be agnostic about what kind of value the function represents. We can give examples of what the value function is not (at least not necessarily): individual actions need not be *utility-maximizing* (for plausible definitions of utility), nor need they be *welfare-maximizing*, nor need they be *self-regarding*, et cetera. Any attempt to to equate value functions with one of these functions, then, is doomed to failure. Nor must individual action be taken in ignorance of expected actions by other agents; individual actions can be (and often are) sensitive to the agent’s beliefs about the expected actions of other agents, and individual actions can depend on the contributions of other agents in order to bring about the desired outcome.

Suppose we take such a definition of individual action. We then face a question: is *all* action that we engage in individual intentional action? Many philosophers (including all
of the philosophers whose views we analyze in Chapter 3) argue that we also engage in a distinctive kind of action—collective action—which is not reducible to individual action as defined. I argue here that such philosophers are correct—we cannot explain collective action away by fully reducing collective action away into a set of individual actions; the resources that a theory of individual action provides us are simply not sufficient to account for collective action.

What resources does a theory of individual action provide us? Though the details are different in many cases, we can speak in general about ‘theories of individual action’. In a set of individual actions, we have a set of agents, we have a set of behaviors which maximize value functions, and we have the outcome produced by the set of behaviors. In order for collective action to be fully reducible to individual actions, there must be some arrangement of behaviors, value functions, and outcomes produced which jointly constitute a collective action.

The typical way of reducing collective actions to individual actions is by identifying some outcome as the collective goal by way of the value functions of the individual agents. In order for such an explanation to work, we need the individual value functions to be such that they can pick out some outcome which is the collective goal; if agents choose behavior so as to bring about the collective goal, they are then thereby engaged in collective action.

There are several different ways that such theorists might try to identify a collective outcome on the basis of individual value functions. They might argue that we can identify the collective goal on the basis of the value functions that agents actually do have; such a theory would argue that the collective goal is the one which is (e.g.) Pareto-maximal or otherwise Pareto-efficient. Or, we might argue that we can identify the collective goal on the basis of the individual value functions that the agents should have; such a theory (e.g. collective utilitarianism) would argue that if everyone holds the same individual value function—maximize utility—then the outcome that each individual wants to bring about will be the collective goal. Another possibility is that collective action occurs when each individual has the same (or sufficiently similar) value function as applied to the decision problem under question, regardless of what the value function is.

None of these approaches work; they all run into what we can call Searle’s challenge. Searle’s challenge is that collective goals by themselves are never sufficient to account for collective action, because agents might do their part in bringing about a collective goal through the exercise of their individual agency. Indeed, they might bring about a collective
goal in spite of (or perhaps because of) an explicit disavowal of any collective intentionality.

2.2 Individual action

Let us start by defining individual action more precisely. Individual action is individual intentional action; as we will use the term, the claim that some behavior \( x \) is agent \( X \)’s action entails that \( X \) intentionally did \( x \).\(^1\) This does not lighten our explanatory burden, of course, but merely shifts it to providing an explanation of what it means for an agent to intentionally perform some behavior.

The outline of a plausible approach to understanding intentional action was offered by Davidson, beginning with his ‘Actions, Reasons, and Causes’ paper in 1963.\(^2\) Davidson suggested that individual action is distinguished by two features: first, action is behavior that can be rationalized according to some values held by the agent at the time of the behavior; and second, that rationalization is a species of causal explanation—that is, the reasons an agent \( X \) perceives for doing some action \( x \) are also the cause of her doing \( x \).\(^3\) For Davidson, rationalizations of behavior appeal solely to the agent’s beliefs and desires held at the time of action.

Now, a Davidsonian theory of action has been frequently criticized, and rightly so.\(^4\) As has been noted, Davidson’s theory does not allow for a volitional component to action. Nor can Davidson easily account for the mental state of having an intention; since having an intention is important for planning and future agency, a theory of intentional action which fails to account for the state of having an intention is an impoverished theory.\(^5\)

Importantly, however, none of these criticisms attack the central idea that intentional

---

\(^1\)There are contexts in which we might want to attribute behaviors to agents without making the stronger claim that the agent performed the behavior intentionally. In particular, we might attribute some behavior to an agent in a legal sense, or in a moral sense, without making the further claim that the behavior was the agent’s action. If agent \( X \)’s irresponsible and negligent behavior causes person \( Y \) to die, we might well say that \( Y \)’s death is morally and legally attributable to \( X \). Nevertheless, we need not hold that \( Y \)’s death is \( X \)’s action—\( X \) might not have killed \( Y \) intentionally, perhaps because she was ignorant of the causal effects of her behavior.

\(^2\)Davidson 1963.

\(^3\)I follow Davidson throughout in calling reason-based explanations of behavior ‘rationalizations’. In calling such explanations rationalizations, I do not intend to suggest that they are false explanations, or that they necessarily fail to describe the actual causal explanation of the behavior in question. Our ordinary use of the term ‘rationalization’ sometimes suggests such distinctions; my use of the term here is not meant to perfectly follow ordinary language usage.


\(^5\)See Michael Bratman in Bratman 1999a and Bratman 1984.
behavior is behavior that can be rationalized in some way. The common feature of the criticisms mentioned above is that we cannot have a complete theory of action which appeals solely to the facts that an action is rationalizable, and that the rationalization stand in the right causal connection to the behavior. A complete theory of action must do more than sort behavior into instances of action and instances of non-action; it also must provide an account of those mental states and processes that lead to action. The most common objections to a Davidsonian theory is that the beliefs and desires are not together sufficient to account for all relevant mental states and processes that lead to action.

The rationalizability thesis—that a necessary condition on X’s behavior x being X’s action is that X can rationalize performing x—is thus still up in the air after the various criticisms to Davidson’s view mentioned above have been considered. I think we should accept the rationalizability thesis, although we need not follow Davidson in accepting that it is solely beliefs and desires which rationalize behaviors.

Davidson suggests that we can ‘[define] an intentional action as one done for a reason.’6 This suggestion is on the right track, although it should be acknowledged that this formulation does not rule out very much; having a reason is a matter of merely having some appropriate belief and pro-attitude. If I intentionally perform some action—spontaneously jumping in the air, for instance—Davidson suggests that I have an appropriate pro-attitude of jumping in the air, even if I have no other pro-attitude in favor of the jumping. As he writes,

> When we [treat wanting as a genus including all pro-attitudes as species] and when we know some action is intentional, it is empty to add that the agent wanted to do it. In such cases, it is easy to answer the question ‘Why did you do it?’ with ‘For no reason’, meaning not that there is no reason but that there is no further reason, no reason that cannot be inferred from the fact that the action was done intentionally; no reason, in other words, besides wanting to do it.7

Importantly, Davidson does not explicitly require that my pro-attitude in favor of jumping in the air be stronger than my pro-attitudes in favor of other options available to me at the time.8 My desire to jump in the air might be strongly outweighed by, for instance, my desire to act with propriety. If I nevertheless jump, I have a reason to jump—my weak

---

6Davidson 1963, p. 688.
7Davidson 1963, p. 688, emphasis original.
8Although not explicit, Davidson’s causal theory of action can be interpreted as having such a requirement.
desire to jump—but, it seems, I have a much stronger reason to not jump. Accordingly, any behavior can be my action—just as long as I have *some* desire to do it, no matter how weak, or how outweighed by other desires.

Whether agents ever intentionally act contrary to their strongest perceived reason is a matter of philosophical dispute. If we are skeptical about the possibility of akratic action—if we believe that agents *never* intentionally act contrary to their strongest perceived reason for acting—then we can define actions to be behavior rationalized by one’s strongest perceived reason for acting. If, on the other hand, we accept that akratic action is possible, then we should define action to be behavior which is rationalized by *some* perceived reason for acting, though not necessarily the strongest such reason.

In either event, actions which run counter to an agent’s strongest perceived reason for acting are *internally irrational*. We appeal here to Bernard Williams’s distinction between internal and external reasons for acting: if an action is justified by elements of an agent’s subjective motivational set, it is an internally rational action; if the action is rationally justified by things that are not part of the agent’s subjective motivational set, it is externally rational. Like Williams, our concern is with internally rational and irrational behavior. Though it is possible to imagine actions which are internally irrational for agent *X* but externally rational (*X* saves the busload of toddlers from the burning wreckage, though she thinks herself much more justified in bypassing the accident and going to the casino), we should accept that such actions are irrational *by the agent’s own lights* at the time the action is performed. Further, an agent’s behavior might be externally rational even though it is not internally justified by *any* element of the agent’s subjective motivational set. *X*’s behavior might have the effect of saving the busload of toddlers, even though she sees absolutely no rational justification in doing so. In such cases, we should be properly hesitant to call it an action at all—at least under the description which makes the behavior externally rational. If saving the busload of children is not justified at all by *X*’s subjective motivational set, then we should be hesitant to call saving the busload of children *X*’s action. The behavior might be her action under a different description—*X* might have been attempting the steal the candy that the toddlers were eating, and saw it first necessary to first remove the babies from the flames—but given that the act of *saving the babies* simpliciter was not rationalizable according to *X*’s subjective motivational set, her saving the babies could not have been her action under that description.

---

9Williams 1981.
We will take intentional action, then, to be behavior which is rationalizable (under some description) according to an agent’s subjective motivational set. Further, we will take rational intentional action to be behavior which is (under some description) the most rationalizable according to the agent’s subjective motivational set.

In saying that rational action is action which is the most rationalizable, we are adopting a maximizing view of rational agency. Such a view is challenged by the claim that certain types of bounded reasoning can produce rational behavior, as for example in Herbert Simon’s concept of satisficing.\textsuperscript{10} The philosophical justification for the claim that satisficing reasoning is rational reasoning is unclear, however. Some assume that satisficing reasoning is justifiable only insofar as it leads to outcomes which are more rational than outcomes which would have been produced without satisficing reasoning, as might happen when deliberation time is costly, or when there are environmental constraints which prevent the possibility of unbounded reasoning. This strategy, though, just takes rational satisficing reasoning to be another form of maximizing reasoning. There are other justifications of satisficing reasoning, such as ethical justifications or justifications grounded in practical reasoning.\textsuperscript{11} These approach aim to justify satisficing while admitting that satisficing does not maximize an agent’s underlying value function. Non-maximizing justifications of satisficing are less persuasive than maximizing justification; it is hard to identify in virtue of what satisficing is rational if it is sub-optimal.\textsuperscript{12} Nevertheless, the strong intuition remains that some actions are ‘rational enough’ to count as rational actions, in spite of the fact that they are not the most rational action that could be done.

We will, however, not argue against a satisficing conception of rational behavior here. We will adopt a maximizing view of rational action, such that $X$ acts rationally so long as her behavior is the most rationalizable behavior available to her given her subjective motivational set. If a satisficing picture of rational behavior is the one that we ought to adopt, then the arguments presented in this chapter will not directly apply. We should be skeptical, however, of the claim that a satisficing conception of individual rational action will permit us to account for collective action in ways that a maximizing conception of

\textsuperscript{10}See, for instance, Simon 1955 and Simon 1959. For a sample defense of the rationality of satisficing reasoning in artificial intelligence, see Pollack 1991; for psychology, see Gigerenzer and Goldstein 1996.
\textsuperscript{12}Again, we are looking at satisficing as a normative theory of rational action and not as a descriptive theory. We thus do not evaluate the claim that we as human beings do frequently engage in satisficing reasoning.
individual rational action cannot. Although some of the arguments I present below attack the maximizing nature of individual rational action, the more substantive worry comes from the individual nature of individual rational action. And, although the account of collective action I provide later shares with satisficing theories a rejection of the claim that the only rational action in a decision problem is one that maximizes a value function, we should nevertheless reject the view that we can account for collective action through agents who justify their actions according to satisficing conceptions of individual rationality.

2.2.1 Rationalizations

In claiming that individual intentional action is action which is rationalizable according to some subjective value function, we are not giving a complete story. That some behavior is rationalizable does not yet mean it is someone’s action; the rationalizability of the action might be accidental to its performance. To put it in Wittgensteinian terms, the fact that my arm goes up might be rationalizable; I might recognize the behavior as providing great benefits to me. This does not suffice to make it my action, however; my arm’s going up might have been simply a spasm. To address this worry, Davidson argued that the rationalization must cause the action. Others agree with Davidson that actions must be related causally to the reasons the agent possesses, but they argue that the action must be causally related in the right way to the rationalizing mental states. To cite an example by Robert Audi, I might have rationalizing beliefs and desires in favor of waking you up; these beliefs and desires might cause me to be nervous, which in turn might cause my dropping a tray of dishes outside your door, which might in turn cause my waking you up. Though the rationalizing mental attitudes cause the behavior, the causal pathway is deviant, and so we do not have a genuine case of action.\(^{13}\)

At the same time, some have been critical of the claim that actions are caused by rationalizing mental states; the worry is that causal explanations and reasons explanations are incompatible. Jennifer Hornsby and Jaegwon Kim have each presented arguments against the explanatory sufficiency of a causal theory of action.\(^{14}\) Hornsby’s concern is that any events-based theory of agency is going to necessarily omit the role of the agent in causing action; actions are not mere events in search of (e.g. rationalizing) causes, since some behavior’s being an action presupposes an agent. Kim’s criticism of Davidson’s causal theory

---

\(^{13}\) Audi 1986.

of action consists primarily of the claim that Davidson seems to posit two distinct kinds of explanation of the same action: one is a rationalizing explanation (appealing to beliefs and desires), and the other is a causal explanation (appealing to brain states and their causal powers). Both explanations cannot be independently sufficient, Kim argues; the principle of explanatory exclusion dictates that we cannot have two unique and complete explanations of the same explanandum. Now, Kim does not make the further argument that we should therefore accept or reject a causal explanation of action. However, it is plausible to think that there is some causal explanation of action available to us. If there is, and if Kim is correct, then it seems we should be skeptical that a rationalization-based explanation of action could be complete or independent. However, we should note that a physiological causal explanation of behavior need not appeal to rationalizations at all—rationalizations are psychological, and we can have a physiological account of behavior which does not reference the psychological at all. If we want to define actions as properly rationalized behavior, then, we cannot exclude rationalizations from our theory of action.

Rather than directly address the question of whether there needs to be a causal connection between rationalizations and behavior in order for the behavior to count as action, I propose, in effect, that we avoid the question altogether. We should draw a distinction between a rationalization causing an action, and a rationalization factoring in a causal explanation of an action. Rationalizations are token mental states; if rationalizations cause behavior, then there must be general mental laws which entail that those token mental states, insofar as they fall under particular mental types, cause the relevant token behavior. However, the causal relationship, if it exists, holds not between the mental type and the event type, but rather between the mental particular and the event particular (i.e. the action).

As Davidson notes, however, whether a causal relationship holds between two particular events is independent of whether we have a causal explanation of the behavior. Explanations in general, whatever they are, do not exist at the level of causal relationships between particulars; the fact that this billiard ball caused that billiard ball to move in a certain direction does not explain the movement of the second ball. What we need for a proper explanation are general laws of physics covering the interaction of medium-sized objects. This holds even more for explanations of actions: the existence of a causal relationship between a mental particular and some particular behavior does not suffice to explain the behavior as

\[15^{Davidson\ 1993.}\]
action (even if it might suffice to explain the behavior as behavior).\footnote{We might still be skeptical that the existence of such a causal relationship between a mental particular and some particular behavior constitutes an explanation of the behavior if the causal explanation is always at the type level.} Whether my raising my arm is an action does not depend on whether my arm-raising was caused by some particular rationalization mental state; what matters is whether there exists some general law (or sufficiently robust regularity) connecting that mental state qua rationalization to the behavior qua action. In explaining actions, the way we describe the relevant mental states matters, as does the way we describe the choice situation. That I see my raising my arm as a way to say hello, and as the best available way to say hello, are relevant in explaining the arm raising as my action. Further, that I saw that particular arm-raising as of the type of action licensed by my reasons is relevant for whether the arm-raising is my action.

Once we see that the task of finding an explanation of action is distinct from determining whether a causal relationship exists between a mental particular and behavior, we can bracket the question of whether a causal relationship exists at all between the rationalization and the action. The reason is that even if a causal connection between a token rationalization and a token behavior is a prerequisite for that behavior’s being an action, understanding that causal connection is irrelevant if we wish to give an explanation of the action. Since our interest is in giving an explanation of individual rational action, we need not answer here the question of whether a causal relationship must exist, and if so, what it must be like.

On this point, we are adopting a version of what Neil Campbell calls \textit{explanatory epiphenomenalism}: the identification of agent X’s mental properties does not provide a causal explanation of X’s actions.\footnote{Campbell 2005.} Campbell argues that Davidson’s anomalous monism entails explanatory epiphenomenalism. Regardless of whether this is true—and regardless of whether we accept Davidson’s anomalous monism or not—we should agree with Campbell that explanatory epiphenomenalism is a plausible and benign epiphenomenalism, and one which we might have reason to accept independent of our beliefs about anomalous monism. Accepting explanatory epiphenomenalism only increases the pressure on us to focus on those mental states and properties which rationalize an agent’s action, rather than wonder whether the right sort of causal relationship holds between the mental states and the behavior.
2.2.2 Constraints on value functions

If the preceding argument is correct, then we can explain a rational agent’s performing some action by appealing to the agent’s subjective value function, the maximization of which suffices to rationalize the action. This picture of rational individual agency is a familiar one in philosophy and the social sciences; it is the consequentialist genus of which the utilitarian and *homo economicus* are species.

Since we are modeling individual rational action, we might wonder whether there are any constraints on the value function that the individual rational agent aims to maximize. Though there might be *structural* constraints on how an agent’s values over outcomes relate to one another, there are no (or almost no) *contentful* constraints on which outcomes an individually rational agent can value over others.

What do we mean by the claim that there might be structural constraints on an agent’s value function? We can, for simplicity, argue that every rational agent’s value function will be *weakly complete* and *transitive*. A value function is *weakly complete* if, for any two outcomes in the domain of the agent’s deliberation, the agent values one outcome over the other or is indifferent between them. It is *transitive* if, whenever an agent values one outcome $o_1$ over a second outcome $o_2$ and also values the second outcome $o_2$ over a third outcome $o_3$, the agent values $o_1$ over $o_3$. We adopt these assumptions for simplicity, although they are not essential for the argument that follows.\(^\text{18}\)

The preceding are *structural* constraints, in that the constraints on the value functions appeal only to formal properties of the valued outcomes, and not to the content of the valued outcomes. It is important to note that in order to capture individual action, we do not need to impose *contentful* constraints on value functions, either—an agent’s rationality does not depend on her or his valuing certain outcomes over other outcomes.

In particular, a rational value function need not be one that maximizes agent-independent utility; it need not maximize an agent’s welfare; it need not track the agent’s self-regarding preferences; and so forth. Now, this claim might seem especially strong; one might worry that it is the Trojan horse within which we are smuggling all we need to successfully attack rival theories. The claim is not meant to be controversial, however; in fact I think that it is a claim that we can easily accept. An agent’s deeply held values are rarely captured in welfare functions; the agent who values outcomes which do not maximize welfare (as, for

\(^{18}\text{There is not agreement in the literature about whether these conditions are even intelligible, let alone whether they are binding for the value functions of rational agents. See Rabinowicz 2000, Levi 2002.} \)
example, the agent who holds a trivial but satisfying hobby) is still an individual reasoner. Further, insofar as the agent acts to maximize that value function, the agent is rational; an agent’s reasons for acting derive from the values the agent holds. Likewise, the agent who values altruistic acts over self-interested acts does not therefore cease to be an individual reasoner; consistently bringing about outcomes in which other people benefit is rational for the agent who values bringing about precisely those outcomes.

Now, some methodological individualists would reject our claim that there are no contentful restrictions on value functions. Evolutionary game theorists, for example, tie individual value functions of rational agents to welfare and fitness, so that an agent is not rational if her or his (or its) value function does not value outcomes in which the agent has increased fitness over those outcomes in which the agent has decreased fitness.\[^{19}\] Insofar as utilitarians offer a theory of rational choice (as opposed to moral choice, which many utilitarians see themselves as offering), they tie value functions to agent-independent utility.\[^{20}\] Neoclassical economics ties value functions to agent-dependent utility, i.e. to the psychological satisfaction produced by the production of an outcome or the acquisition or consumption of goods.\[^{21}\] All of these approaches restrict a rational agent’s value function according to the content of the outcomes which agents might deliberate over, and so we might think that the methodological individualist can accept restrictions on value functions.

The methodological individualists who make this argument are mistaken, however. Each of the examples listed above presupposes a substantive view of the nature of rationality: the evolutionary game theorist presupposes that to be a rational agent is maximize one’s own reproductive fitness; the utilitarian presupposes that to be a rational agent is to maximize the amount of utility produced by one’s actions; the neoclassical economist presupposes that to be a rational agent is to maximize one’s own personal utility; and so forth. These presuppositions are rarely, if ever, defended with argument. Accordingly, rather than present arguments against these substantive views of rationality, I merely note that we can label these different views of rationality as ‘rationality-sub’: rationality-sub-evolutionary game theory, rationality-sub-neoclassical economics, et cetera. I am presenting here a view of rational individual agency which appeal to the general notion of rationality; if necessary, we can substitute a rationality-sub without great consequence for the overall argument.

\[^{19}\text{Skyrms 1996, Hofbauer and Sigmund 1998.}\]
\[^{20}\text{Mill 1979.}\]
\[^{21}\text{Sen 1973.}\]
2.3 Is collective action reducible to individual action?

The definition of rational individual action we arrived at in the previous section is of behavior that is rationalizable according to some individual agent’s value function—that is, a rational action is one which maximizes the value of the outcome produced according to the agent’s value function within the context of a decision problem.\(^{22}\) Those who accept what we can call strict methodological individualism argue that we can account for every kind of action, whether collective or individual, by appealing to individual intentional actions. Strict methodological individualism should be distinguished from metaphysical methodological individualism, which argues that we can only appeal to individual agents in accounting for individual and collective behavior, although no claim is made about whether the individual agents are performing individual intentional actions or not.\(^{23}\) Occam’s razor requires us to prefer methodological individualism, whether strict or metaphysical, whenever we can do so while holding the range of explananda constant. Whether we can accept metaphysical methodological individualism will be considered in Chapter 3. Here, we consider strict methodological individualism. The strict methodological individualist argues that every action that we engage is in a species of individual intentional action: that is, there is no action which is not either an individual action, or comprised of a series of individual intentional actions.

As stated, the strict methodological individualist position is not yet criticizable; the position does not rule out the possibility that instances of collective action are caused by systematic irrationalities on the part of the participants. Suppose, however, we think that participation in a collective action can be rational behavior on the part of the participants—that agents in a collective action can rationalize their doing their part in a collective action. Can the strict methodological individualist account for such instances of rational collective action?

I argue here that the strict methodological individualist cannot account for collective action. We can describe a general way that strict methodological individualists attempt to account for collective action: namely, the collective outcome approach. This approach, I argue, cannot be successful. If we are to account for collective action, then, we must

---

\(^{22}\)For simplicity, we can speak of ‘maximizing a value function’ rather than ‘maximizing the value of the outcome produced according to a value function’.

\(^{23}\)Many, although not all, of the theories of collective action we will analyze in Chapter 3 endorse metaphysical methodological individualism, as does the theory I propose later.
broaden our conception of agency.

Let us first define the collective outcome approach: individually reasoning agents are engaged in a collective action so long as they bring about an outcome which is identified as a collective outcome through their intentional actions. Collective outcomes can be identified either by appeal to the value functions that the agents do have, or by appeal to ideal value functions, regardless of whether the agents hold them or not. Agent-independent utilitarians such as Donald Regan identify collective outcomes by appealing to ideal value functions; they argue that collective outcomes should be defined to be those that entail the production of the greatest amount of utility. Insofar as agents adopt the utilitarian value function and act accordingly, then, they are engaged in rational collective action.

Another option for the collective outcome theorist is to define collective outcomes according to the value functions that the agents in the choice situation actually possess. We might argue, for instance, that collective outcomes are those which are Pareto maximal, or otherwise Pareto efficient compared to a baseline expected outcome such as an equilibrium. This is the view proposed by, among others, David Gauthier.

Collective outcomes can be defined even more narrowly: we might define collective outcomes as, essentially, the mutual cooperation outcome in a Prisoner’s dilemma game. In defining collective outcomes this way, we would be highlighting the characteristic features of Prisoner’s dilemmas: they feature a Pareto-superior coordination outcome which is strongly dominated by a unique Pareto-inferior equilibrium. As a result, in order to attain the benefits of the ‘collective outcome’, agents must act contrary to a dominant strategy.

All of the above approaches define collective outcomes as derived by the value functions that agents do or should hold during a decision problem. Note that the collective outcome approach does not beg the question of whether the agents are engaged in a collective action. We might worry that it does; we might think that by identifying some outcome as a collective outcome, we are begging the question as to whether bringing about that outcome is a collective activity or not. It does not beg the question, because we have an independent way of knowing whether an outcome is a ‘collective outcome’ or not, corresponding with the particular definition of collective outcomes adopted. The term ‘collective outcome’ is just a label, and it picks out some outcome or outcomes which possess some interesting

---

24See Regan 1980.
26Luis Fernando Medina seems to define collective outcomes as cooperation outcomes in Prisoner’s dilemma games in Medina 2007.
properties. We need not even call such outcomes ‘collective outcomes’; what is important is not that they are identified as ‘collective outcomes’ by the agents, but rather that the agents intentionally bring the outcome about.

Once the collective outcome has been determined, the only thing remaining to be done is to solve the motivation problem: how do you motivate each agent to act so as to bring about the collective outcome? If the collective outcome maximizes each agent’s (actual or ideal) value function, as is the case for agent-independent utilitarians, then the motivation problem is effectively solved; we can assume that rational agents are motivated to do what they have most reason to do. If the collective action does not maximize each agent’s (actual or ideal) value function, then each agent must somehow find it rational to act on behalf of the collective outcome even though it does not maximize the individual value functions; this is the classic problem of the rationality of collective action.

If, then, we accept the claim of collective outcome theorists that collective action occurs when agents jointly produce a collective outcome, then we face the additional problem of figuring out how to make it rational for agents to produce the collective outcome. This problem is only relevant, however, if the definition of collective action proposed by the collective outcome approach is a good one. Unfortunately, it is not. As we will see in the remainder of this chapter, the collective outcome theorist does not have a satisfactory account of collective action. Collective outcomes might be interesting in their own right, depending on the properties in virtue of which they are identified as collective outcomes. Collective outcomes cannot be the basis of a theory of collective action, however, because of a potent counter-example prominently raised by Searle.

2.4 Searle’s challenge

We can understand this counter-example as follows. In defining collective action as a set of agents acting on behalf of a collective outcome, the collective outcome theorist fails to be able to distinguish between two cases: one in which each agent performs the action required to produce the collective outcome, and they do so collectively; and a second, in which each agent acts so as to produce the collective outcome, but there is no collective action involved in the production of the outcome at all. It is the former case which is a genuine instance of collective action. The latter case, on the other hand, does not guarantee that the production of the collective outcome is, in fact, a collective action.
This argument was most clearly expressed by John Searle in his 1990 paper ‘Collective Intentions and Actions’, and we can call it Searle’s challenge. Searle’s conclusion takes the form of two theses:

**Thesis 1:** There really is such a thing as collective intentional behavior which is not the same as the summation of individual intentional behavior.

and

**Thesis 2:** We-intentions cannot be analyzed into sets of I-intentions, even I-intentions supplemented with beliefs, including mutual beliefs, about the intentions of other members of a group.

Although Searle directs his argument against intention-based analyses of action, his argument applies equally well against the rational choice-based account of action that we have been developing so far. We can assume, for simplicity, that a rational agent $X$ who holds a value function which is maximized by $X$’s doing action $x$ will have an intention (or, in Searle’s terminology, an ‘I-intention’) to do $x$. As a corollary of thesis 1, Searle argues that agents engaged in collective intentional behavior will each have a ‘we-intention’ in favor of the collective activity. Thesis 2 then amounts to the claim that the collective outcome theorist cannot account for we-intentions, and therefore cannot account for collective action.

In order to illustrate thesis 2, Searle introduces his capitalist businessmen example: suppose a group of businessmen each believe in Adam Smith’s invisible hand, so that they each believe that humanity benefits when everyone pursues her or his individual, selfish interest. Each businessman forms the intention to act in pursuit of the collectively beneficial outcome where everyone pursues her or his selfish interest. Further, each businessman has the belief that the other businessmen will act likewise; they thus each believe that their selfish actions will be effective in bringing about the collectively beneficial outcome.

The collective outcome theorist ought to hold that the capitalist businessmen are engaged in a collective activity. Their individual value functions are such that they each most highly value the outcome in which everyone pursues her or his individual selfish interest;

---

27 My use of Searle’s argument does not follow Searle exactly, and I do not claim that Searle endorses the argument I present here.
30 We will look more carefully at intention-based theories of action in Chapter 3.
31 I here intend the terms ‘businessman’ and ‘businessmen’ in a gender-neutral fashion.
that outcome is thereby defined as the collective outcome.\footnote{Note that we have been a bit vague about how exactly we determine the collective outcome; we might, for instance, define the outcome in which everyone acts selfishly to be the collective outcome in virtue of the fact that it maximizes each agent’s welfare, or in virtue of the fact that each agent values selfish action. The first appeals to agents’ ideal value functions; the second appeal to their actual value functions. Regardless of how we define the collective outcome, it seems clear in this example that the selfish outcome is in fact the collective outcome—assuming that the agents have true beliefs about humanity’s benefitting by their each pursuing their individual selfish interests.} We then have a set of individual agents each acting in such a way that they intentionally bring about the collective outcome. Assuming their actions are successful, we ought to conclude that the agents are engaged in a collective action.

As Searle notes, however, we have reason to believe that the capitalist businessmen are \textit{not} engaged in a collective action. Indeed, according to Searle the capitalist businessmen might be guided by an ideology which entails that there should not be any we-intention. This is not to say that it is \textit{impossible} for the businessmen to have a we-intention; the businessmen might, Searle suggests, form a pact to always act in pursuit of their selfish interests, so that they might collectively benefit humanity. In such a case, the businessmen would have a we-intention, and their bringing about the collective outcome (if successful) would be a collective action. If, however, the agents do not have such a we-intention, they are not engaged in a collective action, regardless of what action each agent performs, and regardless of whether the agents successfully bring about the collective outcome.

We will return later to investigate whether Searle is correct to argue that collective actions require non-reducible we-intentions. For now, it suffices to note that Searle’s challenge generalizes to collective outcome theories of collective action. Suppose that $o$ is the collective outcome (however this is determined) for a set of agents $N$ facing some decision problem, and suppose that $A^* = \langle x_1, \ldots, x_n \rangle$ is the set (or, without loss of generality, a set, if there are multiple) of actions which produces the outcome $o$. Searle’s challenge says simply that whether the agents in $N$ perform the actions in $A^*$ does not suffice to determine whether the agents are engaged in a collective action; we can always imagine two similar but distinct scenarios, one in which the agents’ performance of $A^*$ and production of $o$ is a collective action, and one in which it is not. (Again, we leave until later the question of whether performing the actions in $A^*$ is \textit{necessary}, as well as the question of what else a satisfactory theory requires.) The reason is that each agent might have $o$ as her or his individual goal, and see performing her or his action ($x_i$ for agent $i$) in $A^*$ as the necessary
means to bringing about \( o \). Further, each agent might pursue \( o \) with the explicit understanding that her or his action is an individual action, in pursuit of an individual goal, and that in addition it is not part of a cooperative project. \( X \)'s achieving her goal might benefit \( Y \), and vice versa, but this fact does not suffice to make the goal a collective goal, since they might each hold and pursue the collective goal without any concern for whether the goal is, as it were, collectively held.

Collective outcome theories define cooperation in terms of whether a set of agents each performs actions which, collectively, bring about a particular outcome. Searle's challenge is that such theories cannot suffice to account for collective action, because collective outcomes can be brought about through individually-minded agents, each of who rejects the claim that their action is part of a collective action. Now, it might be argued that the agents are simply incorrect—that they are engaged in a collective action because they brought about a collective outcome, even though (like the capitalist businessmen) they do not think of themselves as doing a collective action, and indeed strongly believe that they ought not be engaged in a collective action. This would be a strained interpretation, however. The fact that the agents do not think of themselves as engaging in a collective action ought to give us prima facie reason to believe that the agents are not, in fact, engaged in a collective action; if we want to override the agents' own judgment, we ought to have convincing reason.

Do we have such reason? We do not. The collective outcome is called a collective outcome for convenience only; by calling it this, however, we run the risk of begging the question. Let us instead call the collective outcome the \( c\)-outcome. The collective outcome theorist then needs to explain why agents' acting to bring about a \( c\)-outcome suffices to be a collective activity. It cannot be merely that every agent benefits (according to her or his real or ideal value function) if the \( c\)-outcome is brought about, as there are many mutually beneficial outcomes the production of which does not count as a collective activity. As one example, two agents \( X \) and \( Y \) might be engaged in a market exchange where each unsuccessfully attempts to cheat the other. Because their unsuccessful cheating results in the production of a fair market outcome, they bring about an outcome which benefits each (\( X \) needs sprockets and has cogs, while \( Y \) needs cogs and needs sprockets). We would not call their exchange of cogs and sprockets a collective activity, however, as they each performed their actions with the sole goal of producing the best possible outcome for themselves. The example generalizes to any number of free exchanges; the fact that a particular outcome is intentionally produced and mutually beneficial does not mean that
the activity that produced the outcome must be considered a collective one.

Nor can c-outcome theories avoid Searle’s challenge by adding in agential beliefs about the likely behavior of other agents. Even if $X$ and $Y$ each believe that the other agent will act in the expected way to produce the c-outcome, that does not suffice to make the outcome produced a collective outcome. $X$ might believe that $Y$ will try to cheat her and get two cogs for each sprocket, but will eventually settle for one cog per sprocket; likewise, $Y$ might believe that $X$ will try to cheat him and get two sprockets for each cog, but will settle on exchanging one cog per sprocket. $X$ and $Y$ each act as predicted, and they produce the outcome in which they exchange cogs and sprockets at a rate of one for one. The outcome is mutually valued, in that they each value exchanging over not; the outcome is, thus, a c-outcome. Nevertheless, $X$ and $Y$ are not engaged in a collective action.

Further, this problem is not just limited to situations of market exchanges.\textsuperscript{33} By definition, c-outcomes are produced from a set of actions performed by a set of agents. Whether we consider the production of the c-outcome to be a collective activity, however, depends at least in part on whether the agents consider their own contribution to the c-outcome to be a part of a collective action or a part of an individual action (however this is analyzed). The collective outcome theorist, however, simply does not have the resources needed in order to account for this difference. The only way for the collective outcome theorist to do this would be to build the collectivity into the c-outcome, so that when each agent acts in order to produce the c-outcome, they are thereby acting in pursuit of a genuinely collective outcome (and, presumably, are thereby engaged in a collective action). This strategy is viciously circular, however. In begging the question of what a collective outcome is, it does not help us to analyze the concept of collective action.

Because of Searle’s challenge, then, the collective outcome theorist cannot define collective action strictly in terms of whether a set of agents is successful in producing a collective outcome. Does the collective outcome theorist have any other options? The collective action theorist might insist that what matters is not merely that a collective action is produced, but rather the fact that the agents understand the outcome to be a collective outcome, and thus understand their actions to be contributions to a collective action.

Unfortunately, there is no clear way of fleshing this out without begging the question. For the collective outcome theorist, agents individually justify to themselves the performance

\textsuperscript{33}Or, if we do limit this to market exchanges, we have to acknowledge that every collective action has an equivalent which is analyzable as a market exchange.
of individual actions. The resulting outcome must be a direct result of those (individually justified) actions, as the collective outcome theorist would fail to have an intelligible theory of action otherwise. If an agent is to understand an action to be a contribution to a collective outcome, the agent must first understand the outcome to be a collective outcome, as outcomes are correlated with sets of actions. In virtue of what, though, can an agent believe an outcome to be a collective outcome? The set of collective outcomes $C$ can be defined either extensionally or intensionally. If the agents understand outcomes to be collective extensionally, then they run into the same problem we ran into above: the agents still lack any way of connecting the fact that some outcome $o^*$ is in the set $C$ with the claim that acting to bring about $o^*$ is thereby a collective action. It could be entirely arbitrary that $o^*$ is in $C$. If, on the other hand, the agents understand the outcome $o^*$ to be in $C$ intensionally, then the agents must have a concept of ‘collectivity’ that they apply to various outcomes in order to determine whether it ought to be a part of $C$ or not. Our problem, however, is precisely to understand what that concept is; we cannot merely help ourselves to it without first providing an analysis.

2.5 Answering Searle’s challenge

Collective outcome theories thus have a very difficult time in answering Searle’s challenge. If we are to account for collective action, then, we must find a different framework in which to do so. Further, any proposed account of collective action we give must be sensitive to the challenge Searle poses.

Any theory of collective action which can answer Searle’s challenge must be able to distinguish between when the production of an outcome is a collective activity, and when a similar production of the same outcome is merely a set of individual activities. In this sense, Searle’s challenge is a collective action analogue of Wittgenstein’s question about action. Wittgenstein asked what was left over when one subtracted the fact that one’s arm went up from the fact that one intentionally raised the arm.\footnote{Wittgenstein 1973, I, §621.} Searle’s challenge, then, asks us what is left over when we subtract the fact that agents brought about a collective outcome from the fact that the agents acted collectively. The collective outcome theorist argues that nothing is left over. As we have seen, however, this argument is inadequate. Consequently, we must reject strict methodological individualism, as it does not permit us to account for
the difference in the two cases.

A number of philosophers and social scientists have proposed analyses of collective action which attempt to fill in the bridge between the mere occurrence of a collective outcome and a collective action. In the following chapters, we will look at the most promising of these attempts to analyze collective action in ways which also are adequate responses to Searle’s challenge. As we will see, none of the proposed solutions is entirely satisfactory; some fail to avoid the force of Searle’s challenge, while others fail as general solutions to the problem of collective action. The problems with these theories will prove instructive for the analysis of collective action that I propose in Chapter 4 through Chapter 6.
Chapter 3

Theories of collective action

3.1 Introduction

As we saw above, Searle’s challenge poses a problem for any fully reductive theory of collective action. If the argument is correct, then fully reductive theories are incapable of explaining collective action. We need something else; we need to build the collectivity in to our theory of action at some point if we are to account for collective action. The problem we face is choosing the right point at which to build the irreducible collectivity. There are numerous theories which are not fully reductive, and so might be more successful at answering Searle’s challenge. There is a wide range of ways in which the irreducible collectivity might be built into a theory of collective action. In this chapter, we will look at various theories of collective action, though—as we will see—none of the theories on offer are acceptable as they stand.

The first group of theories to be considered will be called the team preference approach to collective action. The team preference theorists we will consider here are Michael Bacharach, and Robert Sugden, and Natalie Gold. They each argue that there are two primary differences between outcomes produced through individual goals and outcomes produced through collective goals. The first difference concerns the difference between individual and collective goals: collective goals are generated by team preference functions, as distinct from individual goals, which are generated by individual preference functions. The second difference concerns the reasoning process through which the outcomes are produced: agents who reason about what to do on the basis of team preference functions engage in team reasoning, whereas agents who reason about what to do on the basis of individual preference functions
engage in individual reasoning. Being a member of a team, then, requires that agents both hold team team preferences, and also engage in team reasoning.

As we will see, the team preference approach faces one major problem. Agents who share a collective goal are also required to share a team preference function. Regardless of how we interpret this ‘sharing’, however, this condition is too strong; we should allow the possibility of a gap between any individual agent’s team preference function and the team’s collective goal. Preferences are not the same as goals; just as individuals can have goals which are not reducible to the agent’s individual preferences, so too can collectives have goals which are not reducible to a collective preference function.

Later, we will look at what we will call the we-action approach, of which the main proponent is Michael Bratman. Bratman’s analysis of collective action requires that agents in a collective action share intentions to bring about some collective goal or engage in some collective action. In Bratman’s analysis, agents who share an intention must each individually posses an intention which has as its content that they (namely, the agents in the collective) perform the shared action or bring about the collective goal. Ontologically, Bratman’s account of collective action makes reference only to individuals and not to groups. For Bratman, this means that these intentions with collective action as their content must be held by individual agents. This leads Bratman into a problem, however: if individual agents hold intentions about collective actions, then they must also intend the actions of other agents. This is a violation of a very plausible condition on theories of action, namely that agents can only intend actions or outcomes that they themselves directly control.

We might, then, despair of formulating a satisfactory theory of collective action which reduces to individual agents with ordinary intentions. Some theorists, including John Searle and Raimo Tuomela, offer theories of collective action which still respect methodological individualism with respect to agents, though they account for collective actions by appeal to special intentions which are not reducible to ordinary intentions. I call such theories we-intention theories. I argue that both Searle and Tuomela appeal to special intentions which are not reducible to ordinary intentions to account for collective goals, and they analyze collective action in terms of such collective intentions. These theories face problems, I argue, because they fail to explain the existence of conceptual links between the ordinary individual intentions (I-intentions) and the special collective intentions (we-intentions). Further, such theories fail to account for the fact that there are often no clear phenomenological distinctions between collective actions and non-collective actions, whereas the non-reductive
collective action theorist implies that, insofar as there are different mental states involved in the activities, there is likely to be a phenomenological distinction between the two.

Finally, we turn to what we call plural subject theories of collective action. Plural subject theories, of which we consider those of Margaret Gilbert and Abraham Sesshu Roth, reject the idea that we can reduce collective action to individual agents acting on behalf of individually-held mental states; plural subject theories reject methodological individualism altogether. In its place, plural subject theories argue that we can only account for collective action if we reject the claim, which I call reductionism, that collective actions are performed by agents who act upon individually held mental states (regardless of whether the content of those mental states are individuals or groups). I argue that plural subject theories fail on two counts. First, they fail to offer a plausible rejection of reductionism. Second, they also fail to offer a convincing positive theory of collective action. They argue that commitments are constitutive of collective action, which is implausible; further, insofar as they blur the psychological independence of agents in a collective action, they have problems accounting for the kinds of individual reasoning and individual contributions that agents make on behalf of collective goals.

If the arguments offered in this chapter are correct, then we must accept that none of the theories of collective action currently on offer are acceptable. Further, we will have rejected some of the arguments made by Gilbert and Roth which hold that no reductive theories of collective action could be successful. This will clear the stage for a reductive theory of collective action, which we will build in later chapters; for now, we turn to our analysis of the existing theories, and to the team preference approach.

3.2 The team preference approach

As we saw in Chapter 2, Searle’s challenge requires that any successful theory of collective action be able to distinguish between an outcome which is brought about by a set of individually reasoning agents, and that same outcome as brought about by a set of collectively reasoning agents. We also saw that traditional rational choice theory does not have the resources to account for such a difference. One relatively conservative strategy for addressing Searle’s challenge within a belief/preference rational choice framework is to expand the preferences available to agents so as to include team preferences. This strategy has been advocated most forcefully by Michael Bacharach, by Robert Sugden, and by Natalie Gold
(in joint work with Sugden). There are differences in the details of the theories presented, and these differences are sometimes significant; Bacharach, for instance, allows agents to simultaneously hold multiple distinct team preferences, and uses an exogenous distribution function to answer the question of which team preference function actually motivates an agent to act, whereas Sugden only considers cases in which agents hold one team preference function. There is, however, a general approach to analyzing collective action which we can call the team preference approach, and which can accommodate the theories of Bacharach, and of Sugden, and of Gold.

The team preference approach is distinguished by two key features: the goals of a team are determined by team preference functions, and in acting towards the team goal agents rationally justify their performance of some action over another through team reasoning. If either of these two features is missing, we do not have a team preference approach to analyzing collective action. As we will see, however, there are problems with both features of the team preference approach; although it is a promising way to analyze collective action, it cannot successfully account for collective action.

### 3.2.1 Team preferences

To see why, let us turn first to the team preference function feature of the team preference approach. Team preference functions are comparable to individual preference functions; they are both defined over the set of outcomes in an agent’s subjective decision problem, and they both represent ordinal rankings of those outcomes. Further, both individual and team preference functions are employed by individual agents in order to resolve subjective decision problems. It is not the case, then, that team preference functions are necessarily possessed by collective agents or teams. Ontologically, team preferences functions seem to be on par with individual preference functions; as Sugden writes,

> When I say that a person who engages in team-directed reasoning has a team-directed preference for \( x \) over \( y \), I mean that she takes herself to have reason to make those choices among actions that, according to the theory of team-directed reasoning, are implied by that preference. This analysis is a straightforward generalization of the standard analysis of individual-directed reasoning. Against a charge of circularity or emptiness, the two analyses stand or fall together.\(^2\)

---

2. Sugden 2000, p. 199. We will look at the the difference between team-directed reasoning and individual-directed reasoning below.
The primary difference between individual preference functions and team preference functions is this: whereas individual preference functions rank one outcome \( o_1 \) over another outcome \( o_2 \) so long as the deliberating agent \( X \) judges herself to have greater reason to bring about \( o_1 \) than \( o_2 \), team preference functions rank \( o_1 \) over \( o_2 \) so long as \( X \) judges that some team \( G \) (of which \( X \) takes herself to be a member) has greater reason to bring about \( o_1 \) than \( o_2 \). Individual preference functions track the individual’s beliefs about choice-relevant reasons for individual agents, while team preference functions track the individual’s beliefs about choice-relevant reasons for teams.

The Sugden quote above indicates that Sugden believes this definition of team preferences to be free of circularity worries. One such circularity worry is that in order for \( X \) to have a team preference function, \( X \) must first possess the concept of a team such that she can judge whether the team \( G \) (of which she is a member) has more reason to bring about outcome \( o_1 \) or \( o_2 \). If I am the member of a theater troupe facing a decision about whether to put on a production of Hamlet or Oedipus, then in order for me to have a judgment about whether the troupe has more reason to produce Hamlet or Oedipus, I must first have a clear concept of the theater troupe as a collective. And this, it might be argued, begs the question of whether such a collective exists in the first place.

Sugden is correct to dismiss circularity worries of this kind, however. On the one hand, we needn’t presuppose that the team \( G \) is anything beyond the set of individual agents who comprise \( G \) in order for \( X \) to hold beliefs about what \( G \) has reason to do. I can hold beliefs about what the ten people in a burning building have most reason to do (namely, work together to escape the burning building, assuming that escaping requires some coordination on their parts) without holding any further beliefs about whether the ten people in the room represent a collective in any more socially meaningful sense.\(^3\) On the other hand, even if we do allow that for \( X \) to judge that the team \( G \) has reason to do \( o_1 \) over \( o_2 \) requires that \( X \) possess the concept of ‘team’ (such that she believes \( G \) to be a team, rather than a collection of individuals), we still do not fall prey to circularity worries. The reason is that we are trying to explain the difference between an outcome’s being produced by a team or a collective, and that same outcome’s being produced by a collection of individuals;

\(^3\)In Chapter 2 we looked at theories of collective action that defined collective outcomes as those which provided some benefit to the agents involved. Although we dismissed such theories as providing adequate analyses of collective action, this need not lead us to conclude that it is incorrect to define what groups have reason to do in terms of what benefits the individual agents in the group—although this is not the definition of a group’s reasons for acting that I present in this work.
we are looking for the difference between an individual agent’s action being justified by appeal to individual reasons, and that same action’s being justified by appeal to collective reasons. It is not circular to appeal to the concept of a team in explaining this difference, any more than it is circular to appeal to the concept of an individual in explaining the difference between an individual’s action and mere behavior. We are not looking to provide a conceptual analysis of the concept of a collective—we are looking to analyze collective action.\footnote{But cf. Gilbert in Gilbert 1989.}

Just as individual preference functions indicate the reasons that an individual has to bring about some outcome rather than others, then, so too does a team preference function indicate the reasons that a group has to bring about some outcome rather than others. It is also the case that team preference functions, like individual preference functions, indicate the possessing agent’s judgment about the reasons that the indicated agent or group has to act. That this is so is clear in the individual case; if agent $X$ has an individual preference function which favors bringing about outcome $o_1$ over $o_2$, then $X$ must judge herself to have more reason to produce $o_1$ than $o_2$. Whose judgment else would it be? In the case of team preference functions, however, the case is more complicated. Although team preference functions must correspond to perceived reasons that the group has for acting, the question remains: whose perception matters? As for the options, it might be the individual agent’s judgment about the reasons the group has for acting: if $X$’s team preference function (for group $G$) prefers $o_1$ to $o_2$, then $X$ judges that $G$ has more reason to bring about $o_1$ than $o_2$. On the other hand, it might be the group’s judgment which matters: for any agent to have a team preference function (for group $G$) preferring $o_1$ to $o_2$, it must be the case that $G$ judges that it has more reason to bring about $o_1$ than $o_2$.

The latter interpretation is problematic. For us to take such an interpretation, we need an account of what it means for a group to hold a judgment which does not itself presuppose a theory of collective action. Although there are some theories on offer, as for example by Philip Pettit or by Margaret Gilbert, the adequacy of these theories is debatable.\footnote{See Pettit 2002, Gilbert 1989, and Gilbert 1996a. See also Schmitt 2003.} We need not debate these theories here, however, because there is another problem with the latter interpretation of team preference functions: it is implausible as an interpretation of what a preference function is. The latter interpretation entails that an agent cannot reason according to a team preference function unless the relevant group actually holds the
judgments corresponding with the preference function. It is hard to see, however, what could prevent an agent from reasoning according to a preference function which is not actually held by the group. Whether some group $G$ judges that it has reason to bring about outcome $o_1$ over $o_2$ is irrelevant; if $X$ (as a member of $G$) is to have her deliberation influenced by the team preference function, then $X$ must also hold such a judgment. Now, there might be some principle which ensures that whenever $G$ judges itself to have reason to bring about $o_1$ over $o_2$, then every agent $A$ in $G$ will also judge that she or he has reason to bring about $o_1$ over $o_2$. We cannot take the existence of such a principle for granted, however; whether there is one depends largely on how we fill out the details of what it means for $G$ to hold a judgment. In any event, the explanatory primacy is on whether the individual agent judges that the group has reason to act, and so we can focus our attention there, rather than on the judgments of the group.

To repeat, then: for an agent $X$ to hold a team preference function (for group $G$) favoring outcome $o_1$ over $o_2$ is for $X$ to judge that $G$ has more reason to bring about $o_1$ than $o_2$. Team preference functions are thus comparable to individual preference functions. Where individual preference functions track what an agent judges an individual to have reason to do, team preference functions track what an agent judges a group to have reason to do.

### 3.2.2 Team reasoning

The team preference approach is not complete with just a theory of team preferences, however; team preferences only reveal an agent’s judgment about what a group has reason to do. Team preferences do not yet provide an individual agent with any indication of what she or he should do, even if the agent identifies with the group. If I identify with a group of friends and believe that we have reason to all meet up at a concert later on that evening, it is not yet clear what that belief about the group’s reasons for acting entails for my own individual reasons for acting. Do I have reason to purchase a ticket to the concert? Do I have reason to attend? Do my own individual reasons for acting depend on the expected behaviors of the other agents in the group I identify? What is needed, then, is a bridge between the group’s reasons and the individual’s reasons.

This bridge problem does not exist for agents reasoning with individual preference functions. If agent $X$ prefers outcome $o_1$ to $o_2$, then $X$ judges herself to have more reason to bring about $o_1$ than $o_2$; and if we assume that $X$’s decision problem consists solely of those
two outcomes, then \( X \) has reason to perform those actions which are expected to produce \( o_1 \). Although there might be some uncertainty as to what that action is—I prefer the outcome of my being happy to the outcome of my being sad, although I do not necessarily know which actions of mine are more likely to promote my happiness than my sadness—there is a direct connection between the outcome that \( X \)’s individual preference function ranks most highly and the outcome that \( X \) has reason to bring about. The reasoning process which leads from individually preferred outcome to rational action is simple practical reasoning.

The team preference approach solves this bridge problem in the case of team preference functions by advocating team reasoning. Just as team preferences are held by (ontological) individuals and not teams, so too is team reasoning performed by individuals and not teams.\(^6\) When an individual \( X \) engages in team reasoning as part of the group \( G \), \( X \) first determines that \( G \) has most reason to bring about some optimal outcome \( o^* \) (through the use of a team preference function). \( X \) next judges that she has reason to do her part of that set of actions which will produce \( o^* \). Gold and Sugden express the process of team reasoning by way of a formal schema:

\[
\text{Schema 4: simple team reasoning (from an individual viewpoint)}
\]

1. I am a member of \( S \).
2. It is common knowledge in \( S \) that each member of \( S \) identifies with \( S \).
3. It is common knowledge in \( S \) that each member of \( S \) wants the value of \( U \) to be maximised.
4. It is common knowledge in \( S \) that \( A \) uniquely maximises \( U \).

I should choose my component of \( A \).\(^7\)

Compare this formal reasoning schema with another formal reasoning schema that Gold and Sugden present:

\[
\text{Schema 2: collective rationality}
\]

1. We must choose one of \( (\text{left, left}), (\text{left, right}), (\text{right, left}) \) or \( (\text{right, right}) \).
2. If we choose \( (\text{left, left}) \) the outcome will be \( O_1 \).

\(^6\)Some theorists draw a distinction between the kind of reasoning that is carried out by individuals, and the kind of reasoning which is carried out by teams. Sugden, for instance, contrasts team-directed reasoning with team reasoning in Sugden 2000; whereas team-directed reasoning is performed by individual agents, team reasoning is performing by groups (i.e. by sets of agents). Unless otherwise noted, we will focus here exclusively on reasoning performed by individuals, and not on reasoning performed by sets of individuals.

\(^7\)Gold and Sugden 2007, p. 126.
(3) If we choose \((\text{left}, \text{right})\) the outcome will be \(O_2\).

(4) If we choose \((\text{right}, \text{left})\) the outcome will be \(O_3\).

(5) If we choose \((\text{right}, \text{right})\) the outcome will be \(O_4\).

(6) We want to achieve \(O_1\) more than we want to achieve \(O_2, O_3, \text{ or } O_4\).

We should choose \((\text{left}, \text{left})\).\(^8\)

Where Schema 2 talks about the reasons that the \textit{group} has to act, Schema 4 talks about the reasons that each individual in the group has to act. The gap between Schema 2 and Schema 4 is filled in by team reasoning.

We can also see from Schema 2 suggestions of how Gold and Sugden intend that we understand the requirement that agents choose ‘their component’ of the collective action profile \(A\): when \(A\) is specified so as to make precise the contributions of each agent, then each agent should choose to do their contribution. In Schema 2, the group’s preferred action profile was \((\text{left}, \text{left})\); this means that the first agent’s contribution is \text{left} and the second agent’s contribution is \text{left}. Gold and Sugden do not go into details about how agents might determine their component of \(A\) in cases where \(A\) is not fully specified, although we can assume the existence of a mechanism through which the agents can determine their contribution to a collectively preferred outcome.\(^9\)

Note that under the ‘simple team reasoning’ schema captured in Schema 4, for agents in a group to engage in team reasoning, they must hold certain beliefs. First, it must be common knowledge that the agents in the group \textit{identify with} the group. This is a strong requirement. Gold and Sugden write,

We will say that \(i\) identifies with \(G\) if \(i\) conceives of \(G\) as a unit of agency, acting as a single entity in pursuit of some single objective.\(^{10}\)

We consider whether there is a non-circular way of understanding this requirement below. In addition to (true) beliefs required about group identification, there are other requirements that must be fulfilled in order for an agent to engage in simple team reasoning: there must be a common team preference function \(U\) that the agents reason according to, and there

\(^8\)Gold and Sugden 2007, p. 122.

\(^9\)As we will see below, this same issue will pose a much less benign problem for other theories of collective action.

\(^{10}\)Gold and Sugden 2007, p. 125, emphasis original.
3.2. THE TEAM PREFERENCE APPROACH

must also be a unique action profile $A$ that maximizes the value of $U$.\textsuperscript{11} We can interpret the last requirement as claiming that there must be one unique outcome $o^*$ which $U$ prefers over all others, and there must be one unique set of actions which is expected to produce $o^*$.

3.2.3 Problems

The team preference approach to collective action is not without its problems; we will discuss one major one here. Insofar as the team preference approach is a conservative extension of traditional rational choice theories, it inherits many of the weaknesses of traditional rational choice theories in modeling agency—in particular, the limitations that arise from its being restricted to belief-type attitudes and desire-type attitudes. I will say more in Chapter 4 about our need to broaden our rational choice theory beyond just beliefs and desires; here, we can restrict ourselves to noting that the team preference approach requires that agents share identical team preference functions, and that this requirement is too severe. Furthermore, we cannot weaken this sharing requirement so as to make it more palatable; the problem persists so long as we attempt to use beliefs and desires to account for an agent’s goals (whether those goals are individual goals or collective goals).

Shared team preferences

On the team preference account detailed above, in order for a group of agents to engage in collective action, they must all employ the exact same team preference function in their team reasoning: for any two agents in the group $X_1$ and $X_2$, if $X_1$’s team preference function $u_1$ favors some outcome $o_1$ over another outcome $o_2$, then $X_2$’s team preference function $u_2$ must also favor $o_1$ over $o_2$. Thus we have condition (3) of the simple team reasoning schema 4 presented above:

It is common knowledge in $S$ that each member of $S$ wants the value of $U$ to be maximised.

\textsuperscript{11}Gold and Sugden do not explicitly state that $U$ must be a team preference function; they claim that $U$ can be ‘any payoff function’. If, however, we think that team preference functions differ from individual preference functions in representing choice-relevant reasons for a group to act rather than an individual—and Sugden has claimed this—then prototypical cases of collective action would seem to require that $U$ be a team preference function. Whether $U$ could be an individual preference function and still be conducive to collective action will be a question that we can ignore here.
Comparable conditions are repeated throughout the team preference literature; thus Sugden talks of a common utility function $v$ that agents in a team $N$ seek to optimize in Sugden 2003, and Bacharach identifies each group that an agent might belong to with a unique utility function $U$ in Bacharach 1999.

We can call the claim that each agent has a completely identical team preference function to be strong sharing. The team preference approach, then, entails that a necessary condition for a group of agents to act collectively is that they strongly share a team preference function. This strong sharing condition is too strong, however. Remember that team preference functions correspond to an individual agent’s judgments about what the team has reason to do; if $X$’s team preference function $\succ_X$ (for group $G$) prefers outcome $o_1$ over $o_2$, then $X$ judges that $G$ has more reason to bring about $o_1$ than $o_2$. The strong sharing condition, then, entails that a necessary condition for collective action is that every agent in a group must share identical judgments about what the team has reason to do.

One easy way to see that this is too strong is by considering the various outcomes that a group might deliberate between, but which are not favored by any of the agents in the group (though they disagree about just how disfavored they are). Suppose two agents, $X$ and $Y$, are deliberating about going as a group to see one of four movies: Awesome movie, Boring movie, Captivating movie, and Disappointing movie. $X$’s team preference function $\succ_X$ and $Y$’s team preference function $\succ_Y$ are in partial agreement; they both rank $A$ as the most preferred movie, and $C$ as the second most preferred movie. $X$ and $Y$ disagree about whether they have more reason to see $B$ or $D$, however; $X$ thinks that they have more reason to see $B$ than $D$, while $Y$ thinks they have more reason to see $D$ than $B$. Thus,

$A \succ_X C \succ_X B \succ_X D$

while

$A \succ_Y C \succ_Y D \succ_Y B$.

(Perhaps $X$ thinks strongly that the group should reward boringness over disappointment, while $Y$ thinks that they would have more to talk about after the disappointing movie.)

Assuming these team preference functions to be common knowledge, it seems odd to suggest that $X$ and $Y$ cannot engage in a collective action in virtue of these team preference functions. After all, $X$ and $Y$ can certainly coordinate their actions so as to bring about the group’s most highly preferred outcome. (We can suppose that there are no worries about
tickets being sold out, et cetera). There is even agreement in this case about the second most highly-preferred outcome for the group, in case complications prevent them from jointly acting to bring about the most highly preferred outcome. Nevertheless, the agents do not strongly share a team preference function; ≿\_X is not identical to ≿\_Y. According to the team preference approach, then, collective action is not possible given these team preference functions.

The defender of the team preference approach might argue that the strong sharing requirement is too strong—that what is needed is not that agents hold identical team preference functions, but rather that their team preference functions agree on the most preferred outcome for the team (even if they disagree about how disfavored other potential outcomes are). Let us say that two agents X and Y weakly share a team preference function when their team preference functions ≿\_X and ≿\_Y agree on the most preferred outcome for the group in some situation. That is, o\(^*\) ≿\_X o for every outcome o in the decision problem if and only if o\(^*\) ≿\_Y o for all o.

If we return to our movie example, X and Y weakly share a team preference function with respect to their decision about movies; both X and Y judge that the team has most reason to go see Awesome movie. If we were to require as a necessary condition for collective action only that agents weakly share a team preference function rather than strongly share it, we would avoid the counter-intuitive conclusion that X and Y are incapable of engaging in collective action on the basis of their respective preference functions.

Although weak sharing does avoid the problem with the movie-goers, requiring that agents weakly share team preference functions is still too strong. If agents must weakly share a team preference function to engage in collective action, then it would not be possible for a group of agents to engage in a collective action even though one of them believes that the group (as a whole) has more reason to do something else—it would require, that is, that the agents’ judgment about what the group has most reason to do be unified. If we accept weak sharing as a necessary condition for collective action, then we must disregard the possibility of a group of agents engaging in collective action to bring about some outcome o\(_1\) in spite of the fact that they are not in unanimous agreement. Some agent in the group thinks the group has more reason to bring about outcome o\(_2\).

And yet, this situation seems to happen with regularity. The football team might disagree as to whether the team ought to execute pass play p or run play r; some members of the team believe that the team has most reason to do p, while others believe the team
has most reason to do $r$. Their disagreement about the reasons the group has to act does not entail that they are incapable of collectively executing one play or another, however. Similar situations abound: the music group might collectively interpret their chamber music one way, in spite of the fact that several members believe the group has more reason to interpret it another way; the North American corporation might expand their sales into Europe while a vocal but cooperative subset believes they have more reason to expand into Asia; and so forth. Different judgments about the reasons a group has to bring about one outcome rather than another is not incompatible with collective action.

It is easy to see why the team preference approach would require that agreement about reasons for a group be necessary for collective action. In the individual case, as we saw in Chapter 2, action is behavior for reasons. When an individual agent acts, the agent acts upon her or his reasons for acting. It is reasonable to think that this situation also holds true when there is a set of agents acting collectively—that when a set of agents acts collectively, the group acts upon reasons that it has for acting. The intuition, then, is that groups act upon reasons analogously to the way that individuals act upon reasons. And, it is plausible to think, if the group must have a reason to act in order to act collectively, then all of the agents in the group must share an understanding of what the group’s reasons are for acting.

The intuition is mostly correct; insofar as a set of agents is acting (as a group), they are acting upon reasons that the group has. We can also allow that, from a perspective internal to the group’s judgment, when a group acts, it acts upon the strongest reasons that it has for acting at the time. In other words, reasons are important, as is the idea that the group aims at maximizing its reasons for acting. Where the team preference approach errs is in blurring the distinction between the group’s reasons for acting, and the judgments each individual has about the group’s reason for acting. We have not yet said what it means for a group to have reason for doing one action over another from an internal perspective. It is clear from the preceding discussion, however, that a group $G$’s having internal reason to bring about outcome $o^*$ is not reducible to each agent in $G$’s judging that $G$ has reason to bring about $o^*$. Groups, like individuals, face pragmatic pressure to resolve decision problems. But whereas individuals are necessarily unanimous in resolving decision problems—when an individual resolves a decision problem, there are no conflicting judgments by others in the agential unit about what the agent has reason to do—groups are not necessarily so. A group might resolve a decision problem without there being unanimity about judgments about what the
3.2. THE TEAM PREFERENCE APPROACH

A group has reason to do. The resolution of a decision problem is a decision about actions, however, and not a judgment about reasons; groups successfully resolve decision problems when they choose to act. If a group can successfully choose to act without coming to agreement about reasons the group has to act, the group will have successfully resolved the decision problem.

Now, groups can act without coming to agreement about its reasons for acting. So long as a group has a salient goal—a goal that each member of the group is committed to realizing, such that this is all common knowledge—then the group can successfully act even though they remain in dispute about their reasons for action. The group’s goal would be analogous to an individual goal; just as an individual goal is an outcome that an individual is (volitionally) committed to bringing about, a group goal is a goal that the group is committed to bringing about.

Further, this goal commitment distributes over the individuals in a group in a way that judgments about reasons for acting do not. A group G’s judgment that it has most reason to bring about outcome \( o^* \) does not distribute; we cannot conclude that each member of \( G \) also judges that it has most reason to bring about \( o^* \). Goal commitment does distribute, however; if \( G \) is committed to bringing about \( o^* \), then each member of \( G \) must be committed to bringing about \( o^* \). This gives us reason to think that it is goal commitment, and not judgment about reasons for action, which is a necessary condition for collective action; a fuller discussion of goal commitments will be deferred until Chapter 4.

The preceding argument has shown, then, that we need to be able to divorce an account of the possession of goals from an account of the possession of reasons for action (or of judgments about reasons for action). An individual agent might possess a goal which does not accord with her or his perceived reasons for action. (Such an agent might well be irrational; we will consider this point further in Chapter 4.) Likewise, a group might possess a goal which does not accord with the constituent members’ perceived reasons for action. Team preference theorists tie both their account of goal possession and their account of perceived reasons to value functions. This fact prevents the team preference theorist from giving a satisfactory account of a set of agents who each possess a common team goal and yet each hold different judgments about what the group has reason to do.

If we are to have a satisfactory account of collective action, then, we must employ a theory of action which can distinguish between the possession of a goal and possession of reasons for action. In the remaining sections of this chapter, we will turn our attention to
theories that account for goal-possession through psychological attitudes which are distinct from, and irreducible to, beliefs and preferences.

3.3 We-action theories

Intention-based theories of collective action represent an attempt to answer Searle’s challenge while at the same time avoiding the flaws of the team reasoning approach. While the team reasoning approach models an agent’s goals by appealing to her or his preferences (whether individual or team) and her or his mode of reasoning, an intention-based theory of collective action can account for an agent’s goals by appealing to the content of an agent’s intentions.

There are many ways of incorporating intentions into a theory of collective action. We will look here at the theory proposed by Michael Bratman; his theory of collective action is distinctive in that he appeals only to individual agents and to the contents of their intentions, and he further claims that the contents of their intentions do not appeal to non-reductive collective action concepts. Although Bratman’s theory is not fully reductive, in that he does not explain collective action by way of the same kinds of attitudes and contents that explain individual action, his theory is nevertheless very reductive; he appeals to individual agents and to the same mental states as are employed in explanations of individual actions. There are other theories of collective action which also appeal to intentions; insofar as these theories do not appeal to the same mental states as in individual actions (as, for instance, the theories put forth by Searle or by Tuomela), we will consider them below.

We-action theories are not fully reductive, but this in itself is not a problem; as we saw in Chapter 2, we should be skeptical that any fully reductive theory of collective action can avoid Searle’s challenge. The problem facing Bratman’s theory, then, is not that it is not reductive enough; rather, the problem is that Bratman does not provide a satisfactory account of how the content of the individually-held intentions can invoke the collective action while not violating the condition that agents can only intend their own actions.

3.3.1 Shared intentions and team goals

In his analysis of collective action, Bratman talks primarily of shared intentions. We saw in the previous section that the team preference approach fails to have an acceptable analysis of team goals. It is worth wondering whether Bratman’s concept of shared intentions is
3.3. **WE-ACTION THEORIES**

sufficient to account for team goals—that is, if a group of agents shares an intention to act, do they necessarily possess a team goal? And, in the other direction, if a group of agents possesses a team goal, do the agents in the group share an intention?

Skepticism about the equivalence between the two concepts stems from the suspicion that the concept of ‘sharing’ entails a practical and interpersonal intimacy that is not necessarily present in all cases of collective action. We can contrast the phrase ‘X and Y went on a walk together’ with the phrase ‘X and Y shared a walk’; although both activities entail the agents walking together, we might suspect that the latter entails a personal connection between the agents that the former does not. If sharing an experience (such as taking a walk) is indeed something stronger than collective participation in an experience, then it could be plausible that sharing a goal (or an intention) is something stronger than merely collectively holding a goal or intention. It is very difficult, however, to say what this ‘something stronger’ is.

Perhaps Bratman’s use of the phrase ‘shared intention’ is intended to invoke the stronger concept of sharing rather than the weaker concept of sharing? This is doubtful, I think, simply because his analysis makes no effort to account for that nebulous ‘something stronger’. For one, Bratman takes ‘we share an intention to X’ to be synonymous with ‘We intend to X’. For another, Bratman’s analysis simply doesn’t make reference to concepts that would suffice to account for the difference between the strong notion of sharing a goal, which entails a particular kind of interpersonal intimacy, and the weaker notion of sharing a goal, which does not. Bratman identifies three features characteristic of shared cooperative activity: mutual responsiveness, commitment to the joint activity, and commitment to mutual support. None of these features seems designed to pick out the difference between the two kinds of sharing we have mentioned, because they do not speak to the interpersonal intimacy requirement; rather, each feature is goal-oriented, and speaks to the actions that agents are committed to performing in pursuit of the goal.

It is unlikely, then, that Bratman’s use of the term ‘shared intention’ is meant to invoke the strong kind of sharing which entails interpersonal intimacy. There is, however, another way in which shared intentions might conceivably be different from team goals. We can see this clearly if we look at the distinction that Bratman draws between jointly intentional

---

12See e.g. Bratman 1993, p. 106.
13Although Bratman identifies these features as characteristic of shared cooperative activity and not shared intention, the account that he gives of the two phenomena are very similar, and we apply this analysis of shared cooperative activities to shared intentions.
activities (JIAs) and shared cooperative activities (SCAs). Bratman describes two singers who, though each intending to do her or his part in bringing about an outcome in which they both successfully sing a duet together, nevertheless each fail to intend to support the other singer should she or he need help in singing her or his part. Each singer, that is, intends to sing her or his part as well as possible, and intends to sing in a manner which is appropriately responsive to the singing of the other. Should one falter, however, neither singer intends to help the other. In such a case, Bratman claims, the agents are engaged in a JIA but are not engaged in a SCA, since SCAs require a commitment to mutual support. To quote Bratman,

\[\text{[A] joint-act type [can] be loaded with respect to joint intentionality but still not [be], strictly speaking, cooperatively loaded . . .}^{14}\]

If this is right, then we might wonder whether team goals are ‘cooperatively loaded’, or whether they are merely loaded with respect to joint intentionality. Bratman’s analysis of shared intentions is of a cooperatively loaded phenomenon; if team goals are not cooperatively loaded, then shared intentions and team goals do not purport to describe the same thing.

I think that team goals are cooperatively loaded in Bratman’s sense. Team preference theorists do not generally talk about a disposition that agents in a collective action have to mutual support, perhaps in part because it is difficult to analyze such dispositions any way other than counterfactually, and rational choice models are not ideally suited to analyzing counterfactuals. Nevertheless, agents who hold a team goal must be committed to more than agents who are engaged in a JIA are committed to. Consider three kinds of outcomes that can be produced: collectively successful outcomes, in which each agent in a joint activity successfully perform their roles and the joint outcome is brought about; individually successful outcomes, in which one individual agent successfully performs his or her action but the joint outcome is not brought about due to failure on the part of the other agent; and jointly unsuccessful outcomes, in which neither agent performs her or his action, and the joint outcome is not brought about.\(^{15}\) Whether an agent is engaged in a JIA or an SCA, the jointly unsuccessful outcome will surely be the lowest valued outcome according


\(^{15}\)The word ‘joint’ here is being used neutrally; we can assume only that the production of a joint outcome require the contribution of more than one agent. We cannot simply assume that joint outcomes are collective outcomes (however we understand the latter phenomenon).
to the agents’ value functions.\textsuperscript{16} According to Bratman’s definition of JIAs, the agent who is engaged in a JIA but not a SCA should value the individually successful outcome over the collectively successful outcome. As Bratman states about his mutually unsupportive singers, ‘were I to stumble you would not help; for you prefer my failure to our success.’\textsuperscript{17}

This is not yet conclusive. What Bratman’s unsupportive singer prefers is not necessarily the outcome in which she performs her action and the other singer fails to perform his action over the outcome in which they both perform their actions. Rather, what the unsupportive singer prefers is the outcome in which she performs her action and the other agent fails to perform his action over the outcome in which she \textit{has to perform additional actions to help him successfully perform his}. This distinction, however, does not affect the main claim I want to make here: namely, that the team goal theorist can distinguish between the agent engaged only in a JIA and the agent engaged in a SCA, according to the pattern that their value functions take over various outcomes. Further, although they rarely make it explicit, team preference theorists clearly assume that the possession of a team goal requires that agents have a preference function which is consistent with their being in a SCA and not a JIA.

We can thus conclude that the shared intentions of Michael Bratman purport to analyze the same phenomenon as do team goals for the team preference theorist. We saw that the team preference theorist has an impoverished definition of team goals on hand. We should thus consider whether or not Bratman’s account of shared intentions is adequate to account for collective action where the team preference account fails.

### 3.3.2 Bratman’s intention-based theory of collective action

At the heart of Bratman’s theory of collective action is his account of shared intentions. He claims that two agents are engaged in a collective action of \textit{J}-ing so long as they share an intention to \textit{J} and their shared intention to \textit{J} is \textit{uncoerced} and \textit{minimally cooperatively stable}. Bratman’s account of collective action is a reductive one; in order to avoid begging the question in his analysis, he characterizes the act of \textit{J}-ing as a \textit{cooperatively neutral joint-act type}; he thus attempts to avoid building into the content of each agent’s intention that the action performed be a collective action. A discussion of whether this is successful

\textsuperscript{16}Our use of the term ‘value function’ here is agnostic about whether we need to posit the existence of team preference functions.

\textsuperscript{17}Bratman 1992, p. 337.
will be left until the following section; here, it will be useful to see how Bratman’s account is supposed to work.

Bratman gives his account of shared intention as follows (remembering that, for Bratman, ‘we intend to J’ is synonymous in contexts of practical reasoning with ‘we share an intention to J’):

We intend to J if and only if:
1. (a) I intend that we J and (b) you intend that we J;
2. I intend that we J in accordance with and because of 1a, 1b, and meshing subplans of 1a and 1b; you intend that we J in accordance with and because of 1a, 1b, and meshing subplans of 1a and 1b;
3. 1 and 2 are common knowledge between us.\textsuperscript{18}

We can call the first condition the \textit{we-intention} condition, the second the \textit{meshing subplans} condition, and the third the \textit{common knowledge} condition.\textsuperscript{19}

As Bratman himself notes, shared intentions are not mental attitudes that are held in the head of any individual; shared intentions are rather a state of affairs that consists of the appropriate mental attitudes being held by the appropriate individuals. Moreover, when a shared intention exists, the appropriate mental attitudes are held by individuals, and the individual mental attitudes do not inappropriately reference the collective activity. Or, so at least Bratman claims; as I will now argue, this claim is an overly optimistic one.

\subsection*{3.3.3 Shared intentions and circularity}

Bratman’s we-intention condition requires that each agent X and Y hold intentions of the form ‘I intend that we J’, where J is a joint act-type. This raises the immediate worry: how can an \textit{individual} agent intend a \textit{joint} act-type?

Bratman identifies several potential worries with an agent’s intending a joint act-type. After rejecting what he claims to be an inadequate own-action condition on intentions

\textsuperscript{18}Bratman 1993, p. 106.

\textsuperscript{19}What we are calling the ‘meshing subplans’ condition is actually broken down further by Bratman into two conditions, which he calls the ‘meshing’ condition and the ‘interlocking intentions’ condition. For Bratman, intentions mesh so long as the intentions are compossible—there must be some way of satisfying all meshing intentions. Intentions are interlocking, on the other hand, so long as the intending agents have supplementary intentions in favor of the efficacy of the other agents’ intentions; agents must intend that the other agents’ intentions be successful. We will, for simplicity, use the single term ‘meshing subplans condition’ to account for both.
(namely, that agents can only intend her or his own actions), Bratman considers two conditions on intentions which seem more plausible: a control condition C, which claims that an agent cannot intend an action that she or he does not control, and a settle condition S, which claims that an agent can only intend that which settles issues that are up to the agent. Since these are plausible constraints on intentions, Bratman argues, the we-intention condition might well violate the C and S constraints.

However, Bratman dismisses worries that the we-intention condition violates either the C or the S constraints on intentions. First, Bratman argues that plausible C and S constraints on intentions should permit what he calls other-agent conditional mediation. An agent X can intend to bring about some outcome o by means of action x even though the production of o requires that some other agent Y perform some action y: so long as X believes that whether o is produced depends on her doing x, she both controls whether o is produced, and she also settles the matter of whether o is to be produced. This is true, Bratman argues, even if Y’s performing y is conditional upon X’s performing x—so long as X knows that Y’s doing y is conditional on X’s intending to perform x, X can control and settle the production of o by forming the intention to x.

It is therefore not illicit, Bratman argues, for an individual agent X to intend a joint act-type J, so long as she plausibly believes that whether the joint action J is done depends (in part) on whether she forms the intention to do her part x of the joint action J. As Bratman writes,

> When I decide that we paint together, I suppose that my intention that we paint will lead you so to intend as well. Does this mean that, strictly speaking, you don’t get to settle the matter of our painting or, at least, I don’t see you as settling the matter? Well, you remain a free agent; it really is a decision that is up to you and without which we really will not paint. I predict that, in part as a result of my intention, you will so decide; but that does not mean that you do not decide. I can predict what I know to be your free decision. I can predict that you will freely, in response to my intention, intend that we paint, and so settle the matter of our painting together. That is why I can now intend that we paint.\(^\text{20}\)

We should note that the content of the agent’s intention here is the joint act-type ‘I intend that we paint’, rather than the individual act-type ‘I intend that I do my part of our painting’. In presenting his claim this way, Bratman is skipping a step. He argues that

\(^{20}\text{Bratman 1999b, pp. 156–157, emphasis original.}\)
viable intentions can allow for other-agent conditional mediation, and this claim seems right. Bratman makes this claim, however, in order to justify the viability of ‘I intend that we J’ as a legitimate intention that an agent might have. Whether other-agent conditional mediation is possible is a different question than whether agents can hold individual intentions with the content ‘I intend that we J’, however, and in fact other-agent conditional mediation cannot get us intentions with that kind of unique content.

Let us say that outcome o is produced by agent X’s doing action x and agent Y’s doing action y. Let us suppose further that agent Y will only do y if he sees X about to perform x, or if he sees X form the intention to do x. As Bratman has noted, it is proper for X to intend to bring about o by performing action x; X can reasonably be said to both control whether o is produced, and to settle the matter of whether o is to be done. However, it is not yet clear whether we can say that X intends that they (X and Y) produce o. X might intend that she produces o (by way of doing X), but this does not yet license the further claim that X intends that they produce o.

So, can Bratman make the jump from other-agent, conditionally mediated intentions to we-intentions? I do not think he can; we-intentions are different than other-agent conditionally mediated intentions in that we-intentions have as their content not just the agent’s action, but the action of other agents as well. Admittedly, other-agent conditionally mediated intentions include in their content outcomes which require contributory actions from other agents; this is very different, however, from having as their content that other agents perform those contributory actions. To see why, we need to more carefully consider the meaning of intentions with the content ‘I intend that we J’.

It is important for the account of collective action presented in the previous section that the content of each individual agent’s intention be a joint act-type J which is cooperatively neutral. A joint act-type is one which is strictly composed of act-types by two contributing agents. To see this, let agent X have some set of behaviors x which she can perform and which constitute an act-type, and let agent Y have a set of behaviors y which he can perform and which also constitute an act-type. Then there is a joint act-type J which consists of the aggregate set x ∪ y. We can make a further claim, which we can call the compositionality of joint act-types: for any joint act-type J which can be performed by agents X and Y, there must be some set of behaviors x and some set of behaviors y such that J is the aggregate set x ∪ y. If all joint act-types satisfy compositionality, then all joint act-types are reducible to some set of behaviors performed by each of the participating agents.
It is debatable whether or not every joint act-type satisfies the compositionality condition. Every cooperatively neutral joint act-type must satisfy compositionality, however. Suppose we have some cooperatively neutral joint act-type \( J_0 \), performed by agents \( X \) and \( Y \). Since \( J_0 \) is cooperatively neutral, there must be some joint act-type \( J_+ \) which is behaviorally indistinguishable from \( J_0 \), but which is a cooperative activity, and there must be some joint act-type \( J_- \), which is behaviorally indistinguishable from \( J_0 \), but which is not a cooperative activity.\(^{22}\) Then \( J_+ \) and \( J_- \) must be behaviorally indistinguishable as well. Since \( J_- \) is a joint act-type which is non-cooperative, it cannot be explainable solely by reference to collective goals (since, as a non-cooperative activity, there is no collective goal to refer to). Instead, \( J_- \) must be explainable according to the individual goals of \( X \) and \( Y \). This, in turn, means that the there must be component activities \( x \) and \( y \) which track the individual goals of \( X \) and \( Y \), and which jointly constitute \( J_- \). \( J_- \), then, is equal to the union of the individual act-types \( x \cup y \). Since \( J_0 \) is a cooperatively neutral joint act-type, it must be behaviorally identical to \( J_- \); this means that it must also be composed of the act-types \( x \) and \( y \). \( J_+ \) must be composed of actions \( x \) and \( y \) as well, although we are not committed to the further claim that \( x \) and \( y \) must be produced by the same individual goals that rationalized the production of \( x \) and \( y \) for \( J_- \).

Every cooperatively neutral joint act-type, then, must satisfy compositionality: the joint act-type \( we \) \( J \) is identical to the joint act-type \( I \) \( do \) \( x \) \( and \) \( you \) \( do \) \( y \). That an action is a joint act-type is not by itself sufficient to make that action a collective action, of course; we must give some further analysis of what makes the production of some cooperatively neutral joint act-type \( J \) a collective action. This is what Bratman aims to do with his meshing subplans and common knowledge condition.

The we-intentions condition is separable from the meshing subplans condition and the common knowledge condition, however, and if the we-intentions condition is not defensible, then Bratman’s account of collective action fails. Because all cooperatively neutral joint act-types must satisfy the compositionality requirement, however, the we-intention condition ends up being circular.

\(^{21}\)Facundo Alonso makes the argument in unpublished work that some collective actions do not satisfy compositionality.

\(^{22}\)There might be some cooperatively neutral joint act-types which, though described in cooperatively neutral terms, nevertheless do not admit of a cooperatively positive description; this might include joint activities such as quarreling, or mutual coercive activities. I make no claims about whether such activities do in fact admit of a cooperatively positive description; if such joint activities do exist, they do not affect the current argument, as we are only interested in that class of joint activities which might be cooperatively positive.
In the original statement of the we-intention condition, 1 (a) X must intend that they J and 1 (b) Y must intend that they J, where J was a cooperatively neutral joint act-type. Because of compositionality, J must be identical to some union of act types x and y. We can then rephrase the we-intention condition: 1′ (a) X must intend that X do x and that Y do y, and 1′ (b) Y must intend that X do x and that Y do y.

This opens Bratman’s we-intention condition up to a worry: how can X intend that Y do some action? If we accept that the C and S conditions represent legitimate constraints on individual intentions, then we should reject the idea that X can intend that Y do y, since X should not see herself as either controlling or settling whether Y does y.

It might be argued that X does control or settle, in some limited sense, whether Y does y; after all, X can predict that Y will do y so long as he comes to believe that X will do x. Bratman speaks more explicitly about the S condition; he writes,

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

Given this confidence, can Abe decide to pump water? Can he, in the relevant sense, “settle” the matter of whether the water is pumped? I think he can, given that he is in a position to predict that Diane will respond appropriately.23

This does not get us what we need, however. Note that Bratman does not explicitly claim that Abe settles the matter of whether Diane turns the pressure valve; rather, what Abe settles is whether the water is pumped. But Bratman is in danger of equivocating on this point. Abe might intend merely to bring about the outcome of the water’s being pumped. In this case, however, it is wrong to describe the situation as Abe intending that they pump the water (by way of his pumping and her turning the valve). Though Abe might be able to settle the matter of having the water pumped (by way of his pumping), Abe cannot settle the matter of their pumping the water (by way of his pumping and her turning the valve), since Abe cannot settle the matter of Diane’s turning the valve.

The distinction here is relevant. If we describe Abe’s intention as having the water pumped (by way of his pumping the water), then Abe must plausibly believe that there is some causal connection between his pumping the water and the water’s being pumped. That this causal connection goes through Diane’s intentions is irrelevant; although Abe

---

knows that Diane’s contribution is necessary, Diane’s contribution is part of the ordinary causal background which connects Abe’s action to the desired outcome. If, on the other hand, we describe Abe’s intention as their pumping the water (by way of his pumping and her turning the valve), then Abe must see himself as able to settle both his pumping and her turning the valve. This means that Abe must see himself as able to settle Diane’s contribution to the action, which is not possible if Abe also sees Diane as an independent agent.

Though an agent $X$ might be able to intend to bring about an outcome $o$ that depends on her doing action $x$ and another agent $Y$’s doing some action $y$, we have seen that it is not correct to say that $X$ intends that they bring about $o$ by way of her doing $x$ and his doing $y$. Rather, it is more correct to say that $X$ intends to bring about $o$ by way of her doing $x$ (with the knowledge that her doing $x$ will lead, through $Y$’s doing $y$, to $o$). What we cannot say is that $X$ intends that they bring about $o$, since that will require that $X$ intend that $X$ do $x$ and that $Y$ do $y$—and yet we know that $X$ cannot either control or settle $Y$’s doing $y$. Bratman’s reliance on we-intentions is thus problematic; he cannot account for collective action without each agent’s possessing we-intentions, and yet his we-intentions violate reasonable conditions on individual intentions.

To sum up, then: Bratman argues that we can provide an account of collective action which reduces to individual agents each deliberating and acting upon individually held but interdependent mental states, and that these mental states are no different in kind from the mental states employed by agents in individual action. The difference, Bratman argues, between instances of individual action and instances of collective action is that, in the case of collective action, agents are acting upon intentions with a unique content: agents intend that the members of the group perform some action, rather than that the agent herself or himself perform some action. Bratman suggests that intentions which make such non-reductive reference to the collective action do not violate circularity because the intentions are other-agent mediated: $X$’s intention that $X$ and $Y$ do $J$ is something that she can control because she is confident that, should she form the intention, $Y$ will form his corresponding intention and they will do $J$. As we saw, however, if $X$’s intention is other-agent mediated, then it will only refer to her own action, and so the content does not refer to the collective action at all. On the other hand, if the intention does make reference to $Y$’s action as well as to her own, then the intention violates the own action condition, since agents cannot intend other agents’ actions. (Agents might be able to intend outcomes
3.4 We-intention theories

Some theorists avoid the problems that afflict Bratman’s we-action theory by arguing that collective goals are reducible to individually-held intentions, but that these individually-held intentions are distinct kinds of intentions from ordinary intentions. In this way we avoid the problem of understanding how ordinary intentions with irreducibly collective content avoid conflict with the own action condition. *We-intention theories* argue that the intentions which constitute collective goals are psychologically irreducible to the intentions which constitute individual goals. We will look at such arguments in this section.

The theorists who most prominently advance we-intention theories of collective action are John Searle and Raimo Tuomela. Their theories are intention-based, in that they think that we must build our theory of collective action upon we-intentions which are not reducible to the agent’s other mental states (such as beliefs and desires). These theories are still reductive theories, however, because we-intentions are held by the individual agents who deliberate and act upon them.

We will first sketch the theories of collective action that Searle and Tuomela offer; afterwards, we will see that we-intention theories fail to adequately account for collective action because the we-intentions they introduce are poorly suited to account for collective action at best, and incoherent at worst. In particular, Searle and Tuomela have a difficult time accounting for the conceptual links between we-intentions and I-intentions. Further, Searle and Tuomela posit that there must be a metaphysical distinction between we-intentions and I-intentions; this metaphysical distinction would, failing special circumstances, be accompanied by a phenomenological distinction between the two intention-states which is simply not present in actual instances of collective action.

3.4.1 Searle

Searle introduced the worry with which we began our discussion: what is the difference between a set of actions performed as part of a collective action, and that same set of actions performed merely as a set of individual actions? We have so far found the various attempts at solving the problem lacking. What is Searle’s answer, and is it a more promising solution?
3.4. WE-INTENTION THEORIES

Searle claims that we-intentions are a primitive form of intentionality which are not reducible to I-intentions plus associated beliefs and desires. These we-intentions, Searle claims, take the form ‘we intend that we perform act A’. We-intentions are held by individual agents, and we-intentions have what Searle calls ‘further intentional content’, such that each agent intends, of some particular act, that it be her or his contribution to the collective action. In reference to an example involving football players, Searle writes,

In cases like that of the football team each individual will have further intentional content, which in ordinary English he might express in the form “I am doing act B as part of our doing act A.” For example, “I am blocking the defensive end as part of our executing a pass play.”

For Searle, actions have two components to them: a mental component and a physical component. To quote Searle,

The mental component both represents and causes the physical component, and because the form of causation is intentional causation, the mental causes the physical [by] way of representing it.

Searle represents this with the following notation:

\[ i.a. \text{(this i.a. causes: my arm goes up)} \text{ CAUSES: MY ARM GOES UP} \]

and he then proceeds to explain his notation:

There is an intention-in-action which has as its conditions of satisfaction that that very intention-in-action causes it to be the case that my arm goes up; and all of that mental stuff really does cause it to be the case in the physical world that my arm goes up.

The mental (which is equivalent to an intentional attitude), then, is represented in the notation through lower-case letters, and the mental component has as its content that it cause an outcome by way of the agent’s physical behavior.

When it comes to collective action and collective intentions, the formula for intentions gets modified slightly in order to account for the irreducible collectivity of we-intentions.

---

25Searle 1990, p. 408.
26Searle 1990, p. 408. See also Searle 1983, in particular chapter 3, for a fuller account of Searle’s theory of individual intentions.
The intention in action that each agent in a collective action possesses is not the ordinary intention in action described above, but is rather an ‘achieve-collective-B-by-means-of-singular-A’ type of intention in action. The ‘singular A’ is the agent’s individual part of the collective action B, and the agent intends to produce B by way of doing A. Searle notates this collective intention in action as follows:

i.a. collective B by means of singular A (this i.a. causes: A stirred, causes: B mixed).\(^{28}\)

The irreducible collectivity is built into the representation of the collective intention in action as an ‘i.a. collective B by means of singular A’; individual intentions are not represented in this way. Thus, while Searle represents a complex means-end individual intention (such as firing a gun by means of pulling a trigger) as:

i.a. B by means of A (this i.a. causes: A trigger pulls, causes: B gun fires),\(^{29}\)

Searle represents complex means-end we-intentions not as an intention in action to B by means of A, but rather as a intention in action to collectively B by means of a singular action A. Nevertheless, the relationship between the means A and the end B is the same in both the individual and the collective case. As Searle claims,

The intention to stir is part of the [we-]intention to mix by means of stirring in the same way that in the gun case the intention to pull is part of the [I-]intention to fire by means of pulling.\(^{30}\)

In spite of this part-whole similarity between actions taken as means to individual goals and actions taken as means to collective goals, Searle clearly draws an ontological distinction between individual intentions and we-intentions; there is no way to reduce we-intentions to I-intentions. The relationship between the parts and the whole might be similar, but the wholes are, in both cases, distinct.

Before criticizing Searle’s account of collective action, let us first look at the theory Tuomela favors. As we will see, Tuomela’s account is very similar in its principals with the account that Searle offers. Given this, the criticism that we will have for Searle’s theory will apply equally to that of Tuomela.

\(^{28}\)Searle 1990, p. 412.
\(^{29}\)Searle 1990, p. 412.
\(^{30}\)Searle 1990, p. 413.
3.4. WE-INTENTION THEORIES

3.4.2 Tuomela

In Tuomela’s account of collective action, the principal action concept is that of jointly seeing to it that \( O \), where \( O \) can be e.g. ‘the participants’ joint action, somebody else’s action, some other agents’ joint action, a state in the world (like that a house is painted or the window is open)’\(^{31} \). Jointly seeing to it that \( O \) is an action which can be performed by a collective, but which cannot be performed by an individual.

The corresponding principal intention concept of collective actions for Tuomela is that of a joint intention. Two agents \( X \) and \( Y \) have a joint intention with respect to some outcome \( O \) when they jointly intend to see to it jointly that \( O \). The content of the joint intention, then, is that they jointly see to it that \( O \).

Joint intentions are held jointly; Tuomela suggests that there is a further intentional concept, that of a we-intention, such that a we-intention is held by individual agents who, as part of a group, jointly intend to see to it jointly that \( O \). Tuomela says that a we-intention is ‘a participant’s “slice” of their joint intention’\(^{32} \).

Tuomela does say that the formation of a joint intention (and hence of the corresponding we-intentions) must be sensitive to the various joint attitudes that the agents hold—in particular, of joint desires and joint beliefs. Tuomela does not, however, explain what he means by these joint mental attitudes, except to say that joint motivations (such as joint desires or joint beliefs) ‘need not be anything like an aggregation of private motivations but may instead be a compromise based on discussion, negotiation, or bargaining’\(^{33} \). It is not clear exactly how we are supposed to understand these jointly held motivating attitudes; in particular, it is not clear whether these jointly held motivating attitudes also have corresponding ‘personal slices’ in the same way that joint intentions do. Tuomela does suggest that there is no corresponding ‘personal slice’ to joint motivating attitudes; he says that

---

\(^{31}\)Tuomela 2005, p. 329.

\(^{32}\)Tuomela 2005, p. 330. Note that Tuomela suggests that we-intentions might be the more fundamental intention concept as compared to joint intentions; he says that ‘it can technically be said that a joint intention consists of the participants’ we-intentions about the existence of which the participants have mutual belief’. However, Tuomela goes on to write that

\[ \text{[even if we assume that a joint intention is (ontologically) composed of the agents [sic] we-intentions about which there is mutual knowledge (or belief), these we-intentions are different from ordinary action intentions \ldots in that they conceptually depend on the joint intention in question.]} \]

I take this ‘conceptual dependency’ to mean that, in some sense, joint intentions are conceptually primary (even if they are ontologically composed of we-intentions, whatever that might mean).

\(^{33}\)Tuomela 2005, p. 331.
CHAPTER 3. THEORIES OF COLLECTIVE ACTION

Joint desires and wants do not similarly involve making up one’s mind and do not lead to intentional jstitting [i.e. jointly-seeing-to-it-that] . . . and desires and wants do not require success beliefs as their normal (or “normal-rational”) accompaniments.\footnote{Tuomela 2005, p. 331.}

This quoted passage fails as a resolution to the question of whether joint motivating attitudes have corresponding ‘personal slices’ in individual agents. Unfortunately, this is the clearest that Tuomela gets to explaining himself on this point. Still, we need not worry about what exactly Tuomela means by ‘joint motivating attitudes’ here; the success or failure of Tuomela’s account of we-intentions is independent of his account of joint motivating attitudes, and we can charitably assume that Tuomela can fill in the content of joint motivating attitudes in a way which is acceptable.

If we consider the specific case in which the outcome produced is a joint action type (as opposed to the general case in which we consider a set of agents seeing to it that some outcome is produced or maintained), Tuomela’s theory of collective action states that joint intentions to $X$ (where $X$ is a joint action type) are typically formed on the basis of a publicly shared joint plan of action.\footnote{Although this Bulletin Board View of joint intention formation is the typical way in which joint intentions are formed, Tuomela emphasizes that it is not the only way in which such intentions can be formed. We will here disregard the other possibilities for intention formation, as the differences are irrelevant for our purposes here.} Agents who form we-intentions on the basis of this publicly shared joint plan of action—agents who accept the plan—form an intention to participate in the joint action which has two components to it: first, there is a we-intention to do the joint action, and second, there is a corresponding (regular) intention to do one’s part of the joint action (qua one’s share of the joint action). Tuomela considers as an example the joint action of cleaning up a park: one agent might publicly propose to the other agents the joint action of cleaning up a park someday. Agents who accept this proposal each form a we-intention to clean the park, and they also each form an intention to do her or his part in cleaning up the park (qua her or his part of cleaning up the park).

Thus, for we-intentions, each agent in a collective action holds a we-intention with the content ‘we together will do $X$’, or ‘we will jointly do $X$’, where $X$ is the joint action type. Tuomela’s more precise definition of we-intentions (WI) is as follows:

A member $A_i$ of a collective $g$ we-intends to do $X$ if and only if:

(i) $A_i$ intends to do his part of $X$ (as his part of $X$);
(ii) $A_i$ has a belief to the effect that the joint action opportunities for an intentional performance of $X$ will obtain (or at least probably will obtain), especially that a right number of the full-fledged and adequately informed members of $g$, as required for the performance of $X$, will (or at least probably will) do their parts of $X$, which will under normal conditions result in an intentional joint performance of $X$ by the participants;

(iii) $A_i$ believes there is (or will be) a mutual belief among the participating members of $g$ (or at least among those participants who do their parts of $X$ intentionally as their parts of $X$ there is or will be a mutual belief) to the effect that the joint action opportunities for an intentional performance of $X$ will obtain (or at least probably will obtain);

(iv) (i) in part because of (ii) and (iii).\(^{36}\)

Tuomela’s account of we-intentions and joint intentionality is best described as a we-intention theory of collective action because Tuomela’s we-intentions are defined partly in terms of an unanalyzed concept of collective actions. In part (i) of Tuomela’s (WI) definition above, Tuomela says that ‘$A_i$ intends to do his part of $X$ (as his part of $X$)’. There are two ways that we might understand ‘doing [one’s] part of $X$ (as [one’s] part of $X$’; we can understand $X$ as a cooperatively neutral description of a joint act type (as it was understood by Bratman above), or we can understand $X$ as a cooperatively positive joint act type. If Tuomela were to analyze the $X$ in (i) as cooperatively neutral, it is not clear whether he would be able to escape the problems we identified in Bratman above.

I want to argue, however, that Tuomela does not understand $X$ to be a cooperatively neutral description of a joint act type. Consider Tuomela’s account of what it means for an agent to intend to perform her or his part of $X$ as her or his part of $X$:

(2) $A$ intends to perform his part of $X$ as his part of $X$ if and only if:

(a) $A$ intends to perform his part of $X$, and

(b) $A$ intends to perform $X$ with the others in part because (of his belief that) the others intend to perform their parts of $X$ and intend to perform $X$ with the others.\(^{37}\)

There are two places in this definition where an unanalyzed collectivity might creep in: the first is in (2) (a), where Tuomela talks of $A$ intending to perform his part of $X$; the second is in (2) (b), where Tuomela talks of $A$ intending to perform $X$ with the others.


CHAPTER 3. THEORIES OF COLLECTIVE ACTION

We need not discuss the second worry, the one about condition (2) (b), here in any great
detail. If Tuomela intends to give an account of intending to perform one’s part of $X$ as
one’s part of $X$ which is reductive, then Tuomela has to ensure that $A$’s intention to perform
joint action $X$ with the others does not entail that $A$ intends to perform $X$ collectively or
jointly with the others. It is not clear whether Tuomela can give such an account. Even if
he can, however, he would still have to address similar worries about (2) (a). And, as we
will see in the discussion that follows, Tuomela is not able to address those worries while
remaining committed to a reductive account of we-intentions.

Let us, then, consider condition (2) (a). When Tuomela claims that $A$’s intending to
perform his part of $X$ is a necessary condition of intending to perform his part of $X$ as his
part of $X$, he might mean that $A$ must merely intend to do those (cooperatively neutral)
actions which constitute her or his part of $X$. If this were the case, then $A$ would satisfy
(2) (a) so long as $A$ intends to perform some action, regardless of her or his beliefs about the
possible actions of other agents and regardless of the consequences of her or his performing
the action. This requires us to read ‘his part of $X$’ narrowly: ‘his part of $X$’ would pick
out only a specific action, and not the intentional or behavioral context of the action. The
narrow reading picks out only the action itself, and not the joint action of which the action
is a part.

In addition to this narrow reading of (2) (a), there is also a broad reading of (2) (a):
$A$ might intend to perform his part of $X$ only in those cases in which $A$ is aware of the
existence of a set of actions which jointly suffice to produce $X$, and of which $A$’s intended
action is a member. On the broad reading, $A$ can only intend to perform his part of $X$ if
he is aware that there are other agents with the option to perform actions which are their
parts of $X$. In other words, the broad reading requires that $A$ view his action as one of a
set of actions; under the narrow reading, $A$ can intend to perform his part of $X$ even if $A$
is ignorant of those other actions which together suffice to produce $X$.

Finally, there is what we can call a very broad reading of (2) (a). The very broad
reading of (2) (a) holds that the content of $A$’s intention to ‘perform [one’s] part of $X$’ is
not some specific action of $A$’s (whether the action is understood in a joint action context
or not); rather, the content of $A$’s intention is, in some way that remains to be specified,
the entire joint action $X$ (though $A$’s contribution to $X$ might be uniquely highlighted).
If, for instance, we were to claim that $A$’s intending to play his part of $X$ also requires
that $A$ intend to do those actions necessary to ensure that the other agents in the collective
action successfully perform their actions, then we would be adopting a very broad reading of (2) (a).

It is not clear whether Tuomela endorses a narrow, broad, or very broad reading of (2) (a). Consider the following analysis by Tuomela:

...[W]hile the preanalytic account does not, to quote Searle again, “construe ‘doing his part’ to mean doing his part toward achieving the collective goal”, it might still be argued, from a theoretician’s point of view, that underlying (2) there is implicit reference to joint intention. This is because in (2) A’s intention to perform his part of X as his part of X can be satisfied only if the participants intentionally jointly perform X and because such intentional joint performance may be argued upon analysis to depend on the joint aim or intention to perform X.38

This would seem to lead us to adopt a narrow reading of (2) (a); if ‘doing one’s part’ does not mean ‘doing one’s part toward achieving the collective goal’, then it would seem as though the content of the agent’s intention makes no reference to the collective goal (and hence the set of actions which jointly produce the collective goal) at all.

When we adopt the point of view of the ‘theoretician’, however, the narrow reading becomes less felicitous. According to the theoretician, the satisfaction conditions for an agent’s we-intention include that the joint action X be intentionally performed by the agents, and this leaves the concept of ‘joint action’ (or ‘collective action’) essentially unanalyzed.

Now, the narrow reading of (2) (a) does lead us to a reductive account of we-intentions. It is unclear whether the broad reading of (2) (a) also leads us to a reductive account of we-intentions, however; the answer to this question depends primarily on how we interpret the intension of the set of actions which produces the collective outcome. If the intension of the set includes that the actions promote the collective outcome, then the broad reading of (2) (a) entails that we cannot reduce we-intentions to ordinary I-intentions. If, on the other hand, the intension of the set under a broad reading of (2) (a) does not include that the set be a set of collective actions, then we seem to be able to reduce we-intentions to I-intentions. The very broad reading of (2) (a) also makes we-intentions irreducible to I-intentions, as the content of the agent’s intention includes actions of other agents.

This might lead us to reject the point of view of the theoretician in favor of the pre-analytic point of view, since the pre-analytic point of view enables us to adopt the narrow reading of (2) (a). The narrow reading of (2) (a) is implausible as an account of collective

---

38Tuomela 2005, p. 357.
CHAPTER 3. THEORIES OF COLLECTIVE ACTION

action, however; at best, the narrow reading is an analysis of mutually responsive interdependent action. Suppose we have two agents, X and Y, and suppose we adopt a narrow reading of (2) (a). X might intend to walk on the right side of the road, and Y might intend to walk on the left side of the road (his right, as they are walking towards one another). Then we can satisfy Tuomela’s conditions for we-intentions: (2) (a) is satisfied, as X intends to walk on the right and Y intends to walk on the left. Further, though neither conceives of their actions in this way, X’s walking on the right and Y’s walking on the left are parts of a joint action type in which X walks on the right and Y walks on the left. (2) (b) is also satisfied, as X intends to walk on the right because of her belief that Y intends to walk on the left, and the same applies, mutatis mutandis, for Y.

It is false to say of X and Y that they are engaged in a collective action, however; X’s intention to walk on the right is formed in part because she believes that Y intends to walk on the left, but this fact does not commit them to participating in a collective action with respect to their walking. Not every instance of X’s and Y’s walking down a pathway, with mutually responsive behavior, counts as a collective walking. If we were to adopt a narrow reading of (2) (a), then we would have to count all such mutually responsive interdependent actions as collective actions, though we ought not do so.

Tuomela can avoid this problem by adopting a broad or a very broad reading of (2) (a). This would put Tuomela’s theory of action at risk of vicious circularity, however. If collective action is defined in terms of we-intentions, and we-intentions are defined in terms of intending to perform one’s part of some joint action X, and X’s being a joint action is defined in terms of collective action, then the account would be viciously circular.

Tuomela appears to recognize this problem; he writes,

I can go along with my critics to the extent that doing one’s part in a sense presupposes collective (or joint) intention but only in an implicit and unanalyzed sense of aiming at the joint action.\(^3\)

As Tuomela goes on to argue, however,

I deny that this sense creates vicious conceptual circularity which commits me to define we-intentions in terms of we-intentions. Recall that my analysis is not meant to be reductive but is rather meant to elucidate the irreducible notion of we-intention in a fundamentally informative way, and this is what the analysis of my (WI) [i.e. we-intention condition] and (2) help to do without strictly

\(^3\)Tuomela 2005, p. 357.
relying on the concept of we-intention—even if some amount of circularity will be there.\textsuperscript{40}

This is a confusing response. On the one hand, Tuomela is arguing that there is no vicious circularity in his definition of we-intentions; on the other, Tuomela is arguing that he is not trying to give a reductive account of we-intentions at all, but is rather merely trying to ‘elucidate the irreducible notion of we-intention in a fundamentally informative way’.

I think that the only consistent way to understand Tuomela here is to accept that he is not attempting to reduce we-intentions to I-intentions at all; Tuomela’s we-intentions are ontologically distinct from, and not reducible to, I-intentions. This means that we can accept a broad or a very broad reading of (2) (a) without worrying about the irreducibility that such readings introduce.

Tuomela gives us further evidence for this reading of him when he writes,

I would like to emphasize that my analysis of joint intentions and we-intentions is conceptually non-reductive, although it is ontically individualistic or, better, interrelational.\textsuperscript{41}

Tuomela frequently writes as though he is giving a reductive analysis of we-intentions, in particular when he is defending his theory against charges laid against him by Searle that he is giving a reductive (and therefore circular) account of we-intentions. Searle’s charge of vicious circularity is clearly a non-starter, however, given that Tuomela takes himself to be giving an account of we-intentions which does not attempt to reduce we-intentions to I-intentions.

The consequence is that, on the question of the reducibility of we-intentions to I-intentions, Tuomela’s position is identical to Searle’s: collective actions occur when individual agents act toward a common goal through we-intentions. Insofar as either account faces problems in virtue of its irreducibility, those problems will be faced by the other as well.

3.4.3 On the conceptual links between we-intentions and I-intentions

We-intention theories of collective intentions, such as Searle’s and Tuomela’s, must make two claims. The first claim they must make is that theories of collective intentions which

\textsuperscript{40}Tuomela 2005, pp. 357–358.

\textsuperscript{41}Tuomela 2005, p. 342.
reduce collective intentions (i.e. we-intentions) to individual intentions (i.e. I-intentions) fail to adequately account for collective action. The second claim they must make is that it is possible to give a non-reductive account of collective intentions which properly captures the various conceptual links which hold between we-intentions and I-intentions, or between collective goals and individual intentions and behaviors which are directed towards those collective goals.

Searle makes a clear argument that reductive theories cannot adequately account for collective action; this was his reason for giving the capitalist businessmen example that we used to initiate our discussion. Tuomela, on the other hand, does not explicitly make a similar argument. That Tuomela (and any other defender of a non-reductive theory of collective intentions) needs to make such an argument is clear, however. Occam’s razor dictates that we prefer more reductive theories over less reductive theories (or non-reductive theories) when their explanatory capacities are identical; any appeal to a less reductive theory must be justified by an unavoidable explanatory gap in the existing reductive theories. Both fully reductive and we-action theories are more reductive than we-intention theories, since both types of theories appeal to ordinary mental states (and thus a more sparse ontology).

Answering this first claim—that there is a problem with existing reductive theories—requires us to present a plausible theory of collective action which is more reductive than that offered by Searle or Tuomela. We will begin to build such a theory in the following chapter, and so we cannot hope to give a proper answer to this claim here. It suffices for us to say here that if we can give a theory of collective action which is built upon the same mental attitudes which are needed to theorize individual action, then we will have thereby given an argument against non-reductive theories of collective intention and collective action.

The second claim is a more difficult one to address. Indeed, it is not clear whether a more reductive theorist can give any convincing argument against the second claim made by the we-intention theorist. The reason is that if the more reductive theorist has a convincing account of collective action on hand, then the we-intention theorist can simply appeal to the account offered by the more reductive theorist to answer any challenges the more reductive theorist might raise. This is particularly true if the we-intention theorist does not offer a detailed account of the differences between the non-reductive we-intentions and ordinary I-intentions.

For example, suppose the we-intention theorist says of we-intentions that they are, by
and large, *exactly like* ordinary I-intentions, except that when agents act upon we-intentions (along with other conditions), they are engaged in collective action, whereas when agents act upon I-intentions, they are engaged in individual action. Then the more reductive theorist does not have a successful argument against the we-intention theorist; any argument which the more reductive theorist can refute can be similarly refuted by the we-intention theorist, who will merely substitute in the (nearly identical) we-intentions for I-intentions.

Although it is thus difficult to develop knock-down arguments against we-intention theories of collective intention and collective action, there are two concerns that we ought to have about such theories which are independent of whether or not we have a more reductive theory of collective action on hand. One concern, about the phenomenological properties of we-intentions and I-intentions, will be treated next section. The other concern we might have is about the conceptual relationship between we-intentions and I-intentions; this concern will be treated here.

Put simply, the conceptual link worry is about the relationship that ought to exist between we-intentions and I-intentions. In the individual case, I-intentions often spawn new I-intentions; my I-intention to prepare dinner generates in me an obligation to form an I-intention to figure out what to cook, to purchase the food from the market, and so forth. Individual goals are, for this reason, *stackable*; as Bratman has consistently argued, intentions are typically elements in larger plans.\(^{42}\) That individual goals are stackable in this respect is not remarkable; intentions are goal-oriented, and they direct the possessor to take those means necessary to bring about the intended goal. Because the means might well be complex, however, involving many action-steps, it is not at all unusual that an agent might need to form several intentions in pursuit of the original goal.

For collective goals, however, it still needs to be explained why an agent who holds a we-intention faces pressure to form I-intentions in pursuit of the collective goal. Why does my holding a *collective* goal, e.g. that we prepare a sauce together, generate in me *individual* intentions, e.g. that I mix the sauce? Why, in other words, don’t we-intentions generate in their possessor *only* subsequent we-intentions?

Note that there are two cases to be accounted for here. The first case involves actions which an individual agent pursues in pursuit of a collective goal, as when agent \(X\) mixes the sauce as her part of preparing the sauce with \(Y\). The second case involves agents who possess a collective goal, and who must engage in subsequent deliberation as to what to

\(^{42}\)Bratman 1987.
do in pursuit of that goal. As an example, X and Y might be soldiers who each hold the collective goal of capturing an enemy position with the other, but they are not in contact with each other and have not already coordinated their plans. As a result, X and Y must each deliberate about which actions to perform such that they will best contribute to the satisfaction of the collective goal.

Searle’s explanation, as we have seen above, is that I-intentions and we-intentions stand in a ‘part-whole’ relationship; my I-intention to mix the sauce is a part of my we-intention that we make the sauce, in the same way that my intention to pull the trigger is a part of my intention to fire the gun. Searle argues that in the gun-firing example, there is only one action, not two, and so there is only one intention; similarly, he argues, in the sauce-making example each agent performs only one action, and so each agent should have only one intention, not two.

This argument might work to account for the first conceptual relationship between we-intentions and I-intentions—namely, that which obtains when the I-intention and the we-intention are both held concurrently, and the agent performs an action in pursuit of the collective goal. It is less clear how Searle’s response works in the second conceptual relationship, however, when the agent possesses the collective goal but has not yet deliberated about which individual action to perform in pursuit of the collective goal. If Searle is correct in arguing that I-intentions and we-intentions stand in a part-whole relationship to one another, it is puzzling that an agent might hold a we-intention but not yet hold any corresponding I-intention; it seems as though, without the I-intention which is a part of the we-intention, there is not yet a whole intention for the agent to possess.

And yet, the collective goal (as, for example, the collective goal to capture the enemy position) is perfectly legitimate as an intention. After all, it does what an intention should do: it structures further deliberation, it represents an agent’s volitional commitment towards actions, and so forth. Searle must therefore allow such we-intentions to be perfectly legitimate as intentions that an agent might hold, and yet it is not clear how he can do so.

If we take a step back, we can see what Searle’s problem is. The question we asked initially was why we-intentions have a such a tight conceptual link with I-intentions, such that we-intentions are characteristically accompanied by I-intentions. My possessing a collective goal also generates in me an individual intention (or a practical commitment to generating an individual intention). Searle’s answer is to posit a very tight conceptual connection, such that we-intentions and I-intentions always accompany each other: I-intentions are parts of
3.4. WE-INTENTION THEORIES

the whole that is the we-intention. As we noted, this answer might work in those cases where the I-intention and the we-intention are both present. However, this answer is problematic when the I-intention is not yet present—how can an agent have the whole, but not have the parts which make up the whole?

Unsurprisingly, Tuomela faces a similar problem to Searle. Tuomela attempts to account for the conceptual link between we-intentions and I-intentions by arguing that minimally rational ‘we-intenders’ reason in accordance with two schemas of practical inference:

(W1):
(i) We will do $X$; therefore,
(ii) I will do my share of $X$.

and

(W2):
(i) We will do $X$;
(ii) $X$ cannot be performed by us unless we perform action $Z$ . . . ;
(iii) We will do $Z$; therefore,
(iv) Unless I perform $Y$ we cannot perform $Z$; therefore,
(v) I will do $Y$ (as my contribution to $Z$).\(^{43}\)

The first schema (W1) holds that agents who possess a we-intention to do $X$ are thereby committed to doing ‘their share of $X$’; the second schema (W2) holds (assuming we are charitable with the occasional errant ‘therefore’) that agents who possess a we-intention to do $X$ have a commitment to possess we-intentions to do those actions $Z$ which are necessary means of doing $X$, and that they are accordingly committed to (I-)intending to do those actions $Y$ which are necessary means of doing $Z$.

For Tuomela, unlike for Searle, the conceptual connection between we-intentions and I-intentions is not \textit{constitutive} (as part-whole relationships are), but is rather \textit{rational}: a minimally rational agent who holds a we-intention will also hold select I-intentions in accordance with schemas (W1) and (W2).

The problem with Tuomela lies in determining just when some action is the agent’s ‘part’, and what exactly the content of the agent’s I-intention is. Does the agent intend to do her part of joint action $X$, \textit{whatever her part may be}? Or does the agent rather intend to

do some particular action $A$ which she understands to be her part of $X$? If it is the latter, then Tuomela faces the same problems that Searle faced: what happens if the agent intends the joint action $X$, but has not yet acquired through deliberation some individual action $A$? If, on the other hand, it is the former, then Tuomela has merely delayed giving the account which is necessary. Suppose agent $X$ we-intends to do collective action $J$. $X$ is a rational agent; therefore, $X$ has the I-intention to do her part of $J$. The question $X$ now faces is, in virtue of what does she understand some individual action to be her part of the joint action $J$? In virtue of what does $X$ apply the concept ‘my part of $J$’ to some individual action? Unless Tuomela can explain this crucial step, then his account of collective action seems empty. The worry, though, is that whatever Tuomela can say here will smack of ad hocery. What we want of we-intentions is that they structure an agent’s deliberation towards acquiring supporting I-intentions. We know why holding an individual intention structures deliberation towards acquiring supporting individual intentions; Tuomela has given us no explanation of why the same is true for we-intentions, except to simply say that it is a rational requirement. This is not explanation, however, but is rather just description.

We might think that schema (W2) provides assistance to Tuomela here. After all, in this schema Tuomela provides a reason why rational agents reason in this way: they are just taking the necessary means towards the production of the collective goal. The problem is that no action is a necessary means independent of what the other agents do. Suppose agents $X$ and $Y$ are attempting to capture the enemy position, and they each have two options: they can attempt to storm the position from the north, or they can attempt to storm the position from the south. They both recognize that storming the position from the north gives them slightly more favorable ground, and so gives them a better chance of success. They also know that if they fail to coordinate on the direction from which they storm the position, they are guaranteed a catastrophic defeat.

This is, of course, a kind of Hi-Lo game; if $X$ and $Y$ coordinate on North, then they have a good chance of success; if they coordinate on South, then they have a less good chance of success; and if they fail to coordinate, they have no chance of success at all. It is important to note, however, that their storming from the north is not a necessary means to the seizing of the enemy position; they might also succeed if they storm from the south. And, importantly, even if Tuomela were to modify (W2) so as to claim that rational we-intending agents must take necessary means towards those collective actions which maximize their chance of satisfying the collective goal, that would not be enough,
as evinced by the case in which storming from the north and storming from the south are successful with equal probability.

Again, we can intuitively see where Tuomela has gone wrong. Unlike Searle, Tuomela does not try to ontologically bind we-intentions to I-intentions; rather, Tuomela wants to insist that the conceptual link between we-intentions and I-intentions is a rational one. Agents who possess we-intentions will face a rational pressure to form the right kinds of I-intentions. The problem is, Tuomela doesn’t say what those right kinds of I-intentions are. In the individual case, practical deliberation is guided by means-end reasoning: agents take those means necessary towards one’s end. When the goal is collective, however, that means-end analogy breaks down; while an agent might face a rational pressure to form those we-intentions which are necessary means to the satisfaction of a we-intention, it’s not clear why an agent faces a rational pressure to form I-intentions in response to possessing a we-intention.

As mentioned earlier, this is not meant to be a knock-down argument against Searle or Tuomela. It is possible that either theorist could respond to these arguments by appealing to the same resources that a more reductive theorist, such as a fully reductive theorist or a we-action theorist, will use to respond to these same worries. However, Searle and Tuomela can only purchase this response at the cost of their theories becoming ad hoc. There is a practical connection between individual goals and the attitudes that an agent forms as means to achieving those goals; this practical connection does not clearly exist when the goals are manifested in distinctly collective mental attitudes but the means must be manifested in individualistic ones.

3.4.4 Phenomenological properties of collective and individual intentions

In addition to the worries mentioned above, there is another worry that we ought to have about we-intention theories of collective intention. As with the conceptual link worries mentioned above, none of the arguments to be presented here are knock-down arguments; they do not compel us to reject we-intention theories. The we-intention theorist is in a precarious position, however. Because the we-intention theorist employs two distinct mental attitudes, I-intentions and we-intentions, to account for the functional role of intentions, the we-intention theorist has to account in some way for those cases in which I-intentions and we-intentions are, in essence, indistinguishable. The task facing the we-intention theorist, then, is to explain why I-intentions and we-intentions are, in so many respects, qualitatively
identical, in spite of their being different mental states. Moreover, the we-intention theorist must explain this similarity in a way which is not unmotivated or ad hoc.

There is, however, one important way in which the distinction between I-intentions and we-intentions can be taken to be a positive feature of a comprehensive theory of action and not a negative feature. It can be argued that actions taken in pursuit of collective goals have different phenomenological properties than actions taken in pursuit of individual goals, and that the only way to account for these different phenomenological properties is to posit the existence of distinct intention-states.

An argument of this sort is implied by Searle’s insistence that all collective intentionality be compatible with brain-in-a-vat scenarios. Searle expresses this insistence as follows:

[Our theory of collective action] must be consistent with the fact that the structure of any individual’s intentionality has to be independent of the fact of whether or not he is getting things right, whether or not he is radically mistaken about what is actually occurring . . . . One way to put this constraint is to say that the account must be consistent with the fact that all intentionality, whether collective or individual, could be had by a brain in a vat or by a set of brains in vats.44

For Searle, collective intentionality is located within attitudes held in the head of individual agents. The agent who intends to act in support of a collective action or a collective goal has a we-intention, and the agent’s possessing the we-intention does not depend in any way on conditions outside of her or his head, as it were—the agent’s possessing of the we-intention does not, in particular, depend on the existence of other agents, or on their mental states.

Suppose an agent $X$ holds an intention to do some action $A$. Let us say that the phenomenological discrimination thesis holds that $X$ knows, in virtue of phenomenological properties of the two mental states, whether or nor she I-intends to $A$ or whether she we-intends to $A$. For Searle, of course, a we-intention to $A$ will be connected to some collective outcome $J$ such that $X$ intends to $J$ by way of $A$-ing. (We are thus being a bit imprecise in talking about a we-intention to $A$; it would be more precise to say that $X$ has a we-intention to $J$, and that her intention to $A$ is a part of her we-intention.) The phenomenological discrimination thesis would be made true if we accepted the thesis that mental states are transparent; if an agent always knows what mental state she or he is in, then the agent must know whether she or he is in a state of we-intending or I-intending.45

---

45Whether mental states are transparent is still hotly debated; see, for example, Williamson 2000 for an
We need not accept transparency of mental states, however, to accept the phenomenological discrimination thesis, which is weaker. Indeed, Searle’s theory suggests that we ought to accept the phenomenological discrimination thesis; if an agent holds a we-intention, then the agent must also hold an I-intention which is a part of the we-intention. Further, of course, we-intentions have the group as part of their content—an executive part of their content, as it were, as the group is regarded as the author of the collective action. Though Searle does not argue that I-intentions cannot have the group as an executive part of their content, we can plausibly attribute this claim to Searle.

Suppose we accept the phenomenological discrimination thesis. Is this a positive or a negative for a non-reductive theory of collective intention and action? The defender of the non-reductive theory might think it to be a positive; after all, agents know whether or not their action is directed towards an individual goal or towards a collective goal, and so that should be reflected in our theory of action. The non-reductive theory builds this knowledge directly in at the ground floor, by allowing that intentions directed towards individual goals and intentions directed towards collective goals are different mental states and therefore have different phenomenological properties.

We might worry, however, that there is not always a clear phenomenological difference between acting on behalf of individual goals and acting on behalf of collective goals. In particular, an agent might simply not know whether or not some goal is an individual goal or a collective goal. There are, I think, actions and goals which are liminal, in that they might be individual goals, they might be collective goals, and they might be both. This situation is most prevalent, I think, in cases where one agent is ignorant about the mental states of others. We can appeal to an example which Margaret Gilbert makes heavy use of—that of a shared walk between two agents. Let $X$ and $Y$ be out on a walk. Suppose further that $X$ and $Y$ are walking along the same path, and they each believe this to be the case. There are clear instances in which $X$ and $Y$’s walk is a collective action: $X$ and $Y$ agreed beforehand to walk together, they each approve of their plans for their shared walk, and so forth. There are also instances in which $X$ and $Y$ are clearly not on a shared walk together: $X$ and $Y$ do not know each other, do not plan to know each other, and merely recognize that they are walking along the same path at the same time. (Perhaps they each correctly believe that they are walking along the same path to a common event, such as a sporting event.)

argument against the transparency of mental states.
In addition to these cases, it seems to me that there are liminal cases: cases in which neither $X$ nor $Y$ are certain whether they are engaged in a collective action or an individual action. $X$ and $Y$ might each desire that their walk be collective—$X$ might find herself appreciating $Y$’s company, and $Y$ might worry for his safety in walking along the road and desire the security of walking with $X$. Further, they might tentatively deliberate as though they were engaged in a collective action: $X$ might slow down if she sees herself pulling away of $Y$; $Y$ might consider altering his route if he sees $Y$ starting to veer off. Their ignorance over whether they are engaged in a collective action or individual actions (perhaps even conditionally interdependent individual actions) seems to be based to some degree on their failing to have certain beliefs about the beliefs and desires of the other agent: $X$ does not know whether $Y$ also desires or intends that their walk be a collective action, and so $X$ does not know whether her own actions are in support of a collective goal or an individual goal.

Note that this does not deny Searle’s claim that the structure of an agent’s intentionality must be independent of whether she or he has gotten facts about the world right. It might be incompatible with $X$’s intending to act upon a group goal that she believe that the other agents in the group (in this case, $Y$) do not intend to act upon the group goal at all. If $X$ does not have a certain belief about $Y$’s intentions, then, $X$ might be uncertain whether her own intention is best described as being an individual intention or a we-intention.

If I am right that such examples exist—that agents might be uncertain as to whether they are engaged in an individual action or a collective action—then the we-intention theory of collective intention faces a problem: how do you account for the fact that an agent might be in a mental state which is phenomenologically indeterminate between being a we-intention and being an I-intention? The we-intention theorist might, of course, deny the phenomenological discrimination thesis. This seems like a high price to pay, however. Most of the benefit of having two distinct mental states of intending comes from something like the phenomenological discrimination thesis: if agents are ignorant about what mental state they are in, then what is the purpose of posting more than one in the first place? Our goal here is not to capture the metaphysical truth of whether agents are engaged in a collective action or an individual action; rather, our goal is to capture the distinctive form of reasoning and justification involved in working towards a collective goal rather than an individual goal. And for this, an agent must know whether or not she or he possesses a we-intention or an I-intention.

A more reductive theorist, on the other hand, does not face such problems. Insofar as
the more reductive theorist appeals to the same mental attitude to account for individual
goals and collective goals, the more reductive theorist can argue that any uncertainty as to
whether an agent is acting on behalf of a collective goal or an individual goal is caused by
uncertainty in other mental states (such as beliefs or desires, whether those mental states
are her or his own, or those of the other agents in the group). We will see in the following
chapters how the theory I propose, which is a version of a we-action theory and is therefore
more reductive than the we-intention theory, can handle such cases.

Searle and Tuomela offer we-intention theories of collective action in that they believe
that collective intentions are distinct psychological attitudes to individual intentions. As
we have seen, both of their theories are ontologically reductive—they both argue that col-
lective action and intentionality can be entirely accounted for through mental attitudes
that are held in the heads of individuals. This is a compromise strategy; it attempts to
remedy the problems of purely reductive theories by positing the existence of non-reductive
mental states (as opposed to non-reductive contents), while still insisting upon an ultimate
reduction of groups to individuals. As we have seen, the problem with this compromise is
that we lose the methodological benefits of a purely reductive theory, and we also stand in
need of an explanation of how the group-oriented attitudes held by individuals relate in the
appropriate ways to individual-oriented attitudes.

In the following section we will consider a strategy for accounting for collective action
which avoids the compromise position by falling even more heavily on the side of non-
reducibility: instead of arguing that collective action can be accounted for by irreducibly
distinct mental attitudes possessed by individual agents, plural subject theories argue that
collective action can be accounted for only by appeal to irreducibly collective agents.

3.5 Plural subject theories

We will call Margaret Gilbert’s view of collective action, and Abraham Sesshu Roth’s view
of collective action, plural subject theories of collective action. As we will see, this is a
somewhat misleading description for Roth; ‘plural subjects’ is a term introduced by Gilbert
and plural subjects play a significant role in her theories of sociality and agency. And,
while Roth is strongly influenced by Gilbert, he does not employ the term ‘plural subject’.
Furthermore, Roth’s theory of collective action is ontologically reductive; Roth explains
collective action through the actions of individual agents. At the risk of some imprecision,
we will here call both of them plural subject theories; their link is that they are distinguished
by the fact that they each reject a view of agency that we can call the reductive view. According to the reductive view of action, agents are motivated solely by mental attitudes
that they themselves individually hold. The reductive position is compatible with a wide
range of contents of the mental attitudes held by the individual agent, just as the reductive
theorist can accommodate a wide range of types of mental attitude. What the reductive
theorist argues is that insofar as a mental state is relevant for an agent’s deliberation and
action, that mental state must be properly possessed by the agent herself or himself.

It is instructional to compare the reductive view of agency with Amartya Sen’s sympathetic agent. Sen distinguishes between sympathy and commitment; the sympathetic agent
is the one whose personal welfare is affected by the conditions of others, while the committed agent is one who is motivated to act in spite of experiencing no personal effect upon her
or his utility by the conditions of others. The sympathetic agent, Sen argues, is egoistic,
in that the sympathetic agent can act only upon her or his own welfare. The committed
agent, in contrast, is not egoistic; the committed agent might choose in a way which is
anticipated to run contrary to her or his maximizing utility or individual interest.

The reductive theorist is not limited to the position that agents are motivated purely
by sympathy. Sympathy appeals purely to utility; the sympathetic agent is one who acts
to maximize utility, though her or his utility might possibly be influenced by the conditions
(including mental states) of others. By contrast, the reductive theorist takes into account
any mental state which is possessed by the individual agent and which might possibly
influence her or his decision-making. Thus, the reductive theorist can appeal to an agent’s
intentions to act, while the sympathetic agent cannot (unless the agent’s intentions can be
converted directly into impacts on utility). If utility were the only mental attitude relevant
for an agent’s decision-making and intentional action, then the reductive agent would be
identical to the sympathetic agent; since there are more relevant mental states than just
utility, however, the reductive position is broader.

All of the views that we have considered so far in this chapter have been reductive views.
Under the team preference approach, agents are motivated by their preferences, whether
team preferences or individual preferences; both types of preferences are mental states held

---

46 see Sen 1977 and, for further discussion, Sen 1985.
47 We are assuming here, for simplicity, that utility is (or can be expressed as) a mental state. Our argument continues to hold so long as impacts on an agent’s utility are representable in the agent’s mental states.
3.5. PLURAL SUBJECT THEORIES

in the head of the individual deliberating agent. We-action theories of collective action take agents to be motivated by their intentions. We-intention theories, like we-action theories, take agents to be motivated by intentions; the only distinction here is that the motivating intentions might have as their content either individual actions or collective actions. All of these theories hold in common that individual agents are motivated by individually-held mental attitudes; while the theories differ on the type and the content of the mental attitudes, they all agree that insofar as they are capable of motivating the deliberating agent, they must be possessed by the deliberating agent.

By contrast, Gilbert and Roth argue that accounting for collective action requires us to take a different tack: they argue that the only theories capable of accounting for collective action are ones which reject the reductive position. Consequently, they each present theories of collective action which reject the reductive assumption that agents can only be motivated in their deliberation and action by individually-held mental states. This position is a tenuous one: as with we-intention theories, Occam’s razor dictates that we should only accept such theories if we lack more reductive theories of agency and collective action. Gilbert and Roth therefore present arguments which purport to show that reductive theories of action are systematically incapable of accounting for the unique features of collective action. We will first look at Gilbert’s and Roth’s arguments that reductive theories are systematically incapable of accounting for collective action; we will next consider their positive proposals. Finally, we will see why Gilbert’s and Roth’s theories are unsatisfactory as accounts of collective action. Gilbert and Roth fail on two counts: not only do they fail to offer a convincing negative argument against reductive theories of collective action, but they also both fail to present a plausible positive theory of collective action.

3.5.1 Gilbert and Roth against reductive theories

Plural subject theories have one primary objection to the claim that reductive theories of agency can account for collective action. The objection depends on the intrinsically normative character of collective action. According to plural subject theorists, participants in collective actions necessarily have normative commitments (or obligations) towards one another. These normative commitments are unique; they cannot be accounted for through reductive theories of action.

Or, so the argument goes. What are these normative commitments, and why is it argued that reductive theories of action cannot account for them?

Directed rights and obligations

One of these commitments is a directed obligation to contribute to the collective goal, and an associated directed right to the contributions of others. These rights and obligations are directed, because they are not universal or general obligations or rights of morality; rather, they are rights and obligations that are owed specifically to the other members of the collective action.\footnote{Gilbert talks specifically of directed obligations in her discussion of Scanlon’s theory of promises in Gilbert 2004. Gilbert does not claim that all collective actions involve promises, and so it is not obvious that her discussion there is relevant for her discussion of reductive theories of agency. Gilbert’s criticism of Scanlon is closely related to her criticism of reductive theories, however, and so we will appropriate her term ‘directed obligations’ here.}

Gilbert characterizes these directed rights and obligations by appeal to an example. Suppose, she argues, that two agents, Sue and Jack, are engaged in a collective action of going for a walk together. While walking, Jack pulls ahead of Sue. Of the directed rights that Sue has towards Jack, Gilbert says:

\[W\]e can imagine Sue taking action in various ways. She might call out, “Jack!” with a degree of impatience. She might catch up with him and then say, somewhat critically, “You are going to have to slow down! I can’t keep up with you.” In both of these cases she rebukes Jack, albeit mildly. She might not do this, of course, but it seems that, again failing special circumstances, her doing so would be in order. In other words, it seems that in the circumstances sue is entitled to rebuke Jack. We would expect both Jack and Sue to understand that she has this entitlement.\footnote{Gilbert 1996b, p. 180, emphasis original.}

Gilbert goes on to characterize Jack’s obligations towards Sue as follows:

The existence of this entitlement suggests that Jack has, in effect, an obligation to notice and to act (an obligation Sue has also) . . . . We would expect those out on a walk together to realize that they have the obligations, and the rights, just noted.\footnote{Gilbert 1996b, p. 180, emphasis original.}

It is important for Gilbert’s account that these rights and obligations are directed towards the other participants in the collective activity. Suppose Jack and Sue are being observed by some third party Alice. Alice is not participating in the shared walk, although
Alice is fully informed about Jack and Sue’s intended shared walk. Alice is also in a position to call out to Jack and inform him that Sue cannot keep up. Alice might tell Jack that, if he is to carry out his intention (and Sue’s intention) that they engage in a shared walk, he must slow down. Alice might even mildly rebuke Jack for walking too quickly; she might look down on people who walk too fast for their walking companions, or who are so self-absorbed as to be unaware that their walking companion has fallen behind. Alice might behave in these ways towards Jack, but there is an important difference between Alice’s behavior towards Jack and Sue’s behavior towards Jack.

Gilbert suggests that we should understand Alice’s behavior in one of two ways. First, Alice might be informing Jack of what prudence requires of him. That is, Alice might be informing Jack that if he intends to walk with Sue, then slowing down is a prudential means for doing so, but that walking at his current pace is not. And second, Alice might be informing Jack about his normative commitments. Jack has a directed normative commitment to Sue, but he might be temporarily unaware that he has this normative commitment, or he might be unaware that his current behavior is violating his commitment.

This should be contrasted with Sue’s rebuke of Jack. Sue is not merely telling Jack what prudence requires of him, nor is she merely informing him about his pre-existing normative commitments towards her. Rather, Sue’s complaint to Jack comes from a privileged position; she has a special entitlement to criticize Jack. This entitlement comes from her status as a co-participant in the collective action—a status that is not shared by Alice.

Roth clarifies this special entitlement of Sue’s as follows:

Sue’s criticism should be seen as part of the shared endeavor. Criticism from nonparticipants cannot be seen in this light. Whether or not welcomed by the group as a whole, such third party criticism is not a part of the exercise of shared agency. So if Sue is in this special position to criticize Jack, this suggests that a distinctive form of commitment is a basis for her criticism, one that does not serve as a prima facie ground for nonparticipants to object to Jack.\(^{52}\)

Roth calls such commitments contralateral commitments. Contralateral commitments, Roth argues, possess a distinctive special and executive nature which are not present in other types of normative commitments, including (as we will see) commitments of a moral nature. The commitments are special, because they only hold between participants in the collective action. The commitments are executive, because any criticism (or rebuke) over a failure to fulfill the commitments is treated as a part of the activity itself. To quote Roth,

\(^{52}\)Roth 2004, p. 364.
The commitment is executive in that the criticism it would license would be a part of the activity in question. It is helpful to draw an analogy with individual cases. Think about when one is performing some task and says to oneself, \textit{wait let’s see—this is supposed to go there, not here}, and then moves the object in question to the appropriate place. The criticism as made by the agent herself (at least as she’s performing the task) has a practical and executive force that is lacking when it is put forth by someone else. The criticism is a part of what it is for her to carry out the task. It’s true that the content of her own criticism is presumably available to others as a part of their criticism of what the agent is doing. Someone else looking at the agent might also judge that the object belongs in one place rather than another and that she ought to move it. But their criticism is just a judgment about the agent’s performance—perhaps a part of some evaluation of her, or at best some sort of advice regarding how she should proceed. It is not a part of what she is doing.\footnote{Roth 2004, p. 402n, emphasis original.}

Roth’s contralateral commitments are an extension and clarification of Gilbert’s directed rights and obligations: they clarify both the nature of the normative commitment that holds between participants in a collective action, and the special standing that participants in a collective action have with respect to the attendant rights and obligations.

\textbf{The alleged failures of the reductive theorist}

As we saw, obligations to conform to collective intentions carry along with them rights that are possessed by the other members of the group; X’s obligation to perform action $x$ (where $x$ is necessary for the success of the group goal) entails Y’s right to X’s action. Collective action thus necessarily has the effect of binding agents together in a web of interconnected rights and obligations.

Reductionist theories cannot account for this interconnected web, Gilbert argues, precisely because reductive theories dictate that all normative obligations must be generated by individually held psychological attitudes. Individually held psychological attitudes must be capable of accounting for both an agent X’s commitment to the group’s goals, and also for the rights that the other agents have to X’s actions. Reductionist theories might be able to account for the former, Gilbert argues, but they cannot account for the latter.

Individual psychological attitudes can impose normative commitments on the actions that an agent performs, as well as provide normative structure to an agent’s future deliberation. Insofar as these individual psychological attitudes are held individually by an agent,
though, they only generate commitments towards the individual agent. Suppose X forms the intention to buy Y a present. This intention might commit X to certain actions, or certain standards of deliberation; X is committed to actually buying the present, and X is committed to not spending the money reserved for the present at the horseracing track. X’s mere possession of the intention to buy Y a present does not grant Y the right to the present, however. The commitment which is generated by X’s individual commitment does not automatically entail a corresponding right in Y. Or, to put it another way which Gilbert would also favor, if X fails to follow through on her intention to buy Y a present, Y could not reasonably argue that he was owed the present. Y might be upset, of course, and Y also might chastise X for failing to follow through on her intention (in the same way that Y might chastise X for failing in her New Year’s resolutions). However, Y’s complaints cannot be about what he is owed, since X’s intention to buy Y a present does not entail that Y is owed anything at all.

Thus, Gilbert argues, reductive theories cannot account for the interconnectedness between obligations to act and rights, and reductive theories cannot explain how these rights might be generated. As a result, the reductive theorist cannot account for directed rights or obligations, or for contralateral commitments.

Now, the reductive theorist might argue that these normative commitments are moral commitments, and they can therefore be accounted for through moral principles. In response, plural subject theorists argue that ethical norms are not directed norms, and so they cannot account for the directed norms that hold between participants in a collective action.

To show that ethical norms cannot account for directed rights and obligations or contralateral commitments, Gilbert and Roth each consider the moral principles of Thomas Scanlon and his ‘Principle F’ (for ‘fidelity’). Principle F is one way of accounting for promissory obligations. Gilbert and Roth focus their attention on Scanlon’s Principle F because it presumes to be able to account for a certain kind of directed norm: the norms that hold between two agents when one agent has led the other to expect certain behaviors. Principle F entails that one agent has a directed obligation to another agent in virtue of intentionally leading another agent to expect certain behavior; this directed obligation binds the two agents together in ways which are not relevant for parties outside of the creation of expectations. Scanlon’s Principle F is as follows:

Principle F: If (1) A voluntarily and intentionally leads B to expect that A
will do X (unless B consents to A’s not doing so); (2) A knows that B wants to be assured of this; (3) A acts with the aim of providing this assurance, and has good reason to believe that he or she has done so; (4) B knows that A has the beliefs and intentions just described; (5) A intends for B to know this, and knows that B does know it; and (6) B knows that A has this knowledge and intent; then, in the absence of special justification, A must do X unless B consents to X’s not being done.54

Gilbert and Roth focus on the conclusion of Principle F: A must do X unless B consents to X’s not being done. This puts B in a unique position; A essentially owes B the behavior, unless B consents to the behavior not being done.

So why isn’t Principle F sufficient to account for the directed rights and obligations and contralateral commitments characteristic of collective action? Roth presents two major worries about the reductive theorist’s doing so. First, Roth argues, Principle F only generates its norms when there is an intentional creation of expectations. However, Roth argues, it is unreasonable to think that in every instance of collective action, there will always have been the intentional creation of expectations.

Second, according to Roth, Principle F is much too concerned with assurances. Point (2) in Principle F states that B wants to be assured of something (and A knows this), while point (3) states that A acts so as to provide the assurance. However, Roth argues, this demand for assurance is frequently out of place in cases of collective action. The agents who are likely to work together towards a collective goal frequently do not need to give one another assurances, Roth argues, because they might trust each other enough that assurances are not necessary. Two friends who are going for a walk together need not have given one another promises that they would continue walking until the end, nor would they think that any such assurances would need to be given.

Because Roth thinks that Principle F fails to adequately capture the contralateral commitments of collective action, he concludes that contralateral commitments are not likely to be captured by moral principles, and instead investigates whether contralateral commitments might be explicable through individual psychological attitudes such as intentions. Roth again concludes that contralateral commitments cannot be captured through individual psychological attitudes, and thus that they cannot be accounted for through a reductive theory of agency.

Gilbert’s argument against Principle F is more nuanced. Gilbert directly argues that

3.5. PLURAL SUBJECT THEORIES

Principle F cannot account for directed norms (and, in particular, for the directed rights of the promisee). According to Gilbert’s argument, if Anne leads Ben to expect her to call him (in full accordance with Principle F), and Anne fails to call Ben, then Ben ought to have a special right to rebuke Anne. However, Gilbert argues, this special right could only come from one of four places: (a) Anne’s violation of a moral principle; (b) Principle F’s consent clause; (c) Ben’s ‘right to rely’ on Anne’s performance; or (d) Ben’s being the intended beneficiary of Principle F. None of these, however, can account for the directed right Anne has to rebuke Ben. We will not give a full account of Gilbert’s argument against these four cases here; instead, we will present a broad overview.

Most substantively, (a) fails because moral violations can be criticized by everyone, and not just the directly affected parties.

Gilbert argues that (b) fails for similar reasons: the fact that Ben’s consent is at issue does not, in and of itself, entail that Ben has a special standing to rebuke Anne. After all, there might be another person Timmy whose consent is a condition of Anne’s promise to Ben; in such cases, Anne’s failing to secure Timmy’s consent does not mean that Timmy has a special right to rebuke Anne, since Anne’s promise was to Ben and so Ben should have the special right.

Against (c), Gilbert argues that there are two ways to understand what it would mean for Ben to have a right to rely on Anne’s calling him. The first way is as a right to a performance: Ben has a right to rely on Anne’s calling him so long as Ben has a right to Anne’s actually calling him. The second way is an epistemic reliance: Ben has a right to rely on Anne’s calling him so long as Ben is epistemically justified in believing that Anne will call him. Gilbert suggests that Principle F guarantees epistemic reliance, but she suggests that it does not entail that any agent has a right to a particular performance.

Finally, Gilbert rejects (d) on the grounds that an agent might have a duty which is grounded in the interest of the rights holder, but—as in case (a)—that duty is a universal, directionless duty; it does not generate any special obligations towards the rights holder, nor is the rights holder in a unique position to complain.

We will assess these arguments below. As we will see, neither Gilbert’s arguments nor Roth’s arguments against moral accounts of directed obligations succeeds, in large part because they argue primarily against Scanlon’s Principle F and not against the capacity of moral theories in general to account for directed rights and obligations. However, let us for the moment accept the negative arguments offered by Gilbert and Roth against reductive
theories of collective action. What theory of collective action should we accept instead?

3.5.2 Positive plural subject theories of collective action

Gilbert and Roth both argue that we should accept a theory of collective action which is not reductive. The details of their anti-reductive theories differ. Gilbert argues that the agent of collective actions is, properly speaking, a plural subject; her theory is therefore dedicated to explaining what a plural subject is, how it comes about, and what membership in a plural subject means for the constituent agents. Roth, on the other hand, argues in favor of an anti-reductive theory in which agents are capable, under certain circumstances, of acting directly upon the mental states of other agents. In this section we look at their respective positive theories of collective action.

Gilbert

According to Gilbert’s theory of collective action, the fundamental concept is that of a joint commitment: joint commitments underlie collective intentions, and collective intentions underlie collective action. We first cite Gilbert’s definition of collective action:

\[ \text{Persons } X \text{ and } Y \text{ are collectively doing } A \text{ if and only if they collectively intend to do } A \ldots, \text{ and each is effectively acting, in light of the associated joint commitment, so as to bring about fulfillment of this intention.}^{55} \]

Collective action depends on two things: first, a collective intention, and second, a joint commitment. Ultimately, joint commitment is the fundamental concept, as joint intention is defined in terms of joint commitment:

\[ \text{Persons } X \text{ and } Y \text{ collectively intend to perform action } A \text{ (for short, to do } A) \text{ if and only if they are jointly committed to intend as a body to do } A.^{56} \]

What, then, is a joint commitment? Agents who have a joint commitment to do some action \( j \) are jointly committed to doing \( j \) ‘as a body’. The commitment to doing \( j \) as a body requires the agents to emulate a single agent when doing \( j \). The joint commitment is typically created, Gilbert argues, through each agent’s expression of readiness to be jointly committed to doing \( j \). Joint commitments bind each agent by giving each agent sufficient

---

55 Gilbert 2006, p. 12. Gilbert’s original names of the agents have been changed from \( A \) and \( B \) to \( X \) and \( Y \) for the sake of clarity.

56 Gilbert 2006, p. 11, emphasis original.
3.5. PLURAL SUBJECT THEORIES

reason to act in accordance with the commitment. And, just as joint commitments require the contribution of each agent to come into being, they also require the contribution of each agent to be removed. For Gilbert, it is this joint commitment which accounts for the directed rights and obligations that participants in a collective action hold towards one another.

Agents who hold such joint commitments—agents who are jointly committed to doing some action \( j \) as a body—constitute a plural subject. And it is plural subjects that engage in collective action; collective actions are not performed by sets of individuals. We needn’t focus on the ontological status of plural subjects; all we need say is that plural subjects exist so long as the requisite joint commitments hold between the participants. According to Gilbert, plural subjects are not ontologically reducible to individual agents. Regardless of the possibilities for metaphysical reduction, it is clear that, according to Gilbert, the reductive theorist cannot account for these joint commitments because reductive theorists can only appeal to mental states held by individuals, and (as we saw above) these mental states are insufficient to account for these joint commitments.

Roth

For Roth, collective action is made possible when practical intimacy exists between the participants of a collective action. Practical intimacy holds that

It is possible for one individual to take up and to act on an intention formed by another without re-issuing the latter’s intention.\(^{57}\)

According to Roth, if agent \( X \) is aware of agent \( Y \)’s intention that \( X \) perform some action \( x \), and if practical intimacy exists between \( X \) and \( Y \), then \( X \) can act directly on \( Y \)’s intention that she \( x \)—\( X \) does not need to ‘re-issue’ \( Y \)’s intention so as to become a new intention, possessed by \( X \) herself, that she \( x \).

Now, Roth acknowledges that if \( X \) has successfully done \( x \), then she must have done so intentionally. It would be odd, then, for \( X \) to intentionally \( x \) without possessing any intention to \( x \) whatsoever.\(^{58}\) \( X \) does have an intention to \( x \), Roth argues—but \( X \) comes to have the intention to \( x \) through preserving and executing \( Y \)’s intention. This means that \( X \)

\(^{57}\)Roth 2004, p. 383.

\(^{58}\)Roth ignores here Bratman’s rejection of this very same principle, which Bratman terms the Simple View, in Bratman 1984. Roth’s argument would still hold, however, if he accepted the weaker Single Phenomenon view that Bratman favors.
engaged in no deliberation about whether to adopt the intention to \( x \): she did not consider reasons for or against forming the intention to \( x \). \( Y \)’s intention that \( X \) do \( x \) is not one reason in favor of \( X \)’s doing \( x \), a reason which is commensurable with other reasons. Nor is \( Y \)’s intention that \( X \) do \( x \) merely a trumping reason—one which automatically takes priority over \( X \)’s other reasons for or against doing \( x \). This is because if \( Y \)’s intention that \( X \) do \( x \) were to be a reason, even a trumping reason, then \( X \) would be engaging in deliberation upon those reasons and re-issuing \( Y \)’s intention. If practical intimacy holds, then \( X \) simply does not engage in the reasoning game with respect to \( Y \)’s intention—\( Y \)’s intention has the same practical authority with respect to \( X \)’s subsequent deliberation and behavior that \( X \)’s own intentions do. However, there is an important difference between \( X \)’s intentions and \( Y \)’s intentions towards \( X \)’s behavior: if practical intimacy holds and \( X \) acts directly upon \( Y \)’s intention, then \( Y \) remains the authority over the intention. That is, \( Y \) has the authority to correct \( X \)’s behavior, to clarify the intention when necessary, and so forth. \( X \) has no such authority over \( Y \)’s intentions, even when practical intimacy holds.

Practical intimacy, Roth argues, grants us the resources necessary to account for contralateral commitments. Roth argues that one agent has a contralateral commitment to another agent if and only if the following conditions are met:

1. \( B \) has the authority to settle what it is that \( A \) will do with respect to the matter of whether to \( F \), and \( A \) has the corresponding entitlements to \( B \)’s intention as a practical conclusion. Moreover, this authority is exercised, and \( B \) intends for \( A \) to \( F \);
2. this intention is communicated to \( A \);
3. \( B \)’s authority is not blocked, superseded, suspended, or nullified, so that \( A \) is in a position to act directly on the intention for him to \( F \), without re-issuing it.\(^{59}\)

A proper analysis of contralateral commitments, and therefore of collective action, depends on practical intimacy holding between two (or more) agents. But why should one agent’s acting directly on the intention of another agent account for contralateral commitments between them? According to Roth, when an agent \( X \) forms an intention to \( x \), then \( X \) has thereby generated a set of practical commitments to herself; Roth calls these commitments *ipsilateral commitments*. These ipsilateral commitments are the commitments typical of intentions: deliberative commitments, planning commitments, rationality commitments on

\(^{59}\)Roth 2004, p. 391.
possessing conflicting intentions, and so forth. X’s intention to x generates in her practical ipsilateral commitments to take those means necessary to x. Similarly, Roth argues, in cases of practical intimacy, where Y has an intention x that X acts directly upon, this intention generates commitments for X—but because Y possesses the intention, the intention generates practical contralateral commitments, commitments that X has towards the owner of the intention (Y).

Because Roth appeals only to mental states possessed by individuals—beliefs, desires, and intentions—we might think that Roth’s theory is a reductive one. Roth’s theory of collective action is best classified as a plural subject theory, however. The reductive theorist has to hold that agents are motivated to act on the basis of mental states that are possessed by the deliberating agent herself or himself. Practical intimacy denies this, however. As a result, the barrier between agents is dissolved; it is no longer the case that we can analyze collective action purely in terms of individuals acting upon individually held attitudes. As mentioned above, this does not entail that Roth is endorsing the same plural subject theory that Gilbert endorses. Roth does reject reductive theories of action, however, and so his theory is sufficiently similar to Gilbert’s to be grouped in the same category.

3.5.3 Against Gilbert and Roth

As mentioned earlier, plural subject theories face a general problem: if reductive theories of action are capable of accounting for collective action, then plural subject theories are unnecessarily redundant; Occam’s razor directs us to accept the reductive theory. It is thus of crucial importance for plural subject theories that they provide arguments against the capability of any reductive theory to account for collective action. If these arguments against reductive theories fail, then much of the support for a plural subject theory of collective action has dissipated. (We still have to come up with a satisfactory reductive theory of collective action, of course; it does us no good to defend the possibility of a reductive theory if we do not have one available.)

In fact, we have reason to doubt the claim made by Gilbert and Roth that collective actions necessarily feature normative commitments that cannot be accounted for by the reductive theorist. The claim is twofold: first, the plural subject theorist claims that all collective actions feature these special normative commitments; and second, the plural subject theorist must argue that reductive theories cannot account for these commitments.

Should we accept these two claims? Let us, for the sake of discussion, accept the
first: every (normal) instance of collective action features normative commitments that hold between the agents, whether we call these commitments directed rights and obligations, or whether we call them contralateral commitments. (We will argue further below that we should not accept even this premise.) The question then becomes whether the reductive theorist can account for such commitments; I want to argue here that the reductive theorist can account for them.

There are two ways in which the reductive theorist can account for directional obligations of one agent to another. The first way occurs when directional obligations are deliberately created between agents. The second way occurs through the harm principle: agents who are uniquely harmed by an agent’s actions have a special standing to complain that agents who are not uniquely harmed do not have.

As for the first way, we have reason to believe that a great many cases of collective action feature obligations that were deliberately created by the participating agents. Consider a standard Gilbertian account of how plural subjects come into being:

In discussing the creation and demise of joint commitments, I focus on the basic case, where there are no special background understandings in play. In order to create a given joint commitment of the basic kind, each party must make clear to the others that he is ready to be jointly committed in the relevant way. Thus each must express a certain condition of his will. Once these matching states of the will are mutually expressed, and this is common knowledge between the parties, the joint commitment has been created.

To make this more concrete, here are some situations of the kind I have in mind. First example: Jessica says, ‘Shall we meet at six?’ and Joe says, ‘Sure’. I have argued elsewhere that this exchange is best parsed somewhat as follows. Each party expresses to the other his or her readiness to be jointly committed to endorse as a body the decision that they meet at six. These expressions are common knowledge between the parties. This interchange constitutes what would ordinarily be termed ‘an agreement’ in English—an informal agreement, not a legal contract.60

It is important for Gilbert, in the standard story of the creation of plural subjects, that each agent publicly states her or his willingness to participate in the collective action. If this expression of one’s willingness to participate takes the form of a promise, or (more likely) of an assurance of one’s action (whether the action is outright or conditional on the other’s action), then the agents have created an obligation to act. Suppose, for instance,

---

60Gilbert 2006, p. 9.
that X says to Y ‘I will x if you y.’ Suppose that Y responds, ‘ok, I will y if you will x.’ X and Y will have thereby created obligations upon themselves to act as they say they will—obligations which are directed towards the other agent (since the obligation was created between them). If X ends up doing x, but Y fails to y, then X will be perfectly entitled to complain that Y failed to live up to his obligations.

There is a reason why such public declarations of one’s intentions might be so frequently necessary in collective action cases. Many instances of collective action can be modeled roughly as coordination games: outcomes in which all agents coordinate are generally preferred over outcomes in which the agents fail to coordinate. Further, agents who perform cooperative actions, but fail to have their cooperative actions reciprocated, incur additional costs. It is thus crucial for all that the agents find a solution to the coordination problem. Language is an excellent way to solve the coordination problem: if the agents each attest their willingness to perform the collective action, and thereby commit themselves to performing the collective action, then the agents can more easily solve the collective action problem.

Normative commitments have an additional force that mere words do not have. If X forms a commitment to Y to do action x, then she has thereby given herself greater reason to do x than she would have had if she had not said anything at all, or if she had merely uttered words but incurred no normative commitment. If agents are capable of creating normative commitment for themselves, and if those normative commitments give agents greater reason to perform some action, then forming such normative commitments is a good way of solving coordination problems.

Why should we think that agents can deliberately create commitments to one another? Such creations of commitments happen all the time, and they happen outside of the context of collective actions. X and Y might create commitments to one another through the act of marrying; they create their commitments by saying ‘I do’ because of their mutual belief that the institution of marriage involves such commitments to one another. (In a society in which marriage entailed no commitments at all, X and Y’s marriage would not create such commitments.) Equally, X might create a commitment to Y by promising him that she will do x; the act of promising creates such commitments, again at least in part because X and Y believe promising to create such commitments.

In other words, if Gilbert insists that plural subjects are frequently formed through the kinds of explicit verbal declarations of willingness to participate in a collective action, it is
easy for the reductive theorist to argue that these verbal declarations of willingness are actually creations of commitments between one agent and another. These verbal declarations of willingness are often ways of ensuring coordination in coordination problems, and the declarations work by committing agents to perform some action, either outright (as when \(X\) says to \(Y\) ‘I will do \(x\)’) or conditionally (as when \(X\) says ‘I will do \(x\) if you will do \(y\)’).

And what of Gilbert’s complaint that we can only understand the directional commitments that agents create towards one another if we invoke plural subjects? Against her complaint we should take note of two important points. First, Gilbert’s claim that we need plural subject theory to account for promising is surely too strong. Plural subject theory requires symmetry between the participants; if one agent \(X\) expresses a willingness to be jointly committed to an activity, the other agent \(Y\) must reciprocate. Promising, however, can create directional obligations unilaterally; if \(X\) unilaterally commits to \(Y\) to do some action \(x\) by making a promise, then \(X\) thereby owes \(Y\) that action, even if \(Y\) has not done anything to accept \(X\)’s promise. Whatever theory we give about the nature of moral commitments, such a theory must be able to accommodate the fact that directional commitments can sometimes be created unilaterally. Plural subject theory does not appear able to do so, at least not easily.

The second reason we should be skeptical about Gilbert’s claim that the reductive theorist cannot account for directional commitments is that agents can have special standing to complain based upon the harm produced by some other agent’s actions. If \(X\) forms the intention to dispose of toxic waste in \(Y\)’s backyard, \(Y\) has a special standing to complain about \(X\)’s intention that some uninvolved party \(Z\) does not have (though \(Z\) might well complain, insofar as \(X\)’s intention is morally objectionable). \(Y\) has a special standing to complain because \(X\)’s intention adversely affects him in ways that it does not affect \(Z\). Though \(X\) is fairly criticized by both \(Y\) and \(Z\) over her intention to dump toxic waste in \(Y\)’s backyard, \(Y\)’s complaint surely has a special standing that \(Z\)’s does not.

In collective actions the parties to the collective action suffer when one party fails to participate in ways that non-participants do not. In typical coordination games, failures of coordination affect those who participate more adversely than it does those who do not participate, due to the costs of participation. If \(X\) verbally leads \(Y\) to believe that she will do \(x\), where \(x\) leads to an outcome which is a resolution of a collective action problem, and if \(Y\) does his part in the collective action, and \(X\) fails to do \(x\), then \(Y\) will suffer harm that some non-participant \(Z\) will not. As a result, \(Y\) has a special standing to complain about
X’s failing to do her intended action that Z does not have.

So Gilbert’s claim that directional commitments can only be accounted for by plural subject theory seems too strong. Roth’s argument, on the other hand, is that Scanlon’s Principle F is inappropriate to account for the contralateral commitments that hold between participants in a collective action. This argument is moot, however. We could accept Roth on this point while still rejecting his claim that no moral-based principle can account for the directional commitments of collective action. Indeed, Roth seems to admit that something like Principle F can generate directional commitments; as he says,

What’s interesting for our purposes is that the element of consent [in Principle F] places B in a special position with respect to A’s obligation. While anyone, including third parties, is in a position at least in principle to criticize A for not living up to his obligation, only B is in a position to release A from the obligation.\footnote{Roth 2004, p. 367.}

Now, Roth does go on to argue that Principle F cannot account for contralateral commitments. But Roth’s argument should not be directed against Principle F specifically; his argument should be directed against every normative principle which is available to the reductive theorist. Once we allow that some normative principle can account for directional obligations, as Roth seems to allow for Principle F, then the plural subject’s argument against the reductive theorist is substantially weakened. Even if Principle F is not the best way of accounting for contralateral commitments, we should have faith that some normative principle, appropriately formulated, can account for such commitments.

This is, of course, if we accept that collective actions always entail directed obligations which hold between the participants. But we should not accept this; I think that we have reason to doubt this claim. Michael Bratman has described a case in which two singers agree to collectively sing a song, while explicitly noting that they each do so with no commitments towards the other.\footnote{Bratman 1999d.} Gilbert has argued in response that the fact that such obligations need to be explicitly dismissed supports the claim that such obligations are present in collective actions by default.\footnote{See Gilbert 2000.} Gilbert’s response misses the main point, however. As we noted above, the creation of a normative commitment to act is a particularly effective solution to coordination problems. It is not the only way, however. What agents in a coordination problem need is assurance that the other agent will act in the way requisite to bring about
the coordinative outcome. The creation of a normative commitment can do this. However, the mere pronouncement that one will *likely* act in a certain way can also provide such reassurance. If \( X \) says to \( Y \), ‘I am telling you that I will do \( x \), but I am not creating any commitment to you at all to do \( x \),’ this can provide the necessary assurance to \( Y \) that \( X \) will do \( x \), even if \( X \) does not thereby incur any obligations. The language used to express one’s intended action is also language which typically creates obligations, at least in the types of situations typical of collective choice scenarios. As such, if agents insist on there being no such commitments, they might need to make the lack of commitments explicit.

However, these commitments are not necessary for collective action, and this fact is clearer when agents are in a collective choice situation in which their expressing their willingness to engage in the collective action does *not* generate such commitments towards the others. When people gather in parks to play pick-up games of soccer or basketball, people might drift in and out of the activity at their own discretion; this does not mean that the agents are not committed to the activity while they are participating. The same holds for musicians in drum circles, or any other such improvisational activity in which participation is fluid. People can join in the pick-up game, or the drum circle, without any explicit creation of commitments to act in certain ways or to participate in the activity for a certain amount of time; similarly, people can leave the pick-up game or drum circle without any explicit termination of the collective action.

Now, it’s not clear to me how Gilbert would analyze the nature of the directional commitments that hold between participants in a pick-up game, or a drum circle. Invoking plural subjects seems to be the wrong way to go, however. The problem is that the lack of certain obligations to one another are entailed by the collective understanding of the activity in question. Furthermore, the obligations which *are* entailed by the activity might well change depending on the social conventions of the local community. We can imagine ‘local rules’ for pick-up games such that in one park, participants must clearly indicate when they will be leaving, while in another park, agents have no such requirement. Someone who participates in a pick-up game at a foreign park might well be admonished by the participants, as they explain that ‘that’s not how we do it here’.

The point is not merely that we can imagine situations in which agents do not have to ask permission to leave the activity. The broader point that I am making here is that the nature of the obligations which hold between participants in a collective action can be determined by social conventions just as easily as it can be by explicit creations of obligations. In both
cases, we should be reluctant to claim that the obligations are created by the collective activity itself—and since this is what Gilbert’s plural subject analysis requires, we should be accordingly suspicious of Gilbert’s theory.

Moreover, in rejecting Gilbert’s analysis of activities such as drum circles and pick-up games, we should not find succor in Roth’s analysis, which holds that the commitments which hold between participants in a collective action are so strong, and of such a unique nature, as to grant one agent authority over another agent’s action, authority which bypasses the agent’s ordinary deliberative faculties. It is true that the reductive theorist has a difficult time accounting for such practical intimacy. However, even if practical intimacy does sometimes hold between agents—a fact which is by no means obvious—practical intimacy is far too strong a condition to require for collective action.

The problem with practical intimacy is that it grants one agent authority over the other agent in the collective activity. Within the scope of the collective activity, this authority is total: it bypasses the receiving agent’s deliberative and rational faculties, and so the agent must act directly upon the intention without rationally endorsing the intention. If X says to Y, ‘I intend that you y’, Y must take X’s intention as his own for the purposes of further deliberation and action.

But why should Y treat X’s intention that he y as necessarily authoritative in the context of a collective action? There are several reasons why Y might properly subject X’s intention to scrutiny by his rational and deliberative faculties. One would be if there were questions about whether Y’s doing y would be the best means available to bringing about some outcome o, where o represents the group’s goal. In other words, Y might be concerned about whether X is an efficient practical reasoner. Another would be if Y were not sure whether X’s intention that he y is relevant for the collective action or not.

Roth’s problems run deeper than this, however. The two above examples suggest that, even if agents often act directly upon the intentions of others without re-issuing them or mediating them by their own rational faculties, there are questions about when such intentions are legitimate ones to act upon, and these questions can only be settled by rational deliberation. The more serious worry, however, is that practical intimacy is simply not the right kind of relationship that should hold between participants in a collective action—that practical intimacy is indistinguishable from other cases of authority which do not entail, or are incompatible with, collective action.

Consider, for instance, a case in which one agent, X, has complete authority over Y’s
behavior, and consequently directs his behavior towards the production of some outcome \( o \). Further, we can suppose that Roth’s three conditions for contralateral commitments are satisfied: \( X \) has the authority to settle what \( Y \) does with respect to those actions \( y \) which produce \( o \), and \( Y \) is entitled to \( X \)’s intention as a practical conclusion; \( X \) communicates her intention to \( Y \); and, lastly, \( X \)’s intention is not blocked superseded, suspended, or nullified.

We cannot conclude that \( X \) and \( Y \) are therefore engaged in a collective action. Suppose \( X \) has a psychological ascendency over \( Y \), such that \( X \) has the (practical) authority to determine what \( Y \) does. When \( X \) forms an intention for \( Y \) to act upon, \( Y \) does it. But suppose further that \( Y \) has engaged in no deliberation over whether he ought to contribute to the production of \( o \); indeed, we can suppose that \( Y \) has no reason to produce \( o \), and indeed might well have overwhelming reason to not produce \( o \). These situations are are entirely compatible with Roth’s three conditions for contralateral commitments. And yet, in this circumstance, we should be hesitant to say that \( X \) and \( Y \) are engaged in a collective action.

As an extreme example, suppose that \( X \) forms the intention that \( Y \) throw himself in front of a moving train. \( Y \) might, prior to \( X \)’s formation of her intention, plead with \( X \) to not form such an intention; \( Y \) recognizes that \( X \) has the authority to determine what \( Y \) does. \( Y \) does not desire such an outcome. \( Y \)’s desiring that the outcome not be brought about is compatible with Roth’s conditions for contralateral commitments, however.

Nor is the problem fixed if we insist that both \( X \) and \( Y \) have the authority to form intentions for the other. \( X \) and \( Y \) could both have psychological ascendency over the other, and they might be engaged in a kind of showdown: each are attempting to form intentions for the other first, so as to direct the actions of the other towards her or his chosen goal for the group. What Roth leaves out of his account is some endorsement of the collective goal towards which the actions are directed.

Could Roth remedy his account by adding in a condition simply requiring that each agent endorse the collective goal? He cannot, I think, and still preserve the spirit of his theory. The problem is that it is meaningless to endorse a collective goal if the acceptance of that goal does not structure one’s deliberation or action in any further way. Suppose \( X \) and \( Y \) both endorse the pursuit of the collective goal \( o_1 \). If \( X \) has authority over \( Y \)’s actions, then \( X \) can still form an intention that \( Y \) should \( y \). Now, suppose \( Y \)’s \( y \)-ing leads not towards \( o_1 \), but rather towards \( o_2 \). \( Y \)’s endorsement of pursuing \( o_1 \) still doesn’t enable \( Y \) to determine whether \( y \)-ing is an appropriate action or not, since \( Y \) cannot use
his endorsement of \( o_1 \) to affect his judgment of whether \( X \)'s authority is appropriate with respect to his \( y \)-ing or not.

I think there is a reason why Roth is inclined towards a stronger view of authority than is necessary. Roth considers a possibility that two agents engaged in a collective activity have an intention to coordinate their intentions together; he calls such a ‘meta-intention’ a *bridge intention*. Roth rejects the bridge intention proposal, on the following grounds:

Take any two-person case . . . where one participant (you) forms an intention that conflicts with some activity-related intention of another (me), and that this new, conflicting intention of yours is a result of how you’ve filled out, modified, or defeated the overall intention concerning the activity. Thus, whereas the original intention was to drive to Vegas with me, your new, revised intention is to *take a bus* to Vegas with me. Or maybe it’s the intention to go with me to *Reno* instead. Whatever the case, let us suppose that you think that this is a perfectly reasonable way to revise the intention, in light of the circumstances . . . . Your revised intention is not crazy, then, and neither of us thinks that it is.

The bridge intention proposal will not adequately capture the normative demand for coordination that I will face in light of how you revise your intention. Recall that on this proposal, the source of the demand to coordinate with your intention, and hence the source of the status or authority your intentions have in my practical reasoning, is my bridge intention. More specifically, it’s my bridge intention to coordinate with your intentions relevant for filling out and implementing the drive to Vegas together. But now you do not have any such intentions for me to coordinate with. Your intention is now to take a *bus* to Vegas, or perhaps it’s the intention to drive together to *Reno*. My bridge intention has nothing to say about these intentions of yours . . . .

Roth seems to contradict himself here, however. On the one hand, he argues that the new intention—take the intention to go to Reno instead of Vegas—automatically has practical authority for both agents in virtue of the fact that they are engaged in a collective action together. On the other hand, he suggests that the revised intention is authoritative at least in part because ‘[the] revised intention is not crazy . . . and neither of us thinks that it is.’ Roth can’t have it both ways, however. If the revised intention is authoritative because it is viewed as reasonable (or ‘not crazy’) by the participants, then it must have been deliberated upon and endorsed by each agent’s rational faculties. Its authority, in this case, stems from the fact that it is a perfectly reasonable revision in the eyes of each agent. If, however, the revised intention is authoritative in virtue of the agents being engaged in a

---

64Roth 2003, p. 78, emphasis original.
collective action, then the reasonableness of the intention revision is irrelevant for whether or not the revised intention has authority.

Roth seems to recognize this dilemma. He argues that the defender of the bridge proposal might suggest that there is some further goal that going to Vegas satisfies—the intention to gamble, for instance—that makes reasonable the revised intention of going to Reno instead. If this were the case, then the bridge intention could account for the revised intention to go to Vegas.

In reply, Roth states that there need not be any reason to go to Vegas other than the goal of going to Vegas. And yet, he continues to insist that the revised intention to go to Reno is a reasonable intention revision, one that ought to be authoritative for both agents. To quote,

On the contrary, we may have adopted this end in order to make the most of going to Vegas. Or perhaps for us the end of gambling comes packaged with going to Vegas, and that we would not otherwise pursue or intend gambling. Either way, the end of going to Vegas is as it were an end in itself, and not merely a means to gambling. If this is the case—and I stipulate that it is in our example—then my bridge intention (and yours) will be to coordinate with your intentions pertaining to going to Vegas (or going to Vegas to gamble); it will not be the broader intention to coordinate with your intentions pertaining to gambling. If this is right, then the bridge intention proposal will not have the resources to account for the authority or status retained by your revised intention to go to Reno.⁶⁵

Note what Roth has done, however: he has forced himself to first stipulate that the agents form the intention to go to Vegas with the sole intention of going to Vegas, and he then claims that, once engaged in a collective action to go to Vegas (with the sole intention of going to Vegas), if one agent forms the intention that they should instead go to Reno, that intention is authoritative for both.

This is clearly far too strong. Indeed, this seems like as much of a reductio as one could ask for against the claim that collective action necessarily involves the kind of practical intimacy that Roth stipulates. And why stop at Reno? Why not form the intention to go to Mexico, or Australia, or the Moon?

The way to avoid these difficulties is to reject the claim that participants in a collective action have the kind of practical intimacy Roth suggests. We have, then, rejected the substantive arguments behind the positive theories of both Gilbert and Roth. This is not

⁶⁵Roth 2003, pp. 79–80, emphasis added.
to argue that there is no plural subject theory that can account for collective action; it is only to say that the two on offer, those of Gilbert and Roth, do not adequately account for collective action. Further, if our earlier arguments were correct that neither Gilbert nor Roth have offered persuasive arguments against the possibility of a reductive theory of collective action, then we can reject all plural subject theories of action out of hand—so long as we have an acceptable reductive theory on hand. We will proceed to build such a theory in the next chapters.
Chapter 4

Deliberation under constraints

4.1 Introduction

Rational collective action requires a lot of individual rational choices. Collective action is goal-oriented action, where the goal is the goal of a collective and not of an individual. Those many necessary individual rational choices, then, must be individual rational choices on the basis of the collective goal; each agent must make a rational choice, in light of the collective goal, of what to do so as to promote the satisfaction of the collective goal.

My project here is twofold. First, I want to argue that the best way to analyze collective action is as I described it above—as a set of individual agents each rationally choosing to perform some action which is justified in light of its promoting the collective goal. Second, I want to give a description of what such a model of deliberation would look like.

The heart of my theory of collective action is that agents can self-impose constraints upon their own deliberation, that these constraints take the form of intentions and goals, and that these self-imposed constraints do convey some normative force. Collective action, then, occurs when each of a set of agents adopts the same set of constraints upon their deliberation, such that they are committed to bringing about some element of a common set of collectively acceptable outcomes. The model of deliberation is essentially individualistic; each agent engages in individual deliberation. The difference between deliberation on individual goals and deliberation on collective goals, I argue, lies in the nature of the goals.

In this chapter, I argue that any satisfactory model of deliberation must satisfy what I call minimal individualism. Minimal individualism holds that goals (whether collective or individual) are deliberated upon by individual agents. Moreover, we must be able to
represent those goals in the psychology of each deliberating agent.

If this is true, then we need a model of deliberation. I argue for a model of deliberation which has two kinds of pro-attitudes: preferences, which are motivational judgments of desirability, and intentions, which are volitional judgments of desirability. Accordingly, I suggest that there are two interesting kinds of rationality that agents can exhibit: they might be preference-rational, which means that their actions are rationalized by their preferences, or they might be intention-rational, such that their actions are rationalized by their intentions. Intentions, I suggest, constrain an agent’s deliberation and rationalize counter-preferential choices. It is this function of intentions to constrain deliberation which helps to explain collective action, as I argue in Chapter 5. It is the fact that a group of intention-rational agents can constrain their deliberation so as to bring about some outcome (or some of a set of outcomes) that explains what collective action is, and how agents can structure their deliberation so as to produce it.

Given that intentions serve to constrain deliberation, I consider whether any other mental state might achieve the same function. In particular, I look at values, as values seem to be clear instances of mental states rationalizing counter-preferential choices. I argue that, in the context of deliberation over decision problems, values and intentions are functionally identical: values serve as intentions, and vice versa. Therefore, I argue, we can represent both kinds of mental states in our model of deliberation as intentions. Moreover, I suggest, we can be more certain about our claim that intentions provide rational constraints upon deliberation.

In this chapter, then, I attempt to develop a theory of deliberation and constraint such that we can account for individual and collective goals. In the next chapters, I analyze how we can account for individual and collective goals within such a model of deliberation, and I compare my theory of collective action with those that we have seen previously.

4.2 Minimal individualism

We want the account of collective action we give here to be compatible with what we can call minimal individualism. Minimal individualism is not a full-blown methodological individualism; we do not presuppose that our final theory of collective action be ontologically reductive, or that all concepts invoking collectives be ultimately intelligible in terms of
CHAPTER 4. DELIBERATION UNDER CONSTRAINTS

concepts invoking only individuals.\(^1\) Minimal individualism is still a sort of individualism, of course. The motivating idea behind minimal individualism is that when collectives act, they act through the *behaviors* of their constituent members. Further, minimal individualism holds that the individuals who make up a collective for the purpose of collective action must sometimes engage in deliberation about what to do so as to best contribute to the collective action.

These two claims are, or ought to be, uncontroversial. For how else could a collective act, except through the behaviors of its constituents? Individual behavior is not identical to individual action. Nevertheless, the thing that is ‘left over’ when we subtract a behavior from an action (to return to Wittgenstein’s felicitous phrase) is an *intentional state*; it is not another behavior, or is it another action. The same applies to collective behaviors and collective actions. When a collective’s behavior is subtracted from a collective’s action, what is left over is not more behavior; what is left over is a set of intentional states. The behavioral component of a collective action, then, can be divided without remainder among the individuals that constitute the collective, since the behavior of the collective is obviously nothing more than the behavior of the individuals who constitute the collective.

Moreover, there should be no argument about the claim that individuals sometimes engage in deliberation about how best to serve collective goals. Sugden’s footballers’ problem is just one example: *individuals* sometimes deliberate about collective goals, and they sometimes do so *independently* of other individuals.\(^2\)

How do we accommodate these two intuitions? I suggest that we do so through three conditions of minimal individualism. The first condition we call the *behavior compositionality* condition, and it has already been discussed: the behavior of a collective engaged in collective action is composed, without remainder, of the behaviors of the individual agents who jointly constitute the collective.

The second condition has also been discussed; we call it the *distributive deliberation* condition. The claim here is that whatever a collective goal is, it must (at least in principle) be capable of being distributed among the individuals who constitute the collective. I am

\(^1\)We will leave our discussion of methodological individualism somewhat incomplete. It may be that we are not using the term in Weber’s original sense; Weber emphasized that any analysis of collectives must ultimately reduce to individual agents, but he is less clear on what resources we can grant to those individual agents. See Volume I of Weber 1968. I hope that the ensuing discussion of minimal individualism suffices to make the distinction with methodological individualism clear, if in fact there is one.

\(^2\)Assuming, that is, that the footballers in Sugden’s example choose what to do through a process of deliberation; as we have noted already, the agents might choose what to do through non-deliberative means.
not claiming that every individual who constitute the collective must actively engage in
deliberation upon the collective goal; such a claim would be far too strong. It is possible,
for instance, for a collective to hold a collective goal, but where only one member of the
collective—the captain, for instance—deliberates upon the goal and assigns actions to the
other members. Rather, what I am claiming is that insofar as the collective can be said
to hold a goal, then there is a collective goal which is capable of being deliberated upon
by any member of the collective if it is appropriate for the individual agent to do so. This
is not to deny that there is a phenomenon called ‘collective deliberation’, nor is it to deny
that collective deliberation might be incapable of being reduced to a series of individual
deliberations—collective deliberation might be the kind of collective action in which, like
tango dancing, the component parts are incapable of being done individually. Rather, what
I am claiming here is that collective deliberation, however we theorize it, outputs a collective
goal, and that collective goal can be deliberated upon by the individuals who constitute the
collective.

The last claim that we will make is a corollary of the previous claim, and we can call
it the individual possession condition. Whatever a collective goal is, if an individual agent
holds a collective goal, then that collective goal must be instantiated in the individual’s
mental states. Again, we make no claims as to what kind of mental states these must
be, or whether the mental states must be the same kinds of mental states which factor in
deliberation upon individual goals; our claim is more modest here. The claim is merely
that, in order for an individual to deliberate upon a collective goal, that collective goal
must be ‘in the individual’s head’.

Minimal individualism is thus prima facie compatible with a wide range of theories
about collective action, both reductive and non-reductive. I propose here that we take
these conditions seriously, however, and see what a plausible model of deliberation looks
like when we accept minimal individualism. The theory of deliberation that we develop here
must be able to accommodate both individual goals and collective goals. We can proceed
to present the essentials of such a theory.

---

3Cf. Scott Shapiro’s expansion of Bratman’s theory of shared cooperative activities in Shapiro 2002.
4Again, these claims are meant to be unproblematic. I think that even Roth would accept the individual
possession condition, in spite of his theorizing collective action in terms of practical intersubjectivity. For
Roth, when practical intersubjectivity holds between participants in a collective action, then it is not true
that both agents are deliberating about how to best satisfy the collective goal. Further, Roth would presumably
accept that either agent in a collective action could possess the collective goal—there might be a
question of which agent actually is, but that is a separate question.
4.3 A proposed model of deliberation

Giving a theory of collective action compatible with minimal individualism requires that we give a theory in which we analyze collective action in terms of a set of actions performed by a set of agents. For some behavior to count as an action requires that we be able to explain how the behavior has been, or could be, rationalizable by the agent.\(^5\)

An action is rationalizable for an agent so long as we can show how the action is a rational resolution of a decision problem that the agent faces. And, an action’s being a ‘rational resolution’ of a decision problem means that the action is optimally supported by the reasons that the agent possesses at the time of deliberation.\(^6\) Thus, rationalizable actions are actions which are optimal resolutions, according to an agent’s reasons for action, of decision problems faced by that agent.

What is meant by ‘optimal resolution’ will be examined in brief detail below, and in greater detail in Chapter 5. We should note here that an ontologically reductive theory of collective action will necessarily explain how the action of each agent in the collective is rationalizable according to the reasons possessed by the agents at the time of the resolution of the decision problem. This, in turn, requires that we have a theory of deliberation. Our model of deliberation will be a functional one, and so it may not be a faithful portrayal of the exact psychological processes that deliberating agents engage in. It should, however, accurately capture the various inputs and outputs of rational deliberation. Insofar as we are concerned with the rationalizability of actions, it is the output of deliberation that we are most interested in—an action is rationalizable so long as an agent chooses that action as the output of rational deliberation.

\(^5\)We do not require that some behavior already have been rationalized by the agent in order for the behavior to count as an action. Instead, we require only that the agent be capable of rationalizing the behavior, even if the agent has not actively engaged in that process of rationalization. See also Chapter 2 for further discussion of rational action and rationalizability.

\(^6\)The claim that rationalizable actions are optimal from the standpoint of the agent’s reasons goes against the claim, advanced by Simon and others, that some actions can be satisficing—that is, that actions are rational so long as they pass some threshold of rationality, even if the action is not an optimific one. The psychological grounds for advancing satisficing theories of deliberation are sturdier; see e.g. Gigerenzer and Goldstein 1996. The philosophical justifications for normative satisficing theories of rationality are more uncertain, however; it is quite plausible to suppose that rational agents ought to choose optimific actions, even if satisficing actions are ‘good enough’ in some sense. Accordingly, I will disregard satisficing theories of normative rational justification here. Simon’s own presentation of satisficing can be found in Simon 1955 and Simon 1959, while further interesting treatments of satisficing can be found in Byron 2004.
4.3. A PROPOSED MODEL OF DELIBERATION

4.3.1 A normative model of decision theory

The model of deliberation to be developed is intended to explain the rationalizability of actions. As such, it is a normative model rather than a descriptive model; the output of deliberation describes what the deliberating agent ought to do, rather than describing what the agent is going to do (or is going to do, environment permitting).

On the one hand, this gives us another reason to think that our model of deliberation is not a psychologically realistic one—especially if we think that deliberation is a naturalistic psychological process which is causally connected with an agent’s behavior. But it is the other hand which is more interesting: our model’s being normative and not descriptive is what enables us to employ it in a theory of rational choice in the first place. Advocates of rational choice theories are split. Some take rational choice theories to be purely descriptive, and so argue that rational choice theories are valid or not in accordance with whether their predictions about the behavior of agents are in accord with the agents’ observed behavior. For descriptive rational choice theorists, however, the term ‘rational choice’ is a bit of a misnomer; an agent’s choice is rational or irrational strictly depending on whether the choice is consistent or inconsistent with the agent’s previous choices, and with the predictions of the theory itself. In particular, the agent’s choice is not rational or irrational depending on whether the choice is maximally in accordance with the agent’s reasons for acting—after all, every choice is maximally in accordance with the agent’s reasons, according to descriptive rational choice theorists. In other words, the rational agent is the one who chooses in accordance with the predictions of descriptive rational choice theory, and descriptive rational choice makes its predictions without appeal to whether an action is rational or not.

Descriptive theories of rational choice do not care whether an agent makes good choices or bad choices; such theories only care about whether the agent makes predictable choices.

This leaves a descriptive theory of rational choice with precious little to do. Descriptive theories of rational choice might be predictive even if agents never engage in a process of deliberation at all, or are entirely unaware of how they reason or even of what they choose; descriptive theories of rational choice also might be predictive even if agents constantly choose contrary to reason—regardless of whether we understand reason subjectively or objectively. Moreover, descriptive theories of rational choice typically base their predictions

---

7 A classic theoretical defense of descriptive rational choice theory can be found in Friedman 1966. For an overview and a discussion of the philosophical limitations of descriptive rational choice theory, see Sugden 1991.
upon the revealed behavior of agents in prior choice situations—a problematic foundation if we are interested in the additional psychological claim that there is a connection between preferences as psychological states and behavior.\footnote{This point has been made most clearly in Sen 1973.}

And, finally, descriptive models of rational choice fail as theories; they largely fail to accurately predict the actual behavior of agents.\footnote{Kahneman and Tversky 1979. Also: Kahneman 2003.} In response to the deficiencies of traditional descriptive rational choice theories, and by way of accommodating advances in psychological theories of choice and behavior, rational choice theorists have developed theories of descriptive rational choice which allow for systematic ‘irrationalities’—that is, for choice situations in which agents systematically choose contrary to what might be predicted from their revealed preferences. These updated theories of descriptive rational choice, which fall under the general term ‘behavioral economics’, might be more successful in predicting behavior.\footnote{But see Kahneman 2003 for critical views of the psychological foundations of behavioral economics.}

At the same time, they place in starker relief the reasons why descriptive rational choice theories are inappropriate for the purposes of a philosophical theory of action. Behavioral economics as a discipline is at its strongest in pointing out the various ways in which agents are irrational, even by their own lights. To take one example, one well known by casinos, agents typically exhibit biases in judgment when it comes to judgments of probability, and choose actions which do not maximize their preferences.\footnote{See Rabin 1998. This is true even when allowing for the enjoyment that agents receive in gambling; the relevant fact here is that agents frequently and systematically misjudge their probabilities of winning, and this this mistaken judgment of probability factors into their decision to play or not play.} As another example, agents frequently choose contrary to their (stable) preferences because their preferences are not time-consistent; the things that we prefer in the short term are not consistent with the things that we prefer in the long term.\footnote{Ainslie 1975. See also Rabin 1998.} Economists might call such choices ‘impulsive’ choices, or choices ‘inconsistent across time or context’; philosophers might call such choices ‘weak-willed’ choices, or even relegate them into the category of non-choices or non-actions. The philosophical strategy is revealing; by calling such behavior weak-willed, or non-choices, we highlight the extent to which these behaviors are not rationalizable by the performing agent, and are therefore of a different category of behavior than typical actions.

We can leave moot, then, the question of whether a descriptive rational choice theory can be satisfactory as a behavioral science. Our interest lies in understanding actions, and here descriptive rational choice theories are of no help to us. If we want a theory of deliberation
that can be used to understand actions—and, in particular, collective actions—then we must appeal to a normative rational choice theory.

With a normative rational choice theory, the conclusion of deliberation imposes normative constraints on what an agent does, but it does not presume to describe what an agent actually will do. Rational agents will act in accordance with the conclusions of their deliberation; indeed, we can take that to be a definition of what it means to be a rational agent. Normative decision theory aims to identify where the balance of reasons lies; whether agents act in accordance with such a decision theory is a separate question.

4.3.2 Subjective normativity versus objective normativity

There is a bit of equivocation in our use of the term ‘reasons’, however. We can identify two different kinds of reasons: objective reasons and subjective reasons. These terms are, of course, quite vague, and have been used to support numerous distinctions in the philosophy of action and moral philosophy. The way in which we will use the terms ‘objective reasons’ and ‘subjective reasons’ can be seen through contrasting them with two similar but distinct pairs of terms: what Nagel (following Parfit) terms ‘agent-relative reasons’ and ‘agent-neutral reasons’, and what Williams terms ‘internal reasons’ and ‘external reasons’.

Subjective reasons for action are considerations which the performing agent does recognize as providing reasons for action, or would recognize as providing reasons for action if the agent considered the question. Something might be a subjective reason for an agent to act even if the agent does not, in fact, have reason to act; agent X might believe that she has reason to drink the can of paint, and so have a subjective reason to drink the can of paint, even though she has no non-subjective reason to drink the paint. Objective reasons for action, on the other hand, are considerations which provide reasons for an agent to act regardless of whether the agent believes them to provide reasons. An agent might correctly believe of some objective reason that it provides her or him with reason to act; agent X might correctly believe that she has reason to rescue the drowning baby from the pool, and so might have both subjective and objective reason to rescue the baby. Central to the distinction between objective and subjective reasons here is the relationship between the agent’s belief that she has a reason, and whether she does in fact have a reason.

Compare this now to agent-relative and agent-neutral reasons. Nagel defines agent-relative and agent-neutral reasons thusly:

\[\text{13Nagel 1986; Williams 1981.}\]
If a reason can be given a general form which does not include an essential reference to the person who has it, it is an *agent-neutral* reason. For example, if it is a reason for anyone to do or want something that it would reduce the amount of wretchedness in the world, then that is a neutral reason. If on the other hand the general form of a reason does include an essential reference to the person who has it, it is an *agent-relative* reason. For example, if it is a reason for anyone to do or want something that it would be in *his* interest, then that is a relative reason.¹⁴

According to Nagel’s definition, all subjective reasons (as I use the term) are agent-relative reasons, since it is essential to something’s being a subjective reason that it be judged by some agent as providing a reason to act; this appeal to the agent doing the judging suffices to make subjective reasons agent-relative. However, not all agent-relative reasons are necessarily subjective reasons. Suppose there were a principle of rationality entailing as a rational obligation that each promote the welfare of their kin. This would be an agent-relative reason, because it makes an ineliminable reference to the agent; it says of some agent *X* that *X* has reason to promote the welfare of those people who stand in a certain relationship to *X*. If this principle of rationality were true, however, then it would remain true even if agent *X* did not accept its authority. In such a case, *X* has agent-relative reason to promote the welfare of her kin, but *X* has no subjective reason to do so, and so the two concepts of rationality come apart.

The same example shows us that something’s being an agent-neutral reason is not equivalent to its being an objective reason. For Nagel, agent-relative reasons and agent-neutral reasons are mutually exclusive; it is not possible for a reason to be both agent-relative and agent-neutral. In the example above, however, *X* has objective reason to promote the welfare of her kin, and she also has an agent-relative reason to do so. Since *X* has an agent-relative reason to act, she does not have an agent-neutral reason to act, and so something’s being an agent-neutral reason is not equivalent to its being an objective reason.

Williams’s notions of internal and external reasons come closer to my use of subjective and objective reasons. After all, Williams’s account of internal rationality appeals to the elements of an agent’s psychological set, and not, as Nagel does, merely to the fact that the agent is referenced in the reason statement. However, Williams does not appeal to just *any* element of the agent’s psychological set; Williams appeals to those elements of the agent’s psychological set which are *motivational* attitudes. For Williams, then, in order for

---

¹⁴Nagel 1986, pp. 152–153, emphasis original.
an agent to have an internal reason to perform some action, that agent must be motivated in some way to perform that action.

In my characterization of subjective rationality, I make no reference to whether the agent is motivated to perform that which the agent has subjective reason to perform. It is thus consistent with my account of subjective rationality that an agent $X$ might have subjective reason to perform some action $x$, but nevertheless not be motivated at all to do $x$. $X$ might judge that she has reason to call her mother, and yet feel no motivation whatsoever to actually doing it; she might not have any elements in her subjective motivational set which are appropriately related to her calling her mother. Judgments, after all, are not in an agent’s motivational set; that an agent judges something to provide her or him with reason to act does not necessarily entail the presence of a corresponding attitude in the motivational set.

We might, of course, have other reasons for believing that $X$’s holding the judgment that she has reason to $x$ always entails the presence of some element in her motivational set which produces a motivation to $x$. Internalists about reason and motivation all hold such a position, and it is far from a clearly false position to hold. If we accept a principle such that normative reasons for action (of the kind entailed by subjective reasons) entail motivational reasons for action (of the kind entailed by internal reasons), then the distinction between internal/external reasons and subjective/objective reasons collapses. But it is important to note that even if internalism is correct, we can still conclude that subjective reasons and internal reasons are conceptually different; there is no necessary link between subjective judgments of reason and motivation.

Against those who worry that there are no such things as objective reasons, it is significant that we do not have to believe in the possibility of objective reasons in order to maintain that there is a difference between subjective and objective reasons. If the internalist argument is correct, then agent $X$’s possessing a reason to do action $x$ requires that $X$ be motivated to do $x$, and if $X$ is motivated to do $x$ then $X$ must have a subjective reason to do $x$. In other words, according to the internalist, the only reasons that exist are subjective reasons; there are no objective reasons. Whether there are only subjective reasons, or whether there are both subjective and objective reasons, will not be argued here. The fact that we do not currently have any plausible comprehensive theory of objective rationality on offer, however, will be relevant when we consider the various weights that

\footnote{For one version of this argument, see Smith 1994.}
rational concepts ought to have in subjective deliberation. Although my argument does not hinge on this point—my argument would not be rendered invalid even if we did have a plausible comprehensive theory of objective rationality—it will be relevant when we later consider the rational constraints that intentions place upon deliberation.

4.3.3 The rational resolution of subjective decision problems

Given the above, we are in a position to claim that any viable model of deliberation must aim at accounting for subjective rationality and not objective rationality. Agents deliberate over subjective decision problems. A subjective decision problem is a decision problem as conceived of by a deliberating agent. Subjective decision problems consist of choices for the deliberating agent; outcomes (which can be considered as generalized states of affairs); a set of agents, each of who has choices, and whose choices are relevant for the production of the outcome; causal links between the choices of all the agents and outcomes; preferences of each agent over the set of outcomes; and intentions of each agent. I will make the argument in the next chapter that these resources will suffice to model deliberation, both towards individual goals and collective goals; here, I want to claim that these elements ought to be subjective elements—the model ought to conform to each agent’s beliefs—rather than objective elements, conforming to whatever the facts of the matter are. In other words, X’s subjective model of deliberation will consist of the choices that X believes herself to have in the decision problem; it will consist in the outcomes that X believes are possible to be produced; it will consist in those agents that X believes are making choices which are relevant to the production of an outcome; it will consist in X’s beliefs about the links between the agents’ choices and outcomes; and it will consist in X’s beliefs about preferences and intentions—her own, as well as those of the other agents.

Every action which resolves a decision problem has two facets to it: it is an action in and of itself, and it is a resolution of a decision problem. When we evaluate the rationality of such actions, then, we should distinguish between two ways of evaluation: first, we can evaluate the rationality of the action itself, and second, we can evaluate the rationality of the agent’s resolution of her or his decision problem. The distinction is easiest to see in the case of false but justified beliefs: consider an example in which an agent X is deliberating about which of two routes to take in order to arrive at her destination. Historically, Route 1 is extremely congested at the time X is going to leave, while Route 2 is slightly longer but, owing to much less traffic, is usually a shorter trip at that time. X has traveled both
4.3. A PROPOSED MODEL OF DELIBERATION

routes frequently, and her beliefs in the respective route times are well justified. When $X$ deliberates about which route to take, she chooses Route 2. Unknown to her, however, an overturned truck on Route 2 has stopped traffic entirely for two hours, leading Route 2 to be a less desirable route.

Did $X$ choose the right action? We should conclude that $X$ did resolve her decision problem rationally, even though her action was not the most rational action from an objective standpoint. It might be argued that it makes a difference whether the truck accident on Route 2 occurred before or after $X$ decided to take Route 2, but I cannot see how this would be relevant. We might be able to conclude that $X$ has some responsibility for not verifying her beliefs; we might think that, if the truck was overturned prior to forming her intention to take Route 2, there was some way of correcting her beliefs available to her that there would not have been had the truck overturned after she formed the intention. At issue here is the question of what makes a belief sufficiently justified. This is an epistemological question, but when we consider it from the standpoint of practical action, we can simplify the question greatly. In this instance, $X$ is surely justified in her belief that Route 2 will be a more efficient route to take than Route 1, regardless of whether the truck accident has occurred prior to her deliberation or subsequent to it. And, if $X$’s belief is justified, then $X$ is rational in using her belief for the purpose of her deliberation about which route to take. Questions about when a belief is justified, or what steps a deliberating agent ought to take to ensure that her or his beliefs are sufficiently justified, can be ignored for this example.

At the same time, $X$’s decision to take Route 2 is not a rational action, if we consider the rationality of the action itself. The outcome produced by choosing Route 2 is less rational than the outcome produced by choosing Route 1—by $X$’s own lights and, if they exist, by the standards of objective rationality.

There is, then, a distinction between choosing a rational action, and rationally choosing an action. Only one of the two is available to the deliberating agent, however. All deliberating agents want to choose actions which are best supported by the reasons available; this, indeed, is in part what ‘deliberation’ means. But deliberating agents can only resolve decision problems according to the best information they possess at the time, and as such, the best that they can hope to do is rationally choose an action. This fact gives agents an incentive to ensure that their information about the world is correct—that their beliefs are not merely justified, for instance, but also are true—but the fact that an agent’s beliefs are justified is the best guide that an agent has to their being true. If agents want to choose
rational actions, then, the best they can do is rationally choose actions; only the latter can appropriate guide the agent’s actions—and deliberation itself is an action. As such, an agent’s deliberation must aim at rationally choosing an action, rather than choosing a rational action.

This means that agents must deliberate upon the justified beliefs that they have, rather than on those facts of the world of which the agent is unaware. An agent’s justified beliefs constitute her or his subjective decision problem, however—so long as the agent is sufficiently rational to not deliberate upon unjustified beliefs. Again, I am here taking beliefs to be justified or unjustified according to the standards of practical action; these standards are context-dependent, and so an agent’s epistemological duties for the purposes of practical deliberation might be lower than they are in other epistemological contexts. If we assume that rational agents only deliberate upon justified beliefs—or upon one’s *most justified* beliefs, in those instances where action is required and so *some* beliefs must be used—then we can conclude that rational agents will deliberate upon subjective decision problems—namely, upon decision problems which are constituted by their justified beliefs about the available choices, outcomes, and so forth. And, given that agents deliberate over subjective decision problems, and given that our model of deliberation is normative and not objective, our model of deliberation ought to account for when these subjective decision problems are rationally resolved. Our model of deliberation, in other words, attempts to show what action an agent should choose when faced with a subjective decision problem.

There are two worries about such a model which we can address here. The first has already been raised; the first worry is that, in basing our normative theory of deliberation upon the beliefs that an agent *does* have, rather than on the beliefs than an agent *should* have, we are omitting something rather important. This worryer argues that it is hard to see what could be normatively binding about the results of rational deliberation upon unjustified beliefs. The answer to this worry is that we are not ignoring the responsibility that agents have to ensure that their beliefs are sufficiently justified; to the extent that an agent’s beliefs are *not* justified, the agent’s deliberation will necessarily be irrational. Deliberation, like inference drawing, is ‘garbage in, garbage out’; the agent with unjustified beliefs who faces a decision problem has more serious worries than how she or he should resolve that particular decision problem. All agents, then, have a standing rational obligation to hold beliefs which are sufficiently justified for the context of decision-making. However, even for those agents who do *not* have justified beliefs, and who do not have the time to justify
4.3. **A PROPOSED MODEL OF DELIBERATION**

their beliefs prior to resolving a decision problem, the best they can do is to resolve the subjective decision problem they face as best as they can.

The second worry is that, by modeling the decision problem *that the agent herself or himself faces*, we run the danger of importing into our model any deficiencies of reasoning that the agent might possess. Suppose our agent holds false beliefs, as for instance about which action is rational to take given the subjective decision problem that the agent faces. Suppose, in other words, that agent $X$ is *essentially irrational*—she is incapable of rationally resolving a subjective decision problem. We should worry that *if* our model of subjective deliberation were sufficiently subjective, then our model would also capture $X$’s essential irrationality—it would capture the fact that $X$ falsely judges it rational on the basis of her (sufficiently justified) beliefs to perform irrational actions—actions that rational agents facing the same subjective decision problem would recognize as irrational.

Our model does not build in that degree of subjectivity, however, and so we need not worry that our model recommends that essentially irrational agents choose irrationally. Essentially irrational agents face the same subjective decision problem as potentially rational agents do; their subjective decision problems consist of choices, outcomes, agents, links between choices and outcomes, and so forth. The essentially irrational agent is incapable of recognizing which choice maximizes the value of the outcome produced, but this inability is not represented anywhere in the subjective decision problem itself. So long as the essentially irrational agent holds correct beliefs about the subjective decision problem, then any subsequent errors in deliberation cannot be normatively justified through any beliefs that the agent possesses.

These worries having been addressed, then, we should conclude that if our purpose is to build a model of deliberation, we ought to build that model of deliberation upon subjective decision problems. It might, however, be argued that we have not considered any other options. We will consider below whether we should use objective decision problems rather than subjective decision problems in our model of deliberation; as we will see, we should not.

4.3.4 **Why not objective decision problems?**

It still might be argued that the best model of deliberation tracks the normative obligations that come about through the *objective decision problems* that agents face. Objective decision problems are decision problems that agents face which are independent of the
agent’s beliefs. Objective decision problems include all of the elements that subjective decision problems do; they include choices for each agent, and outcomes that are producible through the choices of the contributing agents, and preferences and intentions, and links between choices and outcomes. For objective decision problems, however, these elements correspond to the way that the world is, rather than to the beliefs that agents have.

Advocates of using objective decision problems to model decision problems, a group that includes most traditional rational choice theorists, generally argue that the only true norms on behavior come from objective decision problems, and so we should focus on those. If there is an action available to an agent which the agent is not aware of, but which is superior to all actions known to the agent, then it is rational for the agent to choose that action. Looking at the objective decision problem can explain why it is more rational for the agent to choose the unknown action.

Moreover, there exist theoretical perspectives in which we should clearly be modeling an agent’s objective decision problem. From the standpoint of evolutionary game theory, for instance, we must look at objective decision problems. The reason here is that, assuming a large enough population size, every available action will be tried by some agent in the population; those actions which are more rational will be tried more, and those actions which are less rational will be tried less. The actions which are more rational from the standpoint of the objective decision problem will increase in frequency until an equilibrium is reached. From the standpoint of the population, then, we need to include every available action in our representation of the decision problem, and we must faithfully model the actual links between choices and outcomes.

This fact helps us to see why, from the standpoint of the individual deliberator, we cannot employ objective decision problems. For any deliberating agent, there are simply more actions available at any given moment than can be included in an agent’s subjective deliberation. Indeed, at any given moment we might say that there are an overwhelmingly large number of actions which might be performed; the number is so large that it becomes impossible to distinguish any actions in the mass. If I am deliberating about what to do, it is possible for me to decide to walk to any location in North or South America; it is possible for me to decide to read any book which has ever been published to which I have access; and so forth. Most deliberating agents who face decision problems do not consider such options in their deliberation; rather, they consider only those options which are feasible for the situation. If I am deliberating about what to do for the evening, I generally deliberate
between a finite set of options: read one of the books on my bedside table, for instance, or see one of the movies playing nearby.

Moreover, my deliberation is generally constrained according to ends I possess. For instance, I have already settled on cooking something for dinner, but I need to deliberate about what to cook; moreover, I have already decided to cook a broccoli dish, and so my deliberation is about what to do with the broccoli. Modeling my decision problem here as an objective decision problem is not only overkill, but it also seems to get the normativity wrong. Suppose I have decided to cook something involving broccoli, and I am deliberating about what to cook. (We can suppose here that my choice will be more or less rational according to how many new ingredients I need to buy, how much time it will take to cook, and so forth.) The fact that it might be more rational for me to spend my time pursuing the general betterment of humanity seems quite irrelevant for my deliberation about what to cook; and the objective decision theorist telling me that every dish I might cook involving broccoli is therefore irrational given the rational obligation I have to better humanity is odd to say the least.

The objective decision theorist cannot account for such constrained decision problems at all, because the objective decision theorist must include every available action and outcome into its representation of the decision problem. The objective decision theorist must then insist that every decision that we make be evaluated according to whether it is the very most rational choice available. Such attention to the rational properties of every action renders the deliberating agent no less of a rational fool than the sort warned against by Sen.\footnote{Sen 1977.} We should conclude that we ought to base our model of deliberation upon subjective decision problems rather than objective decision problems if we are to model the kind of deliberation performed by actual individual agents.

### 4.4 The possibility of constrained rational deliberation

Our model of deliberation, then, aims to account for the rational resolution of subjective decision problems. What constitutes the rational resolution of a subjective decision problem?
4.4.1 Preferences

Traditional rational choice theories argue that subjective decision problems are rationally resolved through the unconstrained maximization of an agent’s preference function. It is important for traditional rational choice theories that preferences be rationalizing elements in a subjective decision problem; preferences must be able to rationalize the choice or choices that an agent makes. Preferences do so, according to traditional rational choice theories, by establishing a ranking of outcomes: an agent’s preference function ranks outcomes, and it is rational for agents to choose to bring about those outcomes which maximize the agent’s preference function. Because agents in subjective decision problems have beliefs about the links between choices and outcomes, deliberation requires agents to engage in practical reasoning about which action is connected to the most highly preferred outcome.

Why should preferences rationalize choices? Why, in other words, should we think that an agent’s preferences over outcomes have anything to say about which outcome is rational for an agent to bring about?

Not all rational choice theorists subscribe to the view that the rationalizing attitudes are psychological attitudes on par with preferences. As Sugden has pointed out, rational choice theorists since Savage have argued that behavior is justified not by psychological preferences, but instead by revealed preferences, which can be equated with prior choices. Sugden has also pointed out many of the flaws that come along with attempting to disassociate utility from psychological preferences: Savage’s utility depends on implausible axioms of rationality, such as the claim that preferences are complete, or that preferences satisfy transitivity. Savage’s axioms, in other words, are implausible from a psychological point of view, and are thus ill-suited as grounds for a theory of subjective rational choice.

We need something with greater psychological plausibility to rationalize choices, especially since our model of subjective rational choice requires that everything relevant to an agent’s deliberation be filtered through the agent’s own psychological states. I propose that we use simple psychological preferences—with the understanding that we are merely characterizing some properties of psychological preferences, without any pretense to having given a complete characterization.

What is a simple psychological preference, then? Such preferences have two primary properties of interest: first, they are motivational; and second, they are comparative judgments of desirability. It is the first feature that distinguishes preferences from other mental.

\[17\] Sugden 1991.
states, such as intentions, that will be considered later. The first condition merely states that in order for a mental state to be a preference, it must motivate the agent to act—it must exert some kind of motivational pull over the deliberating agent.

We should be careful in explaining what it means for a mental state to be motivational—especially since we want our proposed theory to be compatible with the possibility that the mental states of which we are phenomenologically aware might be epiphenomenal to any successful causal account of action. In particular, the account we give here of deliberation must be compatible with the possibility that my consciously choosing to perform some action is not itself the cause of my performing that action.18

It is not necessary for us to give a complete account here of how preferences might entail subjective feelings of motivational pull even if they are or might be epiphenomenal to the true causal story of action. It suffices here to note that subjective feelings of motivational pull are just that—subjective feelings. When an agent X holds a preference for outcome \( o_1 \) over outcome \( o_2 \), this means that X experiences a motivational pull towards \( o_1 \) when faced with a choice between \( o_1 \) and \( o_2 \). (Motivational pulls are, in this sense, context-dependent; I do not claim that a preference for \( o_1 \) over \( o_2 \) translates into a general, context-independent motivational pull towards \( o_1 \).

This motivational pull corresponds to a feeling X has of being motivated to act upon \( o_1 \). Whether this feeling is a true feeling about what leads X to act or not is a question which we can leave open.

Besides preferences being motivational attitudes, preferences are also rationalizing attitudes. Holding a preference for outcome \( o_1 \) over \( o_2 \) makes it more rational ceteris paribus for the agent to choose \( o_1 \) in any choice situation which features both \( o_1 \) and \( o_2 \). Preferences are rationalizing attitudes, I argue, because preferences correspond to comparative judgments of desirability; preferring outcome \( o_1 \) over \( o_2 \) entails that the agent judges \( o_1 \) to be more desirable than \( o_2 \) in at least some metric of desirability. This does not imply that the agent judges the first outcome to be superior to the second in every metric of desirability—it just means that there is some metric of desirability according to which the first is superior to the second.

Suppose, for instance, that agent X prefers cake to apples. (More precisely, we should speak of X’s preferring the outcome of her eating the cake to the outcome of her eating the apples, or even more precisely, of preferring the outcome of eating the cake to that of eating

18 My goal here is to stay in accord with the experimental results obtained by Libet et. al. which strongly suggest that conscious intentions to act are themselves subsequent to, and therefore likely initiated or caused by, unconscious brain events. See e.g. Libet 1999.
the apples given certain conditions that $X$ is in at the time of course; we can skip over this additional level of description unless it is necessary.) $X$’s preferring cake to apples means that there is some way in which $X$ judges cake to be more desirable than apples—taste, perhaps. Further, the motivation condition on preferences means that this way in which cake is more desirable than apples exerts some motivational pull on $X$.

Preferences entail a judgment of comparative desirability according to some metric, but there might be other metrics of desirability. $X$ might prefer cake to apples in virtue of their taste, but $X$ might judge it more healthy to eat the apples than to eat the cake. Further, the fact that apples are more healthy than cake might, in $X$’s eyes, provide her with reason to choose the apples rather than the cake in spite of her preference for the former. There might, in other words, be another metric of desirability which, in the agent’s eyes, makes it more rational to choose the apples rather than the cake. Rational counter-preferential choice is possible. I will make the argument for this claim below; here I want to argue that, in the absence of complicating factors, preferences do rationalize choices.

Indeed, it makes sense that preferences would rationalize choices in the absence of other complicating factors. Suppose agent $X$ faces a choice between outcomes $o_1$ and $o_2$. And suppose further that all $X$ has to go on in her choice is the fact that she prefers $o_1$ to $o_2$—she has no further judgments about the desirability of one outcome as opposed to the other, and so she has no other judgments about the rationality or irrationality of bringing about one or the other, she has no intentions to bring about one or the other, et cetera. In this case, $X$’s preference rationalizes her choosing $o_1$ rather than $o_2$. This is not merely because $X$ has nothing else to base her decision on (although this fact is relevant); more importantly, $X$’s preference corresponds to a judgment of comparative desirability, and this comparative desirability is enough to justify choosing the one over the other in the absence of complicating factors. Judging something desirable entails a recognition of some kind of value in the thing judged desirable; this judgment of value is defeasible, but it is a judgment of value nonetheless. Absent other judgments of value, and absent concerns about the irrationality of that judgment of value, that judgment of value suffices to rationalize a choice of the thing judged valuable.

The fact that preferences rationalize choices is what has led to the traditional model of rational choice, in which agents are rational insofar as they maximize, without constraints, their preferences with respect to outcomes produced. It is this model that I want to criticize in the remainder of this section.
4.4. THE POSSIBILITY OF CONSTRAINED RATIONAL DELIBERATION

4.4.2 Intentions and values

As we saw above, in the absence of other mental attitudes relevant for deliberation, preferences rationalize choices—for the agent who only holds preferences over outcomes, the rational choice is the one which maximizes the preferability of the outcome chosen. In the example of agent X facing a choice between cake and apples, we suggested that there might be situations in which X rationally chooses the apples in spite of the fact that she prefers the cake to the apples. I want to suggest in this section that agents can impose rational constraints upon their own deliberation, such that there might be outcomes which are rational for them to choose to bring about in spite of the fact that they are not the most preferred outcome in the decision problem.

It matters that we are talking about constraints, because constraints can only operate on what is already present. If an agent’s deliberation is rationally constrained, then certain outcomes which are already present in the decision problem will be judged irrational to bring about, and other outcomes which are also already in the decision problem will become rational to bring about. It is not the case that the constrained agent’s deliberation will recommend an outcome which was not previously a part of the agent’s decision problem. The agent X who faces a choice between cake and apples might constrain her deliberation so as to prefer the cake, or she might constrain her deliberation so as to prefer the apples. Her constrained deliberation will not, however, recommend an outcome which is neither cake nor apples—pie, for instance. This is important, because if constrained deliberation is to render rational the production of an outcome which was not rational to bring about prior to the constraints, it should not do so by making rational some outcome which is not part of the agent’s original conception of the decision problem.

There are four properties that a mental attitude must satisfy in order for it to be a rational constraint upon preferences in an agent’s deliberation. First, the attitude must constrain deliberation. Second, the attitude must be reason-providing. Third, the attitude must be, at least potentially, counter-preferential; it must float freely of an agent’s preferences. Finally, the attitude must be volitional; it must entail a volitional commitment on the part of the possessing agent to bring about the target outcome or target action. There are two kinds of mental states which satisfy these four conditions: there are intentions, and there are values. I argue that these two mental states satisfy the four conditions within the context of decision problems, and, further, that these two mental states can be taken to be functionally identical—again, importantly, at least within the context of decision problems.
That these four conditions suffice for a mental state to impose rational constraints upon deliberation will be argued for below.

**Intentions**

What are intentions? Since the wane in influence of the Davidsonian stance, there has developed a general consensus around what intentions are. Intentions are *volitional pro-attitudes*. To be volitional is to indicate a practical commitment on the part of the holding agent towards bringing about the content of the intention; the agent who intends to do action $x$ does not just desire that $x$ be done, but is rather settled on and committed to doing $x$. And intentions are pro-attitudes in that they reveal something of value that the agent sees in doing the action or bringing about the outcome. Agent $X$’s intention to $x$ shows that $X$ sees *something* of value in $x$, even if $x$ is something as seemingly devoid of value as drinking a can of paint.

The fact that intentions are volitional directly contributes to other important features of intentions. For one, intentions are not reducible to beliefs and desires; an agent’s sum total of beliefs and desires do not suffice to determine the agent’s intentions, *even if the agent is fully rational*. Moreover, agents who hold intentions face a normative requirement of *means-end coherence*: an agent’s intentions must cohere with that agent’s other intentions and with the agent’s beliefs about the world.

Much has been said about intentions as independent *mental states*, but we are focusing our attention here on the *functional role* of intentions in subjective decision problems. Intentions serve two roles in decision problems: they serve as *inputs* to an agent’s decision-making, and they serve as *outputs* of deliberation. In their role as inputs, intentions affect the rationality of agents’ resolutions of subjective decision problems. We will see how below, but it is worth noting here that in doing so, intentions are no different than an agent’s other decision-relevant mental states such as beliefs or preferences. As outputs, intentions are the products of deliberation. In Broome’s words,

> Forming an intention [through a process of practical reasoning] is making a decision. Making a decision is as close to acting as reason can possibly get you.\(^{19}\)

Now, we should be careful here; *forming an intention* and *making a decision* both seem to be actions of some kind (reflected here in the active verbs ‘forming’ and ‘making’), and so

\(^{19}\)Broome 1999, p. 407.
it would seem that reason can get us all the way to a certain kind of action. But we can set aside for now the kind of mental actions which result in the formation or modification of a mental state. Broome’s point is that the only conceivable output of deliberation is a choice or decision; the output of deliberation cannot be a (non-mental) action, because deliberation by itself cannot carry the agent all the way to action; action requires something more than mere reasoning ability.

There is a tension in these two roles that intentions play. As inputs to deliberation, intentions impose rational constraints upon an agent’s decision-making; they help to determine whether an agent’s choice is rational or not. As outputs of deliberation, however, intentions are themselves subject to rational evaluation; the formation of an intention is an action, and as such is either rational or not.

We can alleviate much of this tension by recognizing that intentions are not much different from an agent’s preferences in their being potentially rational or irrational. Consider a typical Humean model of deliberation in which the agent’s preferences determine the rationality of her or his choices, and where preferences are not themselves subject to rational evaluation. Even in such a case, an agent’s preferences might be irrational, as would occur when one preference goes against the bulk of the agent’s other preferences. X’s very strong preference of having one more drink might conflict with the remainder of her preferences, such as her preference to arrive home safely, or her preference of avoiding a dependency on alcohol. Although X’s holding a preference of taking another drink might itself be reason-providing, the preference can also be considered irrational in that it promotes an outcome which is irrational to bring about after a full consideration of the agent’s reasons.

The analogy between preferences and intentions is not perfect. We can talk about whether the formation of an intention is rational, since the output of deliberation is, properly speaking, the formation of an intention. It is more awkward to talk about whether the formation of a preference is rational or irrational, because we don’t typically think of preference-formation as under an agent’s volitional control. Intentions, then, insofar as they are outputs of deliberation, are akin to actions, and they can be rationally criticized as such. Preferences are not rationally criticizable as actions; we don’t think an agent irrational for coming to hold a preference, though we might think the preference itself irrational.

Nevertheless, the analogy does reveal something important about intentions: namely,

---

\(^{20}\)Agents might, of course, be able to control their preferences indirectly, as when people control their environment, or undergo certain kinds of education, or submit to certain training regimens.
that it is not inconceivable that they should both be a source of reasons for the deliberating agent, and that they should be rationally evaluable according to the subjective reasons that the agent possesses. The way in which intentions can be rationally evaluated will be discussed below. Before doing so, we will first establish that intentions satisfy our requirements for a mental attitude that imposes rational constraints upon an agent’s deliberation.

Let us, then, consider whether intentions can rationally justify counter-preferential behavior by seeing whether intentions satisfy the four conditions laid out above. The first condition that intentions must satisfy in order for them to generate rational constraints upon deliberation is that intentions must actually constrain deliberation. This condition is easily seen to be satisfied. Intentions constrain deliberation because rational intentions must satisfy norms of coherence; holding an intention to $\phi$ or to bring about outcome $o$ rationally compels an agent to not do any action which is incompatible with doing $\phi$ or bringing about $o$. If I have an intention to go running early in the morning, then I cannot rationally stay out late, if my staying out late is incompatible with my running early in the morning. Intentions thus filter my available options; every action which is not compossible with the satisfaction of my intention is filtered out of the decision problem that I face.

In their capacity for filtering options, intentions can be broad or narrow. Agent X might have an intention to prepare a dish with duck, or X might intend to prepare duck a l’orange. X, playing agent Y in a game of chess, might intend to checkmate him; or she might intend to checkmate him by first taking all of his pieces; or, she might intend to checkmate him in particularly humiliating fashion with just a bishop and a knight. Each of these intentions range over a set of outcomes, and they each exclude some outcomes while including others. The chef might intend to cook duck a l’orange, but have several menus to choose between; in this case, there are only a few outcomes that are compatible with her intention—namely, those outcomes that are produced by her following one of her available recipes. The chef who merely intends to cook a duck dish, on the other hand, has many more intention-compatible outcomes to choose between in her subsequent deliberation. The agent who intends to checkmate her partner has a number of outcomes which are intention-compatible (although, of course, not all of them are available at every stage of the game). The agent who intends to take all of her partner’s pieces has fewer outcomes which are intention-compatible, while the agent who intends a bishop-knight checkmate has a different set of intention-compatible outcomes. Broad intentions, then, eliminate fewer options from rational consideration, while narrow intentions eliminate more options from consideration.
However, all intentions, whether broad or narrow, do constrain the deliberation of rational agents. As such, intentions do satisfy the first condition that mental attitudes must satisfy to rationally justify counter-preferential behavior.

The second condition requires that intentions be reason-providing; this is the most difficult point to argue for, and it will accordingly be left until last.

Intentions satisfy the third condition, namely that they be potentially counter-preferential. We claimed above that preferences are comparative judgments of desirability with a motivational affect. Intentions need not have any such motivational affect to them; agents can hold intentions even though the agent has no motivation whatsoever towards doing the thing intended. I might intend to run early in the morning—an intention which has the consequence of making irrational my sleeping in, in spite of the fact that my preferences are strongly in favor of sleeping in rather than running. It is the fact that intentions need not have any motivational affect which allows them to be counter-preferential; because they are different mental states than preferences, they can recommend the production of different outcomes than preferences do.

And, of course, intentions are volitional. As we have already mentioned, intentions are distinguished by their volitional component; in this way, intentions are different from preferences or other non-volitional pro-attitudes. In calling intentions volitional, we are calling attention primarily to their capacity to settle deliberative questions that agents face. Non-volitional attitudes such as desires do not settle deliberative questions; an agent X who desires $x_1$ over $x_2$ still might face the deliberative question of whether to choose $x_1$ or $x_2$. If X intends to choose $x_1$ over $x_2$, however, then X does not face the deliberative question of which to choose.

As Bratman has pointed out, the volitional component of intentions is primarily future-directed and not present-directed.\textsuperscript{21} Present-directed intentions are called by Searle ‘intentions in action’.\textsuperscript{22} Present-directed intentions also share much in common with the traditional philosophical concept of a willing or a volition.\textsuperscript{23} Whatever we choose to call present-directed intentions, for a theory of deliberation and rational guidance we must talk about future-directed intentions and not about present-directed intentions. Present-directed intentions are incapable of providing rational guidance because the agent who fails to act in

\begin{enumerate}
\item Bratman 1984; Bratman 1987.
\item Searle 1980; Searle 1983.
\item See, for example, McCann 1974 or Velleman 2007.
\end{enumerate}
accordance with her or his present-directed intention but is not bound or otherwise physically constrained is not irrational; rather, the agent simply fails to hold the intention in question. Thus present-directed intentions cannot serve as inputs to deliberation. Present-directed intentions might sometimes be outputs of deliberation. Nevertheless, we will follow Bratman in speaking primarily of future-directed intentions; given our dual interest in intentions as inputs and intentions as outputs, focusing on future-directed intentions will enable our theory to account for both functions. We can conclude, then, that intentions are volitional attitudes.

We now return to the question of whether intentions provide rational guidance for deliberating agents. I argue that intentions do provide rational guidance for an agent; adopting an intention to φ (where φ is either the performance of an action or the production of an outcome) generates some (perhaps defeasible) rational pressure on the agent to do φ.

Our argument proceeds by reductio. Suppose that we have a rational agent, and suppose that intentions did not necessarily generate rational constraints upon the rational agent. Let us call the intention which does not generate rational constraints an irrational intention. Intentions are generally formed as outputs of decision problems. The irrational intention, then, must be an intention which was an irrational resolution of the decision problem that the agent previously faced.

Remember that we are concerned with subjective decision problems. In order for an irrational intention to be an irrational resolution of a subjective decision problem, the intention must be an irrational one for an agent to have formed in the light of the agent’s own beliefs about her or his reasons. The agent must, in other words, simultaneously judge herself or himself to intend to φ, and at the same time believe that doing φ is not supported by the reasons the agent had at the time of forming the intention. The agent must be able to say, in other words, both that she or he intends to φ, and also that doing φ is irrational.

Uttering those two phrases in tandem might be possible when we limit our talk to present-directed intentions. For present-directed intentions, such a situation might arise in cases of weakness of will, where an agent simultaneously intentionally acts so as to do φ while recognizing that the balance of reasons supports some other action or outcome ψ.

This does not work when we talk about future-directed intentions, however. In the case of future-directed intentions, an agent must simultaneously intend to do φ in the future, and also believe that doing φ is not supported by the reasons the agent has. This is akin to agent X saying that she intends to have a drink at the bar later on in the evening, in spite
of the fact that she knows that having a drink is irrational in light of the reasons she has
to drink or not drink.

Now, such a scenario is still possible; an agent might intend to φ while still believing that
doing φ is irrational. But note that such an agent is not a rational agent. By intending to
do an irrational action, the agent reveals herself to be not sufficiently responsive to reasons,
which contradicts our initial assumption that the agent be rational.

It might be argued that this reductio does not yet show that intentions must be reason-
providing; it might be argued that we have only shown that one’s future-directed intentions
must be in accordance with one’s perceived balance of reasons, but that the intention itself
might be rationally inert. But intentions do not merely constrain deliberation; they also
set an agent’s ends. And, as Bratman has pointed out, in setting an agent’s ends they also
generate rational pressure on the agent to take those actions which are sufficient means to
the satisfaction of the end.\footnote{Bratman 1987.} This rational pressure is a positive rational commitment, and
is present in the agent who holds the intention but is not present in the agent who does not
hold the intention. Say that an intention φ requires an agent X to perform action x in order
to satisfy the intention. If X does not perform action x, then the agent will be guilty of
failing to take the necessary means to satisfy her end, and she is, to that extent, irrational.
This threat of irrationality provides X with reason to perform action x, so long as she holds
the intention φ. And, as argued immediately above, X must judge her intention φ to be in
accordance with her reasons for action.

I argue, then, that intentions satisfy the four conditions set out above to provide rational
constraints upon an agent’s deliberation in the context of decision problems. My argument
here has three more steps. First, I argue that values are mental states which also satisfy
the four conditions. Second, I argue that intentions and values are functionally identical in
the context of decision problems, and that we can therefore treat them identically. Third, I
will argue that these four conditions suffice for a mental state to impose rational constraints
upon deliberation. I turn first to the claim that values also satisfy our four conditions.

Values

A note on terminology before we begin. When we talk of values, we are talking about values
for an agent. When we talk of agent X’s values, in other words, we are talking about the
things that X values. We are not making the further claim that the things that X values

\footnote{Bratman 1987.}
are actually valuable. Our talk of values, then, does not purport to talk about values ‘in the world’; rather, we are talking only about values ‘for an agent’. It might, then, be more precise to speak of ‘valuings’ than about ‘values’, insofar as the former term more clearly indicates that we are talking about values in the perception of deliberating agents. For ease, we will use the term ‘values’ rather than ‘valuings’, with the understanding that we are talking about values for an agent.

What, then, are values? Engagement with the totality of the philosophical literature on values here will prove impossible. The essence of values, I argue, is that they are comparative judgments of desirability which are also necessarily reason-providing. Strictly speaking, then, agents do not value things; rather, agents value some things (here, outcomes or actions) more than other things (again, outcomes or actions). To say of an agent X that she values the opera, then, is to not make a meaningful statement within the context of a decision problem. It is meaningful to say that X’s valuing the opera gives her reason to attend the opera rather than attend a movie, or to say that her valuing the opera gives her reason to donate money to the opera rather than spend the money on clothes. For a value to be meaningful in the context of a decision problem, the value must provide reason for an agent to do one thing rather than another.

Values, then, are comparative judgments of desirability which are reason-providing. The agent who values outcome o₁ over outcome o₂ judges o₁ more desirable than o₂ with respect to something, and the agent also takes the way in which o₁ is more desirable than o₂ to provide the agent with reason to produce o₁ rather than o₂.

In what follows, we will assume that values are commensurable—that is, that for any two values that an agent holds, and which provide reasons to act in some decision proble, the agent can compare the strengths of the reasons that the values provide in order to arrive at a conclusion about what it is rational to do. This is not an uncontroversial position. This assumption is made in part in order to simplify our model of deliberation. But even disregarding the benefits for our modeling, we have reason to assume values to be commensurable: agents who hold incommensurable values cannot appeal to those values to rationalize their choices in decision problems. If an agent holds two incommensurable values—e.g. between respecting the possibility of sustainable development of some land and respecting the cultural and religious values of native peoples who inhabited the land prior to the current occupants—then the agent must resolve the decision problem by appeal to

---

some rationalizing attitude other than the incommensurable values. For the purposes of
the decision problem, then, it is as though the incommensurable values do not exist—they
do not inform the agent’s decision, so they are of no practical import. For this reason,
we will assume that all of the values which are relevant for an agent’s decision-making are
commensurable, and we can disregard those values which are not commensurable.

So much, then, for what values are. I argue now that values satisfy the four criteria set
out above for mental states which can rationally constrain deliberation: again, that they
constrain deliberation, that they are reason-providing, that they be counter-preferential,
and that they be volitional.

In arguing that values can be modeled as constraints on deliberation, we can directly
follow Levi:

\[ V \text{alue commitments impose constraints on the ways in which feasible options are}
\text{evaluated. The totality of value commitments endorsed by an agent delimit a set of}
\text{permissible ways of evaluation of the feasible options available to the agent. [...] Given his value}
\text{commitments and his knowledge of the feasible options, a rational agent ought to restrict his}
\text{choice to the set of V-admissible options.}^{26} \]

It is clear here that values constrain deliberation in much the same way that we suggested
intentions constrain deliberation above; the rational agent chooses an outcome which is
compatible with her or his values, just as the rational agent chooses an outcome compatible
with her or his intentions.

It is much clearer that values are reason-providng mental attitudes than it was in the
case of intentions. Indeed, we can take it to be definitional; an agent’s values are subjectively
reason-providing. If a mental attitude is not subjectively reason-providing, then it cannot
really reflect the agent’s values. If agent $X$ were to claim to value producing outcome $o_1$
more than outcome $o_2$, and yet claim to not see any reason to produce $o_1$ over $o_2$, we would
think that she was lying, or mistaken about what valuing an outcome entails.

It is equally clear that intentions can be counter-preferential. We defined preferences to
be motivational pro-attitudes. Values are also pro-attitudes, but we have said nothing yet
as to whether they are motivational attitudes or not. It might seem as though values are
necessarily motivational attitudes; it might seem just as odd to say of an agent $X$ that she
values doing $o_1$ but is not motivated at all to doing $o_1$ as it did to say that $X$ values doing
$o_1$ but does not judge herself to have reason to doing $o_1$ at all.

\[^{26}\text{Levi 1986, p. 47.}\]
It might seem as though the two cases are analogous, but they are not. An agent can indeed value the production of one outcome over another without being motivated to produce that outcome; this occurs in those cases where agents have to struggle to act in accordance with their deeply held values against their baser preferences. One agent might value justice and forgiveness, and yet feel no motivation whatsoever to act accordingly when face to face with the raper and murderer of his mother. Another agent might value charity, but feel no motivation to give money to the beggar who confronts her. Finally, to return to a familiar example, an agent might value eating healthily, but feel no motivation to pick the apple in front of her rather than the cake next to it.

It is not plausible, I think, to argue that when the agent does not have the motivation, then the agent does not have the value. For, should the agent fail to produce her or his valued outcome, the agent will feel as though she or he has done something irrational. If my preference to eat the cake is strong enough to cause me to eat the cake rather than the apples, I will not conclude that I had no reason to eat the apples after all. No, I will feel as though I should have eaten the apples, because I judge myself to have reason to eat the apple. That reason comes from my value.

In any event, in each of these cases we have instances in which an agent’s values contradict the agent’s preferences, which is what we were looking for in the first place. It is important, however, that what we have is not merely a conflict in the agent’s preferences. What have instead is a conflict between the agent’s preferences and some other mental state—values, in this case—which are reason-providing in a different way than preferences are.

It remains to argue that values are volitional attitudes. And by themselves, of course, values are not volitional attitudes; it is quite possible to value the production of an outcome (world peace, for instance) but not be committed in any ay to producing the outcome. To be committed to world peace is to take one’s self to have reason to promote world peace at some point; and yet one needn’t be committed to promoting world peace at any particular moment.

Values, then, are not in general volitional attitudes. But values are volitional attitudes for rational agents in the context of decision problems. Take a rational agent X facing a choice between outcomes o₁ and o₂. Suppose that X values the production of o₁ over o₂. If X chooses to bring about o₂, then we must conclude one of two things: either X is not a rational agent, or X does not really value producing o₁ over o₂. By definition, X is a
4.4. THE POSSIBILITY OF CONSTRAINED RATIONAL DELIBERATION

rational agent. Thus, \( X \) must not really value producing \( o_1 \) over \( o_2 \).

This argument hinges upon the fact that \( X \)’s values exhausted her reasons: there were no reasons relevant for \( X \)’s decision which were not captured by her values. This is how we were able to claim that, if \( X \) is rational, then she will choose \( o_1 \). Are we entitled to believe that \( X \)’s values exhaust her reasons here? We have so far identified two sources of subjective reasons for deliberating agents besides values: those generated by preferences, and those generated by intentions. I claim that the reasons generated by values take priority over those generated by preferences; this point will be elaborated upon further below. I further claim that intentions and values are functionally identical; this point also will be clarified below. If these points are true, then the reasons generated by values exhaust \( X \)’s reasons, and so insofar as \( X \) is rational, she will form intentions to act in accordance with her values.

**Intentions and values are functionally identical**

I suggested above that any mental state which satisfies the four conditions (rationality, constraint, counter-preferentiality, and volition) is such as to rationally constrain an agent’s preference-based deliberation. This follows immediately from the four conditions themselves. The first condition establishes that the mental state provides some kind of rational sanction for the selection of some outcome or other. The second condition allows that the mental states constrain deliberation. This condition is important because it ensures that the mental state influences the rationality of an agent’s choice not by introducing new options into the decision problem, but by refining the options which are already available. It should not be a necessary condition for forming an intention, or holding a value, that an agent form a new belief about which actions are available to perform, or which outcomes are available. The third condition ensures that intentions and values have some role to perform. If neither intentions nor values ever diverged from preferences—if they were always reducible in some way to the agent’s preferences or desires—then we would not need to include the mental state in our model; it would be more parsimonious to restrict our model to the underlying preferences.

We should then ask ourselves whether we ought to include both intentions and values in our model. I argue that we need not: in the context of decision problems, intentions and values are functionally identical, and so we can make do with one mental state (which we will call intentions) to serve the role of both. Crucial to this argument is the fact that the
reasons providing by intentions is functionally identical to the reasons provided by values in the context of decision problems. Since this is the case, we needn’t include both intentions and values in our decision model.

Why are the reasons provided by intentions and values the same? The answer is that they are both volitional attitudes in the context of decision problems; it is this fact which makes the reasons provided by intentions and values of a different type than that provided by preferences. Intentions and values have two components to them: they are comparative judgments of desirability, as preferences are, but they are also volitional, in that they represent an actual or ideal choice on the part of the deliberating agent. (Intentions are actual choices, while values are ideal choices—they represent the choice that an agent would make if she or he were rational.) And deliberate choices are important sources of reasons for deliberating agents. The act of deliberation aims at generating choices; deliberation is successful when an agent chooses something to do or an outcome to bring about. Intentions and values reflect the choices that an agent has already made; they reflect those things that the agent already takes to be goods worth pursuing. As such, they provide reasons for action which are rooted in the agent’s agency itself. To be an agent is to be the kind of creature that can make choices, and can respond to the reasons which one perceives and to those things which one takes to be goods. When one has made a choice, that choice provides reasons which are rooted in the agent’s agency.

Both intentions and values, I argue, share in this volitional feature. Further, for both intentions and values, it is this volitional feature which distinguishes the source of reasons that they provide from the reasons that preferences provide. As such, we can treat intentions and values as the same in our model of deliberation. For ease, we will call these attitudes intentions; speaking of intentions is clearer when we consider the mental attitude as both an input to deliberation and as an output of deliberation.

We have not yet said exactly how we should treat the reasons provided by preferences and the reasons provided by intentions. This will be treated in the following section.

4.4.3 Intentions and rational constraints

Central to my claim that intentions generate rational pressure upon deliberating agents is that intentions and preferences both generate different kinds of normative pressures. Accordingly, there are (at least) two different kinds of rationality concepts: there is preference
rationality, and there is intention rationality. I want to argue that both preference rationality and intention rationality are legitimate concepts of rationality, and hence that both intentions and preferences generate some kinds of normative pressure upon the deliberating agent.

Before discussing preference rationality and intention rationality, however, we ought to say something first about why we might think that the two concepts pull apart at all. One way of expressing the worry is as follows: why not think that we can make do with just one single, unified, comprehensive preference function which determines the subjective rationality of an agent’s choices? If an agent X intends to do some action x, why not think that any rational pressure which is derived from X’s intention comes from X’s preferences?

The simple answer is that preferences and intentions are different mental states. Preferences are motivational comparative judgments of desirability, while intentions are non-motivational judgments of desirability (which are, in the context of a decision problem, comparative). Because preferences and intentions are different mental states, then, they should be treated as different in our model of deliberation, and any rational sanction which the attitudes provide should be distinguished from one another.

This simple answer is a start, but it is not sufficient as it stands. Even if preferences and intentions are distinct mental states, this does not mean that our model of deliberation needs to treat them as different. The fact that preferences are motivational while intentions are not might very well be irrelevant from the standpoint of the deliberating agent; what matters is not whether an attitude motivates, but rather whether the attitude rationalizes. And, the argument continues, if preferences and intentions rationalize in the same way, then we ought to treat them the same in our model of deliberation.

I think that it is, in fact, important that preferences are motivational attitudes. At the end of the day, agents must believe of their deliberation that it might result in action, and deliberation can only result in action if it is believed to be connected in the appropriate ways to motivation and behavior. The agent with no motivational attitudes cannot see her or his deliberation as connecting to action. While intentions might be important, then, in providing rational constraints upon an agent’s deliberation, the agent must still have the preferences, or deliberation is simply not possible. The clinically depressed cannot deliberate if they recognize themselves to have no motivation at all to act, regardless of whatever subjective judgments of reason they might possess over their choices.

The further claim I would want to make is that preference rationality and intention
rationality really are distinct rationality concepts, ones that we are familiar with in our everyday experience of deliberation. To make this claim, however, we need to investigate what these two rationality concepts amount to.

Preference rationality is a familiar concept. Traditional rational choice theory, as I defined it above, is based upon preference rationality. The preference rational agent seeks to maximize her or his preferences by choosing so as to bring about the most preferred outcome. Preference rationality is a subjective norm of rationality, and so the rational agent is the one who maximizes those preferences that she or he believes herself or himself to possess.\(^{27}\)

The intention rational agent also maximizes her or his preferences, but the intention rational agent only does so after the outcomes in the decision problem are filtered through her or his intentions. As Bratman has noted,

\[^{28}\text{intentions} \text{ constrain solutions to these [subjective decision] problems, providing a filter of admissibility for options . . . . They narrow the scope of deliberation to a limited set of options.}\]

Intentions do this by filtering out all those actions which lead to outcomes which are incompatible with an agent’s intention. Suppose agent \(X\) intends to be at home throughout the day. If \(X\) later faces a decision problem—how to prepare dinner, for instance—which consists of several options, one of which—going out to a restaurant—conflicts with her prior intention, then \(X\) is intention rational insofar as she chooses that action which is expected to maximize her preferences over outcomes after removing the intention-incompatible action. The intention-incompatible action here is going out to the restaurant, so \(X\) will remove that action from consideration and then choose whichever remaining action maximizes her preferences.

So long as we consider only those intentions which are intentions to perform some action (or one of a set of actions) or intentions to bring about some outcome (or one of a set of outcomes), then we can model intentions by way of such a filter in subjective decision

\(^{27}\)It is possible that our actual preferences might come apart from our beliefs about our preferences; we might simply forget whether, for instance, we prefer vanilla or chocolate ice cream. Given my definition of preferences, however, it is more difficult than we might expect to hold false beliefs about our preferences. Preferences have a motivational pull, and so if I have a preference for chocolate over vanilla, then I must also experience some motivational pull towards the chocolate. In normal circumstances, agents are not mistaken about whether or not they experience a motivational pull. It is possible that agents might be mistaken about whether they experience such a motivational pull, however; similarly, agents might simply not pay the necessary attention to their motivations when forming their beliefs about what their preferences are.

\(^{28}\)Bratman 1987, p. 33, emphasis original.
problems. There are some intentions that are not able to be modeled in this way. Consider performatives, such as the phrase ‘I do’ during a marriage ceremony. It is generally assumed that, in order for the utterance to be effective, the utterer must intend of that utterance that it have its desired performative effect. The man who says ‘I do’ without intending of his speech that he thereby be married is not taken to have caused anything with his words. For another example, consider Bratman’s intention-based account of collective actions discussed in Chapter 3. Bratman requires that agents who share an intention each have intentions that they perform the joint activity in accordance with and because of the we-intention in question and their meshing subplans. Bratman builds more into his concept of intentions than we can do here; Bratman allows agents to intend to perform some action because of some condition. For us, the content of an agent’s intention must be either an action itself or an outcome to be brought about (where ‘outcome’ here is read as a state of affairs).

Regardless of how broad or narrow the intention is, the intention rational agent first evaluates the preferability of those outcomes which are intention-compatible, and then the agent engages in standard instrumental reasoning about which action to take so as to produce the most preferred intention-compatible outcome. Intention rationality thus entails a sort of maximization, but it is a constrained maximization; the intention rational agent maximizes the preferability of those outcomes which are compatible with the agent’s self-imposed constraints.

Neither intention rationality nor preference rationality is equivalent to objective rationality. This is easy to see, if we consider whether it is necessary that an agent who is rational according to one of these two subjective rationality concepts is also overall rational. The answer is clearly no. An agent might have irrational preferences, such as a preference for murdering children or for drinking turpentine. Similarly, an agent might form irrational intentions—intentions to murder children or to drink turpentine are equally irrational from an objective standpoint.

It is equally true, however, that neither intention rationality nor preference rationality is equivalent by itself to subjective rationality. It is conceivable that an agent who is preference rational is nevertheless not overall subjectively rational, just as it is conceivable that an agent who is intention rational is nevertheless not overall subjectively rational. For the first, we can return to our agent deliberating between the cake and the apples. The agent’s preference is for the cake, while her intention is to have the apples. The agent’s intention to eat the apples is revealing of the agent’s underlying values; the agent judges herself to have
more reason to eat the apples than the cake. In this case, the agent’s intention is equivalent to a non-motivational comparative judgment of desirability in favor of the apples over the cake; the apples are seen to be more desirable in terms of some metric of desirability (here, healthiness). Moreover, in this instance the agent judges that the balance of reasons lies with the non-motivational metric of desirability.

It need not be the case, however, that the balance of reasons always lies with the non-motivational metric of desirability. An agent might be so sympathetic as to be genuinely moved by the suffering of others, and so she is strongly motivated to act in favor of her preference to alleviate suffering. The agent might also value promoting the general utilitarian welfare, and so intends to bring about those outcomes in which the general utility is maximally increased. Suppose the agent X faces a choice between physically torturing one person (and thereby creating enough of a deterrent to crime to maximize the general utilitarian welfare) and not torturing the person (thereby satisfying her preference for alleviating suffering). In such a case the agent might find it rational to choose in accordance with her preference to alleviate suffering, rather than with her intention to maximize the general utilitarian welfare.

Now, I don’t expect the preceding examples to have shown much of anything. Not only are our intuitions murky on what the rational choice might be when motivational and non-motivational judgments of desirability clash, but it’s not even clear in virtue of what the two examples are different, and why we might suggest that the balance of reasons lies with preferences in the one and intentions in the other. There are, I think, some general guidelines that we can employ in understanding the puzzle—though I don’t expect these guidelines to be definitive, nor do we need them to be.

First, if we understand the difference between preferences and intentions primarily as a difference between motivational judgments of desirability and non-motivational judgments of desirability, then we might well be convinced that neither preference rationality nor intention rationality has a privileged claim to be the subjective rationality concept. There is, after all, nothing that makes non-motivational judgments of desirability any more subjectively rational than motivational judgments of desirability. That we are more motivated to do one thing rather than another seems a poor basis on which to ground a reason to do one thing over another.

On the other hand, if we understand the difference between preferences and intentions as a difference between non-volitional and volitional attitudes, then it might seem as though
we have reason to privilege intention rationality. As volitional attitudes, intentions are ends; they are choices. To fail to be intention rational is thus to be means-end irrational, and part of our concept of ourselves as rational agents is that we are means-end rational. Under this way of looking at things, agents ought to be intention rational because ends have extra subjective normative weight that mere preferences do not have.

My own leaning is towards the second hand. As we discussed above, one of the characteristic features of agents is their ability to set ends; to systematically choose contrary to the ends that we set is to dissolve the coherency of our agency. Respecting our agency, then, requires that we give a special weight to our choices and our ends—and in our model, those choices and ends are represented by intentions and values. This is not to suggest that the weight we give to our intentions ought necessarily trump our preferences. There might very well be choice situations in which agents ought choose with the balance of their preferences rather than with their intentions; an algorithm for properly weighting the rational force of preferences and intentions may simply not be available.

We need not give a fine-toothed answer here. What we need to accept is the claim that intention rationality is a plausible and viable rationality concept in addition to, or perhaps in place of, preference rationality. If intention rationality should turn out to always trump preference rationality with respect to the subjective rationalization of behavior, so much the better for the theory presented here; our analysis of the nature and rationality of collective action only depends on the claim that intention rationality is a viable rationality concept.

4.4.4 Objections to intention rationality considered and refuted

We might wonder at the claim that acting contrary to one’s intention entails (at least) a pro tanto irrationality. After all, this seems to imply that intentions rationally compel certain behavior regardless of whether the intention itself is rational or irrational. Suppose $X$ holds an irrational intention $\phi$, and faces a decision problem between actions $x_1$ and $x_2$. Suppose further that $x_1$ leads to an outcome $o_1$ which is incompatible with $\phi$ (and is thus inadmissible), while $x_2$ leads to an outcome $o_2$ which is compatible with $\phi$. If acting contrary to intentions entails a pro tanto irrationality, then it would seem that only the action $x_2$ is rational for $X$ to perform. However, $o_1$ might be a rational outcome for $X$ to bring about, while $o_2$ is an irrational outcome for $X$ to bring about (consistent with the claim that $\phi$ is an irrational intention). Should we think that $X$’s rational choice ought to be to perform $x_2$, in spite of the fact that $o_2$ is an irrational outcome?
Broome insists that we draw a distinction between having a reason to perform some action \( x \), and being normatively required to perform an action \( x \).\(^{29}\) Broome’s argument proceeds as follows: in the same way that holding a belief \( p \) does not give us a reason to believe some (non-\( p \)) entailment \( q \) (since, if our belief in \( p \) is a false belief, we do not in fact have reason to believe \( q \)), neither does our intention to \( \phi \) provide us with reason to perform some necessary means \( x \), since our intention to \( \phi \) might be irrational, and we might in fact have no reason to \( \phi \). Although my holding a belief \( p \) might normatively require me to hold all (non-\( p \)) entailments of \( p \), it does not thereby give me a reason to believe the (non-\( p \)) entailments of \( p \). Similarly, Broome claims, \( X \)’s intention to \( \phi \), combined with her belief that \( x \)-ing is a necessary means of satisfying \( \phi \), normatively requires \( X \) to intend to do \( x \)—although it does not give \( X \) a reason to \( x \), since whether or not \( X \) has reason to \( x \) depends on whether \( X \) has reason to intend \( \phi \) in the first place.

Bratman expresses a similar worry which is relevant for the present discussion. For Bratman, if we grant that intentions provide reasons to act (and, therefore, that acting contrary to one’s intention entails a pro tanto irrationality), then we license an illicit form of reason bootstrapping.\(^{30}\) Suppose \( X \) has an all-things-considered irrational intention to \( \phi \). Suppose further that \( X \) is deliberating between three means of satisfying her intention to \( \phi \): \( x_1 \), \( x_2 \), and \( x_3 \). Even if \( X \) reasons that \( x_3 \) is the best (i.e. most practically rational) means of satisfying her intention to \( \phi \), Bratman argues, we still ought to conclude that doing \( x_3 \) is irrational for \( X \), since her intention to \( \phi \) is irrational. However, Bratman says, if we allow for intentions to provide an agent with reasons to act, then we cannot rule out the possibility that \( x_3 \) is, on balance, rational for \( X \). After all, \( X \) has the intention to \( \phi \), and her intention to \( \phi \) provides her with reason to do \( x_3 \). This is an irrational bootstrapping, Bratman argues, because it seems to entail that \( X \) is rational in taking the means to satisfying her intention to \( \phi \), even though her intention to \( \phi \) is itself irrational. Bratman concludes his argument against bootstrapping by noting that there is an asymmetry inherent in the process: if \( X \) were to form the all-at-once intention to \( \phi \) by means of \( x \)-ing, \( X \) would be irrational. However, \( X \) can form the (irrational) intention to \( \phi \), and subsequently form the (now rational) intention to \( x \). This situation is untenable, Bratman argues, and so we should reject the claim that intentions provide an agent with reasons to act.

\(^{29}\) Broome 2002.

\(^{30}\) Bratman 1987.
Bratman certainly does not conclude as a result of the previous argument against bootstrapping that intentions cannot serve as a filter of admissibility over actions in decision problems. Bratman’s objection depends only on the thin claim that intentions provide an agent with reasons of any kind. The model of deliberation that I present here does, in fact, depend on the claim that forming an intention thereby provides an agent with a reason of some kind to act in accordance with that intention. My claim thus seems to be incompatible both with Broome’s objection, and with Bratman’s worries about reason bootstrapping. How should we answer such worries?

Both of these worries seem to depend on an illicit switch between an evaluation of the agent’s choice as a resolution of a subjective decision problem, and an evaluation of the agent’s choice from the standpoint of objective rationality. Consider first Broome. Broome suggests that forming an intention cannot give an agent a reason to act, but rather can only normatively require an agent to act. If by ‘give a reason’ Broome is referring to an agent’s objective reasons for acting, then Broome is correct—the formation of an intention cannot affect the objective reasons that an agent has. However, as we noted above, an agent does not have any appeal to her or his objective reasons in resolving subjective decision problems. The question that must instead be asked is, does an agent’s belief that she or he has formed an intention affect which action is subjectively rational for the agent to perform? Here the answer is yes—a fact which is acknowledged by Broome himself. Broome’s ‘normative requirement’ is, after all, precisely subjective normativity. And, given that our interest is in subjective normativity, we need not take Broome’s objection as posing any threat to our project.

What, then, about Bratman’s bootstrapping worry? Bootstrapping worries can fall into one of two general categories. The first we can call objective bootstrapping. Objective bootstrapping occurs when an action which is not objectively rational prior to the adoption of an intention becomes objectively rational after the adoption of an intention. The second bootstrapping worry we can call subjective bootstrapping; this occurs when an action which is not subjectively rational prior to the adoption of an intention becomes subjectively rational after the adoption of an intention.

If my account of intention rationality were to allow objective bootstrapping, that would be a problem. After all, it is not clear why the formation of an intention ought to change an action from being objectively irrational to objectively rational, unless the intention directly affected the rationality of the intended action. Although there might be cases in which
the formation of an intention affects the objective rationality of an action—Kavka’s toxin puzzle, for instance, or certain interpretations of Newcomb’s puzzle in which the predictor makes her or his prediction on the basis of your intentions at the time of choice—intentions do not generally directly affect the rationality of actions, and so objective bootstrapping is a problem. Or, at least, it would be a problem if my account were to allow it. Because my account of intention rationality is a subjectively normative account, it does not allow objective reason bootstrapping.

What about subjective reason bootstrapping, then? We have two subjective rationality concepts on offer, preference rationality and intention rationality. There are thus four potential strains of subjective bootstrapping: first, when an action is not preference rational prior to the formation of an intention, but becomes preference rational afterwards; second, when an action is not intention rational prior, but becomes intention rational afterwards; third, when an action is not preference rational prior, but becomes intention rational afterwards; and fourth, when an action is not intention rational prior, but becomes preference rational afterwards.

We have no reason to fear that the first option would arise. On our account, intentions do not affect an agent’s preference structure, and so the formation of an intention ought not change an action from being preference irrational to being preference rational. Even if it did, it would pose no challenge to our theory. If it should turn out that forming an intention changes my preferences, we need not worry; preference change in response to environmental stimuli is to be expected, and although this would grant agents a mechanism for affecting their own preferences, it is not clear how this would be subjectively irrational. As an example, suppose I form an intention to eat fewer sweets. My preferences are initially stronger in favor of eating sweets; my intention to eat fewer sweets, however, contributes to my favoring sweets over non-sweets in fewer circumstances. Does this make my resulting preference for non-sweets irrational? I cannot see why this should be the case.

The second option does arise, but again there is no worry. Indeed, this is the whole point of forming intentions—by forming intentions, actions become intention rational which were not intention rational before. X’s choosing the apple is not intention rational prior to her forming the intention to choose it, but it does become intention rational afterwards.

The two hybrid cases, three and four, also do not pose problems. Case three is a simple variation of case two, and case four is vacuous.

Bratman’s worry about intentions rendering irrational actions rational is thus overstated.
Intentions cannot render irrational actions *objectively* rational. Intentions can render irrational actions *subjectively* rational, but this ought not be a surprise. Whether some action is subjectively rational depends on the agent’s mental states at the time of deliberation. Forming an intention is akin to changing one’s own mental states. Why should we be surprised, then, when changing one’s mental states results in a change in what is subjectively rational to do? We would not be surprised to find the outcome of subjective deliberation altered if we were to change an agent’s beliefs, or an agent’s desires; similarly, we should not be surprised to find the results of subjective deliberation changed when we change an agent’s intentions.

There is a more pernicious worry lurking beneath the worry about bootstrapping. If subjective deliberation is supposed to give us our best available approximation as to the real reasons out there, then why should we think that our intentions have any effect on the real reasons? Changing our beliefs is different; when we change beliefs, we might thereby render them less justified, and therefore less trustworthy as a way of picking out an agent’s real reasons. There is no such justification component to intentions, however, and so it seems as though changing intentions might lead the deliberating agent very far afield from where the true reasons lie.

The worry presupposes that all reasons are to be found outside of the deliberating agent, however. Once we accept that agents’ own mental states can provide them with reasons to act, in addition to giving them justified beliefs about what reasons there are to act, then we can see that there is no problem with intentions providing agents with reasons to act.

### 4.5 Rational deliberation over intentions

If intentions are reason-providing mental attitudes, then it is natural to wonder: what sorts of reasons are there for forming or modifying intentions? When is it rational or irrational to adopt an intention, or remove an intention, or modify an intention? Even if intentions are reason-providing mental attitudes once they are possessed, it seems that an agent still might be irrational in forming an intention, as for example when an agent forms an intention to perform an action which is clearly irrational from the agent’s lights.

There is a starker way to see the problem. Intentions, we argued, help to determine what is rational for an agent to do; intentions constrain an agent’s future deliberation. Intentions are also outputs of deliberation. But if intentions constrain future deliberation,
and intentions are also the outputs of deliberation, then it seems as though it is impossible for a decision problem to ever be resolved irrationally. After all, the output of the decision problem will be rationally binding on the agent’s future deliberation, and will determine what is rational for the agent to do in the future. It thus seems to be impossible for an intention to ever be irrational.

This, however, seems quite odd. If this were correct, this would be an even more worrisome recurrence of the bootstrapping problem; not only could intentions sometimes make irrational behaviors rational, but in fact it would be impossible for an intention to ever be irrational. Moreover, it would be impossible to ever resolve a decision problem irrationally. This would be a depressing result. After all, what good is a decision theory if every choice that one makes is ipso facto rational?

Intentions can be irrational, however, and decision problems can be resolved irrationally. What we need is an account of how this is possible, given the fact that intention rationality is a viable rationality concept.

In one sense, all rational evaluation of actions is the evaluation of intentions. After all, the output of a decision problem is not an action but is rather an intention (or choice), and so when we rationally evaluate deliberation we are evaluating not the action but the choice. It should thus not seem odd that it is possible to rationally evaluate intentions; indeed, properly speaking, the only thing we can evaluate is an agent’s intention.

Saying this, however, obscures an important distinction in the kinds of intentions that one might form as outputs of deliberation. Let us draw a distinction between determinate intentions and non-determinate intentions. Determinant intentions uniquely determine an action for an agent to perform; they leave the deliberating agent with no need to engage in further deliberation. Non-determinate intentions, on the other hand, do not uniquely determine an action; they allow for multiple actions which might satisfy the intention, and they leave the deliberating agent with a need to engage in further deliberation. In a chess match between X and Y, if X forms an intention to play the Ruy Lopez opening, that is a determinate intention; X has no further need to deliberate about what moves to play (up to a certain point).\footnote{This intention does, of course, still require some cooperation on the part of Y. If Y does not do his part—if Y evades the Ruy Lopez by playing a Sicilian defense, or a French defense—then X’s intention is frustrated after the first move. The fact that X’s determinate intention might be rendered unfeasible through Y’s action does not make X’s determinate intention any less feasible.} If, on the other hand, X forms an intention to capture all of Y’s pieces before checkmating him, that is a non-determinate intention; although the intention...
does structure $X$’s deliberation by taking some options off the table (either immediately or during the course of the game), it does not structure her deliberation so completely as to leave her with no further need to deliberate about what to do.

We have already mentioned that we cannot calculate the overall rationality of $X$’s intentions without possessing a theory of overall rationality. This applies equally to both determinate intentions and to non-determinate intentions. We do not have a theory of overall rationality. At this point we do have two theories of rationality: a theory of preference rationality, in which an agent is preference rational insofar as she or he maximizes her or his preference function, and we have a theory of intention rationality, in which an agent is intention rational insofar as she or he maximizes her or his preference function after being filtered through her or his intentions. (We do not presume that these two theories of rationality exhaust all there is to say about overall rationality.) Accordingly, we can calculate the preference rationality and the intention rationality of adopting or modifying one’s intentions, whether those intentions are determinate intentions or non-determinate intentions.

It is most straightforward to determine the preference rationality of a determinate intention. A determinate intention yields a unique action for the intending agent, and the agent holds beliefs about which outcome is likely to be produced by that action. The determinate intention is preference rational, then, insofar as the outcome corresponding with the intended action maximizes the agent’s preference ranking over outcomes. This corresponds with the traditional picture of rational choice as the maximization of one’s preferences, where we understand the output of rational choice to be an intention and not an action.

We can evaluate the intention rationality of a determinate intention as well. A determinate intention is intention rational so long as the outcome corresponding with the intended action maximizes an agent’s preference function given the set of intentions that the agent holds at the time of deliberation. Note that determinate intentions do not render themselves intention rational, except in a trivial sense: if a determinate intention $\phi$ is part of the input of deliberation, and is also the output of deliberation, then the fact that $\phi$ is a rational input to deliberation does indeed render $\phi$ rational as an output of deliberation. All this says, however, is that an agent who has made up her or his mind to perform some action ought, absent further deliberation, to perform that action. Deliberating with intention $\phi$ as input is not the same as reopening deliberation on whether the intention to $\phi$ is rational! If agent $X$ were to reopen deliberation on $\phi$, then $X$ would need to have other alternatives.
in the decision problem. A rational agent cannot legitimately deliberate about whether to \( \phi \) while simultaneously intending to \( \phi \).

Determining whether a non-determinant intention is preference rational or intention rational is slightly more difficult. Complicating our calculations is the fact that, with non-determinant intentions, the intention is rational or irrational according to whether the expected solution of the subjective decision problem that results from forming the non-determinant intention is better or worse than the expected solution of the subjective decision problem without the non-determinant intention. In the case of preference rationality, we compare the two solutions according to the agent’s preferences; in the case of intention rationality, we must first filter the decision problem through the agent’s intentions (including the new non-determinant intention).

For non-determinant intentions, there is a set of actions compatible with the intention that an agent can play. Combined with the set of actions that other agents can play (as determined by the deliberating agent’s beliefs about the other agents’ intentions and actions), non-determinant intentions yield a new subjective decision problem which is a subset of the original subjective decision problem. This new subjective decision problem either has a solution, or it does not. If the new subjective decision problem has a solution, then the non-determinant intention is preference rational so long as the solution to the new subjective decision problem is more preferred than the solution to the original subjective decision problem. (The original subjective decision problem might not have a solution, in which case the non-determinant intention is rational.)

If, on the other hand, the new subjective decision problem does not have a solution, then the non-determinant intention is only preference rational so long as every available outcome of the new decision problem is more preferred than the solution (if any) of the original subjective decision problem. This can occur when, for example, an agent forms a non-determinant intention which eliminates from rational consideration a unique but sub-optimal equilibrium, producing a game in which every available outcome is preferred to the original sub-optimal equilibrium. If every outcome of the new decision problem is not preferred, and the new decision problem does not have a solution, then the non-determinant intention is not preference rational.

The intention rationality of non-determinant intentions is a bit more tricky. We can judge a non-determinant intention to be intention rational in two ways. The first way is similar to that discussed immediately above with preference rationality; the non-determinant
4.5. RATIONAL DELIBERATION OVER INTENTIONS

intention might produce a game which has a solution which is at least as preferred as the solution of the original game after being filtered through all the agent’s intentions. Call this strict intention rationality. However, this is not the only way for a non-determinant intention to be intention rational. The resulting game does not need to yield a game with a solution which is more preferred than the original solution after being filtered through the agent’s intentions. The reason is that the agent’s deliberation yields an intention, and that intention then acquires the status of being reason-providing for the agent’s future deliberation. So long as the non-determinant intention is merely compatible with the agent’s previously held intentions, then the agent’s intention can be intention rational. Call this second way loose intention rationality.

The contrast between strict intention rationality and loose intention rationality is not what we might expect it would be. An intention is strict intention rational so long as forming the intention produces a game with a solution at least as good as the solution that the agent originally faced. In other words, strict intention rationality does not require that the solution to the new game be strictly better than the solution to the original game. Any confusion on this point can be allayed if we compare strict intention rationality with loose intention rationality. An intention is loose intention rational so long as the intention is not incompatible with any of the agent’s previously held intentions. It must, in other words, be possible for an agent to fulfill both the new intention and the agent’s previously held set of intentions. Strict intention rationality appeals to ‘better’ outcomes—an intention is strictly intention rational so long as the solution of the resulting game is better (or at least as good) as the solution of the game that the agent would face without the intention. Loose intention rationality, on the other hand, makes no such appeal to a better solution; an intention is loosely intention rational so long as there still exists a solution to game produced by forming the intention.

We can understand loose intention rationality better by thinking about the nature of partial plans or partial decisions. Agent X might have a set of decisions to make in pursuit of some accepted end. She might first have to decide between two actions, a or b. If she decides to perform action a, then she must decide between performing a₁ or a₂; if she decides to perform b, she must decide between b₁ and b₂. Suppose that X’s preferences favor performing a to performing b, and that she prefers to perform a₁ to a₂ and b₁ to b₂. If, as a result of deliberation, X forms the intention to do a, then her intention is both preference rational and intention rational. It is preference rational because the choice between a₁ and
a_2 has a preference-maximizing solution, namely ‘perform a_1’. And the intention is strict intention rational because her intention to do a produces a game (namely, choose a_1 or a_2) in which the solution (perform a_1) is at least as preferred as the solution to the original game (which, in this simple example, is also to perform a_1).

Now suppose that X intends to do b. Her intention is certainly not preference rational; by intending b, X is assured of producing an outcome which is less preferred than the solution of the original choice situation (namely, a_1). And her intention is not strictly intention rational, because the game which results from the intention (namely, choose b_1 or b_2) yields a solution which is less preferred to the solution of the original game that X faced. However, X’s intention to do b is loose intention rational: X’s intention to do b is compatible with her previous intentions, and so produces a decision problem which X can solve.

Loose intention rationality is an important concept, because when X adopts the intention to do b and then faces a decision of whether to do b_1 or b_2, X can rationalize intending to do b_1 by noting that it is intention rational. (Because the intention to do b_1 is a determinant intention, we do not face the question of whether the intention is strictly or loosely intention rational.) Once X forms the intention to do b, that intention reveals something about X’s values, and so it is a bit misleading to say that the intention is irrational in the decision problem that X faces between b_1 and b_2. It is true that forming the intention was not strictly intention rational. It was weakly intention rational, however, which just means that it is possible for the agent to resolve her subsequent decision problem in a way which is rational given the new constraints that she has imposed on herself.

I do not mean right now to arbitrate between these different rationality concepts. One of the benefits of cataloguing these various rationality concepts is that it enables us to identify in exactly which ways an intention (whether determinate or non-determinate) is rational or irrational. An intention might be preference irrational but intention rational; likewise, an intention might be preference rational but intention irrational. In the case of non-determinant intentions, the intention might be strictly intention rational or loosely intention rational. The weight that we accord these rationality concepts determines whether an intention is rational or irrational for an agent to form, and whether an action is rational or irrational for an agent to play.

The leeway is important, however. It is important in particular when we talk about collective actions; collective action sometimes occurs even though the collective action is
preference irrational for all of the participants. That agents act in a collective is not a guarantee that the action performed be a rational one! To see this, however, we must have a theory of collective action on the table. It is to this task that we turn in the following chapter.
Chapter 5

Collective goals and deliberation

5.1 Introduction

The theory of collective action that I want to develop here is an extension of what we can call the standard analysis of collective action: namely, agents are engaged in a collective action when they do their part of a collective goal.\footnote{The standard analysis is not universally accepted, of course. An interesting argument against comes from Susan Hurley in Hurley 2005 and, to a lesser degree, in Hurley 1989: Hurley argues that collective actions do not require collective goals at all. We will not consider Hurley’s argument in depth here; it suffices to say that, even if Hurley’s argument is correct, this only suggests that Hurley and I are analyzing different phenomena—my interest here is in deliberative collective action.} Advocates of the standard analysis face a challenge: namely, how are we to understand the collective goal? It seems as though there are two possibilities. The first possibility is that the content of the collective goal is collectively inclusive—that is, it includes or makes reference to actions of each of the members of the collective. Suppose $X$ and $Y$’s collective goal is to paint a house together. If ‘painting the house together’ makes reference to actions both of $X$ and $Y$—for example, $X$ paints half of the house, while $Y$ paints the other half—then the collective goal is collectively inclusive. The second possibility is that the collective goal is collectively exclusive—that is, it does not make reference to actions of each of the members. If we understand ‘collective goals’ in a collectively exclusive way, then $X$ and $Y$ each hold different collective goals: $X$’s collective goal might make reference to her painting half of the house as her part of the collective project, while $Y$’s collective goal might make reference to his painting half of the house as his part of the collective project.

Under the collectively inclusive interpretation of the collective goal of painting the house...
together, both \( X \) and \( Y \) hold the same collective goal: namely, ‘\( X \) paints one half and \( Y \) paints the other half’. Under the collectively exclusive interpretation, \( X \) and \( Y \) each hold \textit{different} collective goals; \( X \)’s collective goal is ‘\( X \) paints one half of the house, leading to the house’s being painted’ while \( Y \)’s collective goal is ‘\( Y \) paints one half of the house, leading to the house’s being painted’.

We should say more here about what it means for goals to make reference to agents’ actions. We will make the argument below that goals and intentions are functionally identical in the context of deliberation; if this argument is correct, then goals make reference to agents’ actions in the same way that intentions do (whether those intentions are collective or individual). We can understand this as saying that goals make reference to the actions of some agents insofar as they constrain the deliberation and actions of those agents.

To return: both ways of understanding collective goals—as collectively inclusive or as collectively exclusive—are fraught with problems. If we analyze collective goals as collectively inclusive, then we run the risk of violating one very plausible constraint on intentions, namely that agents can only intend to perform their own actions, and they can only intend to produce outcomes that they directly control.\(^2\) The constraints are different depending on whether the intention is an action intention or an outcome intention, but we can subsume both constraints under the general term \textit{own action constraint}. Intentions, in other words, have an own action constraint: agents can only intend their own actions. If we analyze collective goals as collectively inclusive, then it is hard to see how they satisfy the own action constraint. If \( X \) intends the collective goal of her and \( Y \)’s painting the house, then it would seem as though she is intending both her own actions and also \( Y \)’s actions, and this violates the own action condition.

If, on the other hand, we analyze collective goals in a collectively exclusive fashion, then it is hard to see where the collective content comes in. If \( X \)’s collective goal makes no reference at all to \( Y \)’s actions, then it would seem as though \( X \)’s goal is not a \textit{collective} goal at all. If \( X \) intends to paint half of the house, why should we not analyze this as a purely individual goal, rather than a collective goal? The instinct is to reply that \( X \) is painting half of the house as \textit{her part of the collective goal}, but then we still lack a proper analysis of collective goals, which was what we set out to do in the first place.

There is a further problem with a collectively exclusive analysis of collective goals. If

\(^2\)Bratman discusses the own action condition in Bratman 1999c. See also Velleman’s discussion in Velleman 1997.
collective goals are collectively exclusive, then it is a bit imprecise to say that agents in a collective share some common collective goal. Rather, each agent has her or his collective goal, which contains as a part her or his individual contribution to the collective goal. If we think that collective goals must be shared among the members of a collective, then we should be worried about collectively exclusive analyses of collective intentions; it is harder to say with precision what the collective goal is that collectives pursue.

We have seen both collectively inclusive and collectively exclusive analyses of collective goals already. Bacharach, Sugden, and Gold advocate a collectively inclusive analysis of collective goals, as did Michael Bratman. Searle and Tuomela analyze collective goals exclusively. All of the authors propose solutions to the difficulties that their analyses face, but as we have seen, none of the solutions are entirely successful as they stand.

It is particularly important that we distinguish between Bratman’s account of collective action and the one that we present here, since both theories are we-intention theories of collective action. As we saw in Chapter 3, Bratman’s we-intention theory violates the own action condition because of his theory of intentions: Bratman’s intentions are mental states which entail a positive commitment to perform the content of the intention, and since actions of other agents are part of the content of Bratman’s we-intentions, agents who intend collective actions also thereby intend actions of other agents in the collective. Like Bratman, I propose that agents in a collective action each possess intentions which contain as part of their content the collective action. I must, then, be able to explain whether the intentions that each agent possesses makes reference to the collective action in a collectively inclusive fashion, or a collective exclusive fashion.

In this chapter, I propose an account of collective goals which is collectively inclusive—collective goals do make essential reference to the actions of all of the agents in the collective. In spite of this, I argue that our account of collective goals does not run afoul of the own action condition. The own action condition imposes plausible psychological restrictions on what the content of individual intentions can be, and I argue that our analysis of collective goals does not violate those plausible restrictions. The key, then, will be to see how agents can conceive of their actions as part of a collective project while remaining their own actions. I first talk about what we mean by ‘goals’, and how goals are related to intentions. I then go on to talk about how agents solve decision problems when they hold individual goals. Individual goals can be understood either as intentions to perform some specific action, or as intentions to bring about some outcome; I argue that collective goals must be understood as
outcome intentions. Conceiving of collective goals as outcome intentions can, I suggest, help us to understand how agents hold collective goals, how collective goals structure individual deliberation, and how collective goals do not run afoul of the own action condition.

After establishing what collective goals are and how they structure individual deliberation, I go on to make some further points about collective goals. One thing we will do is talk about rationality conditions for possessing and revising collective goals. Another is that we will talk about about how holding collective goals promotes, but does not guarantee, coordination. And, finally, we will talk about the distinction between collective goals and what I call *common goals*.

### 5.2 Intentions and goals

As we saw in Chapter 4, intentions embody the two ‘faces of intention’ that Bratman identifies. Intentions serve as inputs to deliberation, and as such they impose rational constraints on deliberation over future actions. Intentions also serve as outputs of deliberation, however, and as such they represent a commitment in the present on the part of the deliberating agent to perform some action or bring about some outcome. As Bratman has noted, these two faces of intention can come into conflict, especially when we consider the relationship between intentions to act and intentional actions. Our philosophical attempt to reconcile the two concepts requires us, Bratman argues, to draw a distinction between *intentions* and *goals*—notably, intentions are subject to strong consistency requirements, while goals are not.

To see this, let us consider Bratman’s original argument in which he argues for the need to posit a distinction between goals and intentions. According to Bratman, an agent $X$ might have two goals which she knows to be incompatible. In Bratman’s original example, $X$ tries to strike two targets in a computer program simultaneously with her left hand and right hand, while knowing that, because of the design of the program, the program will not permit her to hit both. $X$ might aim to hit both targets, however, with the belief that by doing so, she is maximizing her chances of hitting one target or the other. If $X$ successfully hits a target—target 1, say—then $X$ must have hit the target intentionally; after all, $X$’s behavior was guided by her aim of hitting target 1. We cannot say, however, that $X$ has an

---

3Bratman 1984.

4See Bratman 1984 and Bratman 1987.
intention to hit target 1; doing so would require us to also admit that \( \text{X} \) has an intention to hit target 2 (since \( \text{X}' \)'s actions are guided as much by her aim of hitting target 2 as they are by her aim of hitting target 1), but \( \text{X} \) cannot hold these two intentions simultaneously, since \( \text{X} \) knows that the two outcomes are not compossible.

Bratman uses this example to show that \( \text{X}' \)'s having intentionally performed some action \( x \) does not necessarily entail that \( \text{X} \) had an intention to do \( x \); and this argument, as far as it goes, seems correct. What remains less clear is the status of the volitional attitude that \( \text{X} \) has towards performing those actions that lead towards striking target 1, and those actions that lead towards striking target 2. Let us say of our example that \( \text{X} \) holds a goal of striking target 1, and that \( \text{X} \) holds a goal of striking target 2.

It is clear so far that intentions are not simply goals; \( \text{X} \) is not irrational in simultaneously holding the goal of striking target 1 and the goal of striking target 2, although \( \text{X} \) would be irrational were she to simultaneously hold the intention of striking target 1 and of striking target 2. But what, exactly, is a goal? Bratman does not say. We can only conclude that, whatever goals are, they are not subject to the consistency requirement that intentions are.

The consistency requirement matters, because it is this consistency requirement that accounts for the fact that intentions provide a filter of admissibility over options. \( \text{X}' \)'s intention to do \( x_1 \) filters out of \( \text{X}' \)'s decision problem any action or outcome which is incompatible with her doing \( x_1 \), because of the consistency requirement; if it were not irrational for \( \text{X} \) to bring about outcomes incompatible with her intention to do \( x_1 \), then \( \text{X} \) would have no reason to exclude those outcomes from her rational deliberation. Goals, insofar as they are not subject to this consistency requirement, do not filter options in the same way. This point will become especially relevant when we attempt to build a formal model of deliberation, as our model cannot treat intentions and goals in the same way.

We would like to treat goals and intentions similarly in our model, however. Not only will it greatly simplify our model, but goals and intentions are functionally very similar: they are both volitional pro-attitudes, and they both provide some constraints on deliberation. It is important to note that, though they both provide constraints, the constraints are different. As we have seen, intentions provide negative constraints on deliberation; an agent who holds an intention to \( x \) (whether \( x \) is an action or an intended outcome) thereby eliminates from rational consideration those actions or outcomes which are incompatible with the agent’s doing \( x \). Goals, on the other hand, do not seem to impose negative constraints on deliberation. An agent who has the goal of hitting target 1 does not thereby rule out those
actions which are incompatible with her hitting target 1. In Bratman’s example X holds a goal of hitting target 1, but is also rationally considering actions which are incompatible with her hitting target 1 (namely, those actions which lead to her hitting target 2).

What holding a goal seems to do is to impose strictly positive constraints upon deliberation. That is, holding a goal of doing \( x \) or producing outcome \( o \) requires that an agent take those effective means that promote doing \( x \) or producing \( o \). If my goal is to strike target 1, then in my deliberation on behalf of that goal I must determine which means are best suited to bringing about my goal, and then select those means. This is a positive constraint and not a negative constraint, because it tells me which actions I \textit{should} take, rather than which actions I \textit{shouldn’t} take. The negative constraints that are entailed by one’s intention are constraints over which actions an agent \textit{should not} take.

Now, it is true that intentions seem to entail both positive and negative constraints. If \( X \) intends to bake a cake in the evening, then \( X \) thereby incurs two kinds of constraints: she has both negative constraints, in that it must be irrational for her to do any actions which are incompatible with her baking the cake (such as using up the last of the milk), and she has positive constraints, in that she must determine some positive means of satisfying her intention of baking the cake (such as deciding what kind of cake to bake, deciding where to get any missing ingredients, and so forth). Intentions appear to be distinct from goals, then, in that intentions have \textit{both} positive and negative constraints, while goals only present positive constraints.

I have said ‘seems’ throughout, because we can in fact model intentions and goals in the same way, in spite of what appears to be differences in the kinds of constraints that the two mental states offer. We often talk colloquially about being ‘single-minded’ in pursuit of some goal or other, and I think that this talk accords with a productive way of modeling goals: a goal is a very restricted type of intention, such that insofar as one is deliberating and acting on behalf of a goal, one evaluates options \textit{only} with respect to whether they promote the goal. Goals are, in effect, localized intentions; they are intentions to bring about some outcome, but they are intentions which pursue an outcome completely independent of other mental states or attitudes that the agent might have. And, insofar as goals are localized intentions, they can be modeled as providing purely negative constraints—holding a goal rationally requires the deliberating agent to disregard every action that does not promote the goal outcome \textit{insofar as the agent is deliberating about that goal}. Because goals are localized, however, the goal does not continue to impose those negative constraints while
the goal is not actively pursued.

Consider X’s goal of hitting target 1. X’s goal might structure her deliberation, in that she has a goal of hitting target 1, though she still needs to engage in further deliberation about the best means to take to hit the target. (Should she aim directly for the target? Should she try to corner the target first? Should she engage in random firing?) Insofar as she is deliberating on behalf of her goal, she is considering as irrational every action or outcome which is incompatible with hitting target 1; once she has eliminated all incompatible options, she can evaluate the remaining options according to how well they promote the production of the goal.

X’s goal of hitting target 1 is just a goal, however, since the negative constraints are removed whenever she is not actively deliberating on behalf of her goal. When she is deliberating about her goal of hitting target 2, for instance, she does not take as rationally binding the negative constraints imposed by her goal of hitting target 1; insofar as she is actively deliberating on behalf of her goal of hitting target 2, she considers only those options which promote the hitting of target 2.

It might be objected that, in Bratman’s original example, the agent is not deliberating on behalf of the goals independently; rather, the agent is deliberating and acting on behalf of both of her goals simultaneously. There are two ways that we might understand this objection. If the objection is merely that the agent in Bratman’s example is engaging in two separate instances of deliberation (one directed towards the goal of hitting target 1, the other towards the goal of hitting target 2) at the same time, then the objection is without teeth; nothing we have said so far precludes the possibility that an agent might engage in two different instances of deliberation at the same time. This seems to be based upon the thought that we are only ever phenomenologically aware of deliberating over any one topic at a given time, and that this phenomenological awareness reflects the truth of our deliberation.

On the contrary, however, it seems that we can and do engage in deliberation about distinct things at the same time; and even if we are not actively aware of the contents of one of our deliberative processes, we still possess the ability to bring that deliberative process to the fore of our conscious perception should we need to. This first way of understanding the objection, then, doesn’t pose a real challenge; agents can deliberate about two distinct things at the same time, and so the fact that the agent is doing so in the video games case is not exceptional.
5.2. INTENTIONS AND GOALS

The other way of understanding the simultaneity worry is that there might be simultaneity of outcome. This objection is that the agent in the video games example must be deliberating about one common outcome, rather than about two outcomes at the same time. Unlike the first worry, this worry does indeed pose problems for our proposed interpretation of goals. It is not necessarily irrational for an agent to hold two incompatible goals at the same time; nor is it necessarily irrational for an agent to act towards two incompatible goals at the same time. (As Bratman’s example shows us, acting towards two incompatible goals is sometimes the best way to bring about some other, higher-level goal.) However, it is irrational for an agent to deliberate towards two irrational goals at the same time, if these deliberative processes are not isolated from one another. (In the previous paragraph we looked at simultaneous deliberation towards incompatible goals, but the deliberative processes were isolated from one another.)

Let us call two processes of deliberation coextensive if their goal outcomes are represented in the processes of deliberation as being concurrent—that is, if the two goal outcomes are viewed as one goal. The agent’s deliberation, then, could be represented as one process of deliberation towards one goal. I argue that it is irrational for an agent to engage in coextensive deliberation over incompatible goals. If this is true, then this would pose a challenge for the agent in Bratman’s video game example if we interpret the agent as necessarily engaging in coextensive deliberation. I will argue, however, that rational agents never engage in coextensive deliberation over incompatible goals, and that we can always interpret such deliberation as the benign simultaneous but non-coextensive deliberation analyzed above.

The agent who holds a goal either knows what actions she or he intends to take so as to produce the goal outcome, or the agent does not. It is only in the latter case that deliberation is necessary; in the former case nothing is needed except for the agent’s performing the intended action. If the agent does not know which action she or he is going to perform so as to satisfy her goal, however, then the agent must engage in practical reasoning towards the goals independently of one another in order to properly evaluate the effectiveness of the available options. To see this, consider agent X, who holds two goals: the production of outcome \( o_1 \), and the production of outcome \( o_2 \). These two goals are incompatible. X knows that she cannot both produce \( o_1 \) and \( o_2 \), but she has not yet decided what action she will perform to bring about \( o_1 \); she sees her options as \( x_1^1 \), \( x_1^2 \), and \( x_1^3 \). Likewise, she has not settled on the best means to take to bring about goal \( o_2 \); her options here are \( x_2^1 \) and \( x_2^2 \). If
X were to engage in coextensive deliberation about the production of her goals, then she would be deliberating with the aim of producing the joint outcome $o_1 \cap o_2$—that is, towards the outcome which consists of $o_1$ and $o_2$ together. X’s deliberation is reasoning about which action best promotes the goal outcome, and so X would be deliberating about which of $x_1^1$, $x_2^1$, and $x_3^1$ best produce the outcome $o_1 \cap o_2$. Because $o_1$ and $o_2$ are incompatible, however, their intersection is the empty set; there is no possible scenario in which both outcomes are produced at the same time. All of the available options thus produce the joint goal outcome equally well, because they each completely fail to produce the joint goal outcome; the joint goal set is impossible to produce regardless of which action X chooses.

In ordinary cases of deliberation, whether X chooses to do $x_1^1$, $x_2^1$, or $x_3^1$ depends primarily on which of those actions best produces $o_1$—and this means that, in deliberating about those actions, X deliberates about them with only the goal outcome $o_1$ in mind. Also, when deliberating about which of $x_1^2$ and $x_2^2$ to do, X is considering only the extent to which either effectively produces the goal outcome $o_2$; she is not consider the constraints that her goal of producing $o_1$ imposes on her. If X’s deliberation about the goals $o_1$ and $o_2$ were simultaneous but not coextensive, then X could reason about which action best promotes $o_1$ independently of her reasoning about which action best promotes $o_2$. This is, in fact, what X ought to do—she maximizes her chances to hit target 1 if she chooses to perform those actions which are most conducive to hitting target 1, just as the same holds mutatis mutandis for target 2.

The objection thus misses the mark. An agent can rationally deliberate towards two incompatible goals at the same time, so long as the two processes of deliberation are distinct—that is, so long as the goal outcomes are kept distinct, so that the agent can evaluate the effectiveness of various actions towards bringing about the incompatible goal outcomes separately. It is true that an agent cannot rationally deliberate towards two incompatible goals coextensively; if the agent represents the two incompatible goal outcomes as one common outcome to be brought about, then the agent has no basis upon which to evaluate the effectiveness of any particular option in bringing about the goal outcome. However, this worry is unfounded; the agent who holds two incompatible goals simultaneously might be engaged in two distinct processes of deliberation, in which case there is no irrationality. Such is the case with the agent in Bratman’s example; the agent’s deliberation towards target 1 and target 2 might be distinct processes of deliberation.

Now, in arguing that incompatible goals which are simultaneously held by an agent
can be modeled as two decision problems with two distinct intentions, we must allow that
decisions which are made in one decision problem can still constrain the deliberation over
the other decision problem. Suppose X is applying to two colleges, college a and college
b, and she knows that the two schools coordinate their admissions: if she gets accepted to
either one, she will be rejected from the other. It is not guaranteed that she will be accepted
to either one, but X knows that she will maximize her chances of being accepted to one of
the schools if she applies to both of them. According to Bratman’s original argument, we
cannot say that X intends to get accepted into either college a or college b, since the two
intentions cannot be combined into one consistent plan; rather, we should conclude that
X has two goals. According to the argument that I presented above, X’s goal of getting
into college a can be represented as an intention which is limited in scope: insofar as X
is deliberating about the best means to satisfying her goal of getting accepted into college
a, she disregards her goal of getting into college b, and so for the purposes of deliberation,
X deliberates as though she has an intention to get into college a. X cannot disregard
everything about her incompatible intention to get into college b, however. Suppose both
colleges require X to give an interview, and that she gets to choose the interview date: she
can either select the early date or the late date. If X were to keep her two processes of
deliberation completely isolated from one another, then X would not be able to take her
request for one interview date with one college as imposing rational constraints upon her
deliberation for the other college, and she might agree to the same interview date with the
two separate schools.

We must, then, be able to build some of the practical conclusions of an agent’s delib-
eration into the agent’s other deliberations. In the case mentioned above, X should allow
her intention to interview early at college a to constrain her deliberation about when to
interview at college b, but X should not allow her intention to get accepted into college a to
constrain her deliberation about whether to apply to college b. Is there a principled way of
drawing the distinction between which conclusions are legitimate practical constraints and
which are not? It is not clear.\(^5\) However, even if we do not have a principled way of drawing
the distinction, we should be optimistic that, in each particular case, there is some way of
characterizing the agent’s decision problem such that the right constraints are included and
the wrong constraints are excluded.\(^6\)

\(^5\)One possibility is that the only practical conclusions which are not accepted are the initial incompatible
goals.
\(^6\)As a last resort, we can merge the two incompatible goals into one higher-order intention: X intends
I propose, then, that we treat goals as intentions which are limited in scope. A goal to bring about some outcome \( o \) is equivalent to an intention to bring about outcome \( o \) in the contexts of deliberation. Agents who have a goal bring about \( o \) deliberate as though they have an intention to bring about \( o \), since the goal and the intention restrict their further deliberation in the same ways. And, agents who intend to bring about \( o \) deliberate as though they have a goal of bringing about \( o \). If we need to model a deliberative situation such as Bratman’s video game case, we can do it in one of two ways: we can either isolate the two processes of deliberation, such that the deliberating agent first deliberates about how to strike target 1 and separately deliberates about how to strike target 2, or we can represent the agent’s intention-cum-goal as one of striking one of the two targets, without being specific about which one. Either way we choose to model the deliberative situation that the agent faces, we can model the constraints that the agent faces upon her or his deliberation as coming either from intentions or goals—the two attitudes can be treated identically.

5.3 Goals and rational decisions

We cannot give a complete theory of rational decision making in the presence of goals without having a theory of overall rationality. We cannot, in other words, know what effect the possession of a goal ought to have on an agent’s deliberation without first knowing whether the agent’s goal is a rational or is an irrational one.

We can bracket the concern about the rationality of an agent’s goals, however, if we focus our attention on intention rational agents. If we assume that agents are intention rational (or, equivalently, goal rational), then we need not worry about whether the production of the goal outcome is a rational thing or an irrational thing; we can focus our attention on whether the agent’s choices are rational given the assumption that the agent’s goals have rational force.

The assumption that agents are intention rational has two practical upshots. The first is that agents will enter into deliberation with a consistent set of intentions. Being intention rational requires that agents possess intentions which are jointly compossible; it must be possible for the intention rational agent to act in such a way as to bring about all of her

---

to get into either \( a \) or \( b \), for example. This solution would not serve Bratman’s original argument rejecting what he terms the Simple View, but it serves our purposes here perfectly well.
or his intended outcomes. The second upshot is that the output of deliberation must be compatible with an agent’s previously held intentions. When the output of an intention rational agent’s deliberation is a determinant intention, then the outcome produced by the determinant intention must be one of the outcomes permitted by the agent’s previously held intentions. When the output of an intention rational agent’s deliberation is a non-determinant intention, the non-determinant intention must be compatible with the agent’s previously held intentions—that is, all of the permissible outcomes in the non-determinant intention must be compatible with the agent’s previously held intentions.

Intention rationality, then, is akin to being means-end rational; the agent’s previously held intentions determine the agent’s ends or goals, and the agent’s choice must produce an outcome which is compatible with the agent’s ends. Insofar as there are multiple ways for the agent to satisfy her or his ends, then the intention rational agent will choose the best way to satisfy her or his ends. In this case, as argued above, the ‘best way’ means the way which best satisfies the agent’s preferences (qua motivational judgments of value).

It is worth repeating that in restricting our attention to intention rational agents, we make no claims whatsoever as to the underlying rationality of the agents’ intentions. Intention rational agents deliberate in order to best satisfy their intentions; they do not necessarily act rationally, because they do not necessarily hold rational intentions.

Restricting our attention to intention rational agents, then, deliberating agents face decision problems. The question is, what is the best way for an agent to resolve their decision problems given the set of intentions which serve as inputs to their deliberation? Answering this question requires us to draw a distinction between single agent decision problems (SADPs) and multi-agent decision problems (MADPs).

5.3.1 Rational resolution of SADPs

A single agent decision problem is a decision problem in which the deliberating agent believes that she or he is the only agent whose choice affects the outcome produced. This claim requires some unpacking. We should not understand this to mean that the agent believes that she or he is the only human being whose behavior is relevant for the production of an outcome. This interpretation is far too strong; it would rule out, for instance, situations in which one agent X deliberates about which shirt to buy from a store. Buying a shirt requires not only that X make a decision about which shirt is best, but it also requires that another person, Y, the merchant, sells the shirt to her. It will be foolish, however, to
conceive of the decision problem as one involving choices by two agents, the buyer and the seller; once X decides on a shirt to buy, the decision problem has been resolved. Although Y’s behavior is important for the production of the desired outcome, X does not factor Y’s decision making into her representation of the decision problem, since X is entitled to treat Y’s behavior as directly caused by X’s choice—X believes that, whichever shirt she buys, Y will sell her that shirt.

It is a bit too strong, however, to say that in single agent decision problems, agents must believe that the outcome produced is determined by their choice in the decision problem. This is too strong because it ignores the possibility that the outcome is produced through a stochastic process—that the agent’s choice might maximize the chance of a desired outcome’s being produced without determining it completely. Even in the stochastic case, though, we can draw a distinction between when the outcome produced is believed to be a function (whether determinate or probabilistic) of one agent’s choice, and when the outcome produced is believed to be a function of more than one agent’s choice. And now we can see that, in SADPs, the deliberating agent must believe that the outcome produced is a function only of her or his choice, and not of any other ‘agents’ in the decision problem. (The word ‘agents’ is put in quotation marks to indicate that, though they may be agents, they are not in fact exercising their agency in the decision problem—they are not considered to be choosing to act on the basis of reasons.)

For intention rational agents, then, whether a decision problem is rationally resolved will depend on solely on the agent’s choice. Because the agent’s intentions impose rational constraints on the agent’s deliberation, the agent will have resolved her or his decision problem rationally so long as the agent chooses that action which is expected to produce the most desired outcome which is compatible with her or his intentions.

This view has been presented above in Chapter 4. Besides the slightly expanded treatment of SADPs that we present in this chapter, the only interesting thing to add here is that intentions are equivalent to goals; this means that the intention rational agent who holds a goal ought to choose that action which is believed to best bring about the goal, where ‘best’ here means maximizing the agent’s preferences consistent with the agent’s intentions.

5.3.2 Rational resolution of MADPs

Whereas in single agent decision problems the outcome produced is believed by the deliberating agent to be a function of only the agent’s choice, in multi-agent decision problems the
outcome produced is believed to be a function of more than one agent’s choice. To put it briefly, MADPs are games—the same games which are the subject matter of game theory. As such, if we are to determine the rational resolution of a MADP, we should appeal to the solutions offered by game theory.

We will take the main solution concept of a MADP to be the equilibrium. Equilibria are outcomes in which no agent can produce a better outcome for herself or himself by changing her or his choice while keeping the choices of the other agents constant. For traditional games—that is, games in which agents do not hold intentions—agents are rational insofar as they choose actions which are believed to lead to the production of a unique equilibrium or, in the case of multiple equilibria, which lead to the production of a unique Pareto optimal equilibrium insofar as one exists.

When we introduce intentions, the picture becomes more complicated. The solution concept is still that of the equilibrium, but rational agents now must choose those actions which are believed to produce outcomes compatible with their intentions. Each agent in the MADP has intentions, and they also have beliefs about the intentions of other agents. We will make the assumption that all agents are intention rational, and that all agents believe the others to be intention rational. This greatly simplifies our use of intentions in decision problems; given this assumption, we can construct a new decision problem which consists only of those outcomes which are compatible with the intentions of every agent in the MADP. This set of outcomes will not necessarily be convex: there might be ‘holes’ in the strategic game when the outcomes and payoffs are represented as an \( n \times n \) matrix. In the resulting set of intention-compatible outcomes, there might be equilibrium outcomes; if there are, then agents have solutions to the decision problem insofar as they choose actions which lead to the unique equilibrium or to the Pareto optimal equilibrium, if such exists.

Because the set of intention-compatible outcomes need not be convex, it is possible for there to be an equilibrium outcome even though no agent can rationally change her or his choice should they find themselves in that outcome. This does not affect the rationality of acting towards an equilibrium outcome; it merely means that, should one agent fail to act towards the equilibrium, the outcome produced will conflict with some agent’s intention, and that agent will then necessarily be intention irrational. In non-convex sets of outcomes, an outcome is an equilibrium so long as no agent possesses an action different than the action that they have played such that, should they play their alternate action and the other agents play the same action, the agent produces an intention-compatible outcome.
which is more preferable for the agent than the outcome produced by the original set of actions.\footnote{See Appendix A for a more formal presentation of the results here.}

The general case is that the set of intention-compatible outcomes need not be convex. In such cases, the equilibrium is still the general solution concept that we will use for the decision problems that the agents face. There is an important subset of the general case, however, which is that the resulting intention-compatible outcome set is convex. This occurs when all of the agents in the decision problem form intentions to perform actions, rather than intentions to bring about some outcome. The distinction between the two cases is crucial, and we will look at this difference below.

### 5.4 Action intentions and outcome intentions

The distinction between action intentions and outcome intentions is an important one, and it has appeared occasionally in the literature. Tuomela, for instance, talks about ‘seeing to it that’, and contrasts this concept with what he calls ‘action intentions’.\footnote{Tuomela 2005.} Moreover, Tuomela seems to recognize that the distinction between action intentions and outcome intentions is crucial for understanding how collective reasoning is possible. Tuomela has not clearly explained how action intentions and outcome intentions function in a decision theoretic context, however. Olivier Roy is much better in contrasting action intentions versus outcome intentions in a decision theoretic context.\footnote{See Roy 2008.} Roy does not apply his theory of intentions to collective actions, however. In this section, I propose an account of action intentions and outcome intentions which is both compatible with a more extensive theory of deliberation, and is also tailored to an application in collective action.

Recall the account of subjective decision problems we presented in Chapter 4. For each agent in the decision problem there is a corresponding set of actions. Moreover, there is an outcome function which maps vectors of actions onto a set of outcomes. Actions contribute to the production of outcomes. There is, however, a bit of ambiguity in how we are to tell the story of action that our deliberative model proposes. We could think of agents as deliberating about which action to perform. Or, alternately, we could think of agents as deliberating about which outcome to bring about. Preferences, we noted, are defined over outcomes; for any two pair of outcomes, agents either prefer the first to the second, or prefer...
the second to the first, or they are indifferent between the two. This should lead us to think
that what agents are deliberating about is which outcome to bring about; after all, their
choice is rational to the extent that they maximize their preference function (subject, of
course, to certain constraints).

However, agents can only perform actions; they cannot directly bring about outcomes. In
SADPs in which the outcome function is deterministic and one-to-one, the two are identical:
for each action there corresponds exactly one outcome. Now, it is possible that there are
outcomes to which there corresponds more than one action, or no actions at all. (This is
not the usual way of representing decision problems, but it is possible.) Nevertheless, it is
always true in SADPs with deterministic one-to-one outcome functions that there is some
outcome which can be produced, and which maximizes the agent’s preference function,
such that the agent has a set (perhaps a singleton set) of actions which correspond to that
maximal outcome.

In such cases, it does not matter whether we define intentions to be action intentions
or outcome intentions. Let an *action intention* be an intention to restrict one’s choice to
a proper subset of the actions available in the agent’s subjective decision problem. Cor-
respondingly, let an *outcome intention* be an intention to restrict one’s choice to a proper
subset of the outcomes available in the agent’s subjective decision problem. My claim,
then, is that in SADPs with deterministic one-to-one outcome functions, it does not matter
whether we characterize agents’ intentions as action intentions or as outcome intentions.
The reason is that for every action intention we can construct a corresponding outcome
intention (since every action corresponds with some unique outcome), and for every feas-
ible outcome intention we can construct a set of action intentions between which the agent
ought (intention) rationally be indifferent. (I say ‘feasible’ outcome intention here to indi-
cate that we are only interested in those outcomes which an agent can actually bring about
by performing some action.)

So, then, if an agent $X$ in a SADP faces a decision problem in which there are two
actions, $x_1$ and $x_2$ and two outcomes, $o_1$ and $o_2$, and the outcome function is such that
performing $x_1$ leads to $o_1$ and performing $x_2$ leads to $o_2$, then it does not matter whether
$X$ has the action intention to do $x_1$ or the outcome intention to bring about $o_1$.

The same is true in more complicated choice situations. Suppose $X$ faces a SADP in
which there are three outcomes, $o_1$, $o_2$, $o_3$, and three actions, $x_1$, $x_2$, and $x_3$. Suppose,
however, that both $x_1$ and $x_2$ produce the outcome $o_1$, suppose that the action $x_3$ produces
the outcome \( o_3 \), and suppose that no action produces the outcome \( o_2 \). In this case, \( o_2 \) is not a feasible outcome, and so \( X \) cannot rationally hold an outcome intention to produce \( o_2 \) even if the outcome is the most preferred. \( X \)'s action intention of playing \( x_1 \) is equivalent to an outcome intention to bring about \( o_1 \). And, likewise, \( X \)'s action intention of playing \( x_2 \) is equivalent to an outcome intention to bring about \( o_1 \). This means that \( X \)'s outcome intention of bringing about \( o_1 \) is equivalent to both the action intention \( x_1 \) and the action intention \( x_2 \), which is fine—the two actions belong to an equivalence class, among which \( X \) ought rationally be indifferent.

Such is the case for SADPs. In MADPs, on the other hand, we have to treat action intentions and outcome intentions very differently. The reason is that in MADPs, no individual agent can produce an outcome by herself or himself. This is built into the definition of MADPs; if the production of an outcome is believed to be the function of just one agent’s choice, then the choice situation is not a MADP, but is rather a SADP. And since there is no direct correlation between the performance of some action by an agent and the production of some particular outcome, neither can we say that there is a direct correlation between an agent’s holding some action intention and that agent’s holding some outcome intention.

We must, then, interpret action intentions in MADPs as intentions to perform some action—but not as an intention to bring about some particular outcome. Rather, the agent intends to produce one of a set of outcomes, those outcomes being determined by the outcomes produced by the range of actions that the other agents have available to them to perform. As an example, consider a choice situation with two agents, \( X \) and \( Y \), each of whom have three available actions, \( x_1, x_2, \) and \( x_3 \) and \( y_1, y_2, \) and \( y_3 \) respectively. There are, accordingly, nine possible outcomes: \((x_1, y_1), (x_1, y_2), \ldots, (x_3, y_3)\). If \( X \) forms the action intention of performing either \( x_1 \) or \( x_2 \), then \( X \) intends, in effect, to constrain the decision problem so that there are six possible outcomes compatible with her intention: \((x_1, y_1), \ldots, (x_2, y_3)\). If \( Y \) were also to have an action intention—the action intention of doing \( y_3 \), for instance—then the choice situation would be reduced even further, down to two possible outcomes: \((x_1, y_3)\) and \((x_2, y_3)\).

With action intentions, then, the agent \( X \) who intends some action \( x \) can be said to intend all those outcomes \((x, z)\) such that \( z \) is a feasible set of actions for the other agents in the MADP to perform. This claim might seem a bit extreme. After all, if I am in a choice situation—a prisoner’s dilemma, say—and I intend to perform some action—cooperate, say—then it seems incorrect to claim that I intend to bring about either the outcome in
which I cooperate and you also cooperate, or the outcome in which I cooperate and you
defect. If my intention to cooperate makes reference to your action somehow, then it seems
foolish of me to intend to bring about an outcome which guarantees my own personal ruin.
If my intention to cooperate does not make reference to your action, then it seems as though
my intention is only to perform some action, and not to bring about some outcome.

Though the claim seems extreme, it is not. As we mentioned above, there are two
types of intentions that an agent in a MADP might have: determinant intentions and non-
determinant intentions. Determinant intentions are intentions that uniquely determine an
action for the intending agent to perform, while non-determinant intentions do not uniquely
determine an action. Agents who hold determinant intentions (as my intention to cooperate
in a prisoner's dilemma would be) are, in a sense, abdicating some of their ability to choose
the outcome that is produced. What they are doing is presenting the other agents in
the MADP with a modified decision problem and asking those other agents to choose the
outcome which results. The reason is that this determinant intention is unconditional; it
does not depend on the actions of the other agents in the decision problem. As a result,
any outcome which the other agents in the MADP produce must be compatible with the
agent's intention, or else it would be irrational for her or him to form the intention in the
first place. If I intend to perform some action while knowing that the outcome which results
depends on my action and also on your action, then I am essentially declaring that I am
rationally satisfied with whatever outcome you choose to bring about. If I were not, then I
ought not form an action intention in the first place.

Non-determinant intentions are similar to determinant intentions in this way. If I form a
non-determinant intention when faced with a MADP, then by declaring that certain actions
are irrational for me to do, I am also declaring that all those outcomes which result from
my doing those actions are also irrational for me to bring about. My intention to not defect
means that I must find all outcomes to be irrational in which I defect; if not, it would not
be rational for me to form or hold the intention.

In one sense, then, action intentions can be translated into outcome intentions. If agent
X intends to perform some action x, or one of a set of actions A, then we can say that
X holds an outcome intention including all those outcomes in which X performs x, or in
which X performs one of the actions contained in the set A. The colloquial way to put it
would be to say that when X holds an action intention, X also holds an outcome intention
to bring about all those outcomes which might result from X's doing the intended action.
The same does not hold in the other direction, however: outcome intentions cannot in general be reduced to action intentions. In MADPs, outcomes require the contribution of more than one agent in order to be produced. To return to the example of the prisoner’s dilemma, there are four outcomes in prisoner’s dilemma: mutual cooperation, mutual defection, and two non-coordinative outcomes. Outcome intentions could be any proper subset of that set of four outcomes. Suppose X faces just such a prisoner’s dilemma, and holds the outcome intention of producing one of the two coordinative outcomes: coordination upon mutual cooperation, or coordination upon mutual defection. This intention—bring about $(C,C)$ or $(D,D)$—cannot be translated into an action intention, because there is no action X can play which corresponds to the production of one of the actions in the set. What X should have is an intention to do C if Y is doing C, and an intention to do D if Y is doing D.

In general, if an agent X has an outcome intention to bring about some outcome o, where her contribution to o is to perform action x and the contribution of the other agents in the MADP is to perform the set of actions A, then that outcome intention cannot be translated into an action intention unless the agent also intends to bring about every other outcome $o'$ which would be produced by the agent’s continuing to play action x but where the other agents play the set of actions $A'$, where $A'$ is every feasible set of actions available to the other agents.

In MADPs, then, holding an outcome intention requires an agent to intend to produce some specific outcome, whereas holding an action intention requires the agent only to intend to perform some specific action. The latter is within the agent’s control; it is something that the agent can be certain that she or he can do, and so the agent has confidence that she or he can bring satisfy her intention through her actions. In the case of those outcome intentions which cannot be translated into action intentions, however, the same condition does not hold true. (When we speak of ‘outcome intentions’, we will mean those outcome intentions which cannot be translated into action intentions unless otherwise noted.) Agents cannot be certain that they can satisfy their outcome intentions, because they cannot be sure that the other agents in the MADP are going to do their part in order to bring about the intended outcome. The agent who intends the outcome $(C,C)$ in a prisoner’s dilemma cannot be certain that the other agent is going to do her or his part by cooperating.

There is a tension inherent in outcome intentions. Rationally holding an outcome intention in a MADP requires that the agent have some rational control over whether or not the
desired outcome will be produced; one cannot intend outcomes that one has no control over. At the same time, if we think that agents do have control over the decisions or actions of the other agents in the MADP, then that violates the integrity of the MADP; by perceiving one’s self as having some control (or even significant influence) over the actions of the other agents in the MADP, one ceases to perceive of them as agents at all. One characteristic feature of agents is that they choose autonomously.\footnote{This is not, of course, the Kantian sense of autonomous choice; though rational agents might choose in accordance with their self-imposed constraints, they also might simply choose in accordance with their strongest preference.}

It is at this tense spot that the importance of the own action condition is most perspicuous. If agents can rationally only intend their own actions, then it seems that they are limited to only holding action intentions in MADPs: they can only intend to perform some specific action themselves, and then the consequences of their actions must also be intended, whatever they may be. If we want agents to rationally be able to intend outcomes, then we must find a way to blunt the force of the own action condition here. In the next section, I want to suggest that we do need to blunt the force of the own action condition, because a set of agents each holding action intentions cannot get us all the way to collective action.

5.5 Action intentions fail to explain collective goals

We have talked now about intending outcomes, and intending actions; we have also suggested that intentions are equivalent to goals for the purposes of our model of deliberation. We have not yet said anything about collective goals, however. In Chapter 2 we rejected the claim that collective action only requires that some outcome be produced which is identified as the collective outcome (as, for example, by Paretian considerations). Instead, we argued that in order for a set of actions to be considered a collective action, there must be some process of collective reasoning, and collective reasoning requires that there be a collective goal that the agents aim to produce.

What, then, is a collective goal? Just as an individual goal structures the reasoning of the deliberating individual agent, so too must a collective goal structure the reasoning of the deliberating collective. We have already rejected the metaphysical view of collectives, promoted by Gilbert and others, that they not be reducible to individuals.\footnote{See Chapter 3.} Rather, we should make the assumption that collectives are made up of individuals, and thus that any
property that collectives hold can be reduced in some way to properties of the constituent individuals. Applied to collective goals, this means that when collectives have goals, then there is something that can be said about the goal-oriented behavior of each of the individuals that makes up the collective.

Importantly, however, the goal-oriented behavior must be goal-oriented behavior of the collective. In other words, each agent who deliberates on behalf of a collective goal must be deliberating towards a goal that makes some essential reference to the collective. The same situation applies in the case of individual goals; a goal cannot be said to be agent X’s goal if the goal does not make any reference to X’s actions or things that X might achieve. I cannot have as an individual goal that the Earth move closer to the Sun. Nor can I have as my individual goal that there be fewer beetles in the Amazon, if I do not think that there is any causal connection between any of my feasible actions and the population of Amazonian beetles. This constraint on goals does not come from my beliefs; it is not as though I believe it impossible for the Earth to move closer to the Sun, or for there to be fewer beetles in the Amazon. Rather, the problem is that neither of these outcomes are plausible goals of mine, because they do not refer in any way to me or my behavior. So it is with collective goals; an agent cannot hold a collective goal unless the collective goal makes reference to the behavior of the collective—where, again, we understand the collective to mean the set of agents who make up the collective.

There are some situations in which we speak of ‘collective goals’ even though they do not make reference to the behavior of each of the members of the collective. In some cases, the contribution of some of the members of the collective is conditional or counterfactual; a team might have as a collective goal winning a game, even though many members of the team will not contribute because their contributions will prove unnecessary. If, however, they were needed—if there were an injury to one of the other members of the team, for instance—then the non-playing members would contribute. It seems plausible to say, in such situations, that all of the members of the team have share the collective goal of winning the game, even though their contributions might not be actually necessary (though they might be counterfactually necessary).

We also sometimes speak of collective goals even though some of the members of the collective do not have either actual or counterfactual contributions to the production of the collective goal. One such class of cases involve collectives which are large scale and sociological in nature, such as membership in a group, tribe, nation, or so forth.
such large-scale groups, we do sometimes speak as though the group has a collective goal, even though many or most of the members of the group have no actual or counterfactual contribution to the production of the goal outcome. As an example, we might say that a nation has a goal of winning a war against another nation, even though most of the members of that nation have no contribution, either actual or counterfactual, to the success of the war. In general, these cases are best analyzed as ones in which talk of ‘collective goals’ is loose or imprecise. To say that a nation has a collective goal of winning a war, then, is best analyzed as saying that a subset of the nation has a collective goal of winning the war—namely, those members of the nation who are actually engaged in waging the war—while the other members of the nation are best described as socially identifying with the group. Social identification is an important phenomenon and one well worth theorizing about; such is not my goal here, however, and we should restrict our attention to collective action first. Insofar as the two concepts overlap, we can say things about social identification; we should not expect, however, to find extensive overlap between the two concepts.

For true collective goals, then, the collective goal must reference in some way the behavior or actions of everyone in the collective. We can accommodate the counterfactual cases by noting that those agents whose behavior is not actually necessary for the production of the collective goal still have their deliberation and choices and actions structured by the collective goal; the athlete who does not enter the game is still doing something, and that thing is rationalized as being the best thing that can be done to support the team among the available options.

This helps to explain why a set of action intentions cannot account for collective goals. Collective goals must reference the (actual) behavior of all the members of the collective. The collective goal that the football team possesses of winning the game makes reference to the behavior of each of the members of the team; they all have a part to play, and the collective goal acknowledges that fact in some way. (The specific way that it does so will be discussed in greater detail below; what matters here is just that it must make reference to each of their behaviors in some way.) Action intentions, however, do not make the necessary reference to the actions of the other agents. For the agent who intends to perform some

---

12There might, of course, be some possible world in which each of the members of the nation have some role to play in the success of the collective goal. My point here is merely that the relevant possible worlds are so far away as to be practically irrelevant for our discussion.

13This might run counter to some of Gilbert’s claims about the importance of social identification for understanding collective action and vice versa. See Gilbert 1989, Gilbert 1996a, and Gilbert 2000.
Now, it might be claimed that, just in virtue of being embedded in a MADP, the intention will make some reference to the actions of other agents. After all, the intending agent will have beliefs about the outcomes that might be produced as a result of the agent’s intention, and those outcomes can only come about through the agency of the other agents in the MADP. This is not enough, however, for the action intention to make reference to the actions of the other agents. Here, the reference to the actions of other agents does not lie in the agent’s intention itself, but is rather in the agent’s beliefs.

We can put the point another way. Collective goals must be taken to structure the deliberation of every agent in the collective. The effects of a collective goal on the deliberation of agents in a collective will vary from agent to agent; if we consider teammates attempting to win a game, then the pursuit of winning the game will lead each agent to deliberate in different ways; the defenders might position themselves to prevent scores against their team, while the attackers will position themselves to score goals. Nevertheless, the collective goal should be understood as structuring the deliberation of all. With action intentions, however, it is unclear as to how we can get one collective goal to structure the deliberation of every agent if we restrict ourselves to action intentions. The problem is that we seem to run into an alternate version of Searle’s worry: we can’t distinguish between two situations, in one of which the agents all have action intentions and they sum up to a collective goal, and the other in which the agents all have action intentions, but they don’t sum up to a collective goal.

The reason that we cannot distinguish between these two cases is that an agent’s holding an action intention is consistent with that agent’s intending an individual action. Take X and Y in a prisoner’s dilemma again. X’s having an action intention of cooperating means that X intends to cooperate regardless of what Y does. X’s holding this intention is entirely compatible with X’s intending the individual action of cooperating. Likewise, Y’s holding an intention to cooperate is compatible with Y’s intending the individual action of cooperating. If we want to claim that there is some way for these two action intentions to sum up to a collective goal, we have to explain what the key difference is between the agents’ holding a collective goal, and the agents’ holding two individual goals. Whatever the difference is, it cannot lie in the action intentions themselves, as the action intentions are identical across the two cases.
5.5. ACTION INTENTIONS FAIL TO EXPLAIN COLLECTIVE GOALS

5.5.1 Conditional intentions

It might be thought that we can solve this problem through conditional action intentions. Under this view, a set of action intentions can jointly constitute a collective goal so long as the intentions are conditional upon the other agents’ also doing their part. In the case of the prisoner’s dilemma, the conditional intention view would hold that $X$ and $Y$ hold a collective goal of promoting the mutual cooperation outcome so long as $X$ has an action intention of cooperating which is conditional upon $Y$’s having the action intention of cooperating, and the same, mutatis mutandis, for $Y$.

Now, in general, it is very difficult to account for conditional intentions using the model of rational deliberation that we have adopted. Allowing for conditional intentions suggests that agents might form decisions about what to do asymmetrically; one agent first forms an action-intention, and then another agent gets to respond with an action-intention. This isn’t feasible in our model of deliberation, in which agents act simultaneously: assuming that the conditional intentions are inputs to the agents’ deliberation, then either the conditional intentions have not yet been actualized, in which case one of the agents must form the necessary action intention, or the conditional intentions have been actualized, in which case they are not conditional action intentions any longer, but are rather actual action intentions.

We might think that, insofar as this is true, then so much the worse for the model of deliberation that we have adopted. If our model of deliberation requires agents to act simultaneously, then our model cannot be applicable to collective action, since collective action requires that agents be able to act, or form intentions, asymmetrically. Whether collective goals are best analyzed in terms of conditional intentions, in other words, is independent of how we model deliberation.

The worry about asymmetry is a deeper one, however. If we are to analyze collective goals into a set of conditional action intentions, then something has to get all of these conditional action intentions off the ground. Conditional intentions by themselves cannot constrain deliberation; the conditional intentions must be made actual. A set of conditional action intentions by itself cannot account for collective goals, because collective goals are actual; they do in fact structure the deliberation of the agents in the collective. We need, then, some way to render the conditional intentions actual, and this requires that some agent form an actual action intention.

And this, in turn, perpetuates the asymmetry between the agents in the conditional action. In order for the conditional intentions to become actual, some agent must be the
one to form an actual action intention. In every instance of collective action, then, there is some agent who is the ‘leader’. This is unrealistic, however; it seems as though there very well might be cases in which agents possess a collective goal, but there is no clear ‘leader’ in making the conditional intentions actual. Consider, as an example, agents who possess a collective goal, but engage in deliberation upon that collective goal separately from one another—the soldiers on the opposite side of the island who aim to capture an outpost, say. The agents have a collective goal; further, they must make their decision about what to do on the basis of their beliefs about what the other agents are likely to do. Suppose further that there are two routes that the two groups can each take, and that coordination will result in success while a lack of coordination will result in failure. In this case, neither group knows for sure what action the other group will take. It seems unrealistic to suggest that one group is taking the lead in forming some definite action intention; neither knows whether the other group has formed some specific action intention, let alone whether the other group has formed the action intention before or after.

As we have seen, Roth suggests that in all cases of collective action, one agent is in a leadership position with respect to getting the collective action off the ground. Roth, of course, does not make this claim because he think that agents possess conditional intentions; indeed, it should be noted that Roth is skeptical about the possibility of conditional intentions to account for collective action. However, Roth does require that there necessarily be an asymmetry in all instances of collective action. This necessary asymmetry was unlikely for Roth, and it is unlikely in the case of conditional intentions.

Gilbert’s account of collective action depends on conditional intentions, although her account of conditional intentions is slightly more sophisticated than what we have considered so far. For Gilbert, there is a kind of two-step process to forming intentions with respect to collective action: agents first indicate their willingness to jointly commit themselves to forming a plural subject, and then they form those (action) intentions which derive from their constituting a plural subject. As a result, Gilbert seems to accommodate a certain amount of asymmetry; it doesn’t matter in what order agents indicate their willingness to jointly commit themselves to form a plural subject, just so long as they each do it.

---

14 Roth 2003.
15 Roth 2004. Note, however, that Roth is skeptical about conditional intentions because he is skeptical that conditional intentions can account for the interpersonal commitments that are necessarily attendant upon collective action. I argued in Chapter 3 that there are no such necessary interpersonal commitments of the kind proposed by Roth.
5.5. ACTION INTENTIONS FAIL TO EXPLAIN COLLECTIVE GOALS

Once they both indicate their willingness to jointly commit themselves to forming a plural subject, then at that point the plural subject has been formed, along with all the attendant commitments and responsibilities.

Gilbert’s account seems to avoid the worries. Gilbert’s account is compatible both with agential symmetry and with agential asymmetry; perhaps we can model our conditional intentions in such a manner? The problem with Gilbert’s account, however, is that it remains impossible on Gilbert’s account to distinguish between genuinely cooperative reasoning and strategic reasoning. Suppose X and Y are in a prisoner’s dilemma. They both indicate their possession of conditional action intentions of the kind that Gilbert describes: without yet committing themselves, they indicate their willingness to cooperate conditional upon every other agent’s expression of their willingness to cooperate. These expressions of conditional willingness are not yet binding in any way. Nor, it seems, are they automatically binding when the other agents also express their conditional willingness. The reason is that these expressions are not intentions; they are not even conditional intentions, properly speaking. They are expressions of a willingness to form an intention conditional upon other conditions being satisfied. That expression of a willingness to form an intention might be done strategically, however; it might be made in order to get the other agents to actually form the cooperative intention.

Suppose X says ‘I am willing to cooperate if you are.’ Y responds in kind. Should they then cooperate? X might be more confident that Y has the intention of cooperating, but—and this is the important point—Y’s intention of cooperating does not impose any restrictions on X’s deliberation at all. Does X have the intention to cooperate? It is not clear why she should. Her expression of a conditional willingness to form the intention to cooperate is not yet an intention; it does not constrain X at all. In order for X to have constraints upon her deliberation, she needs to possess an intention of her own.

The best way, then, of understanding Gilbert’s two-step process of conditional intentions is as providing a belief update for the agents. In announcing their willingness to conditional form an action intention to cooperate, X and Y are attempting to update the other agent’s beliefs about what each agent is going to do. X hopes to lead Y to believe that she is going to cooperate, and Y hopes to lead X to believe that he is going to cooperate. Now, these expressions might be genuine; X and Y might actually form the intention to cooperate. But they need not be genuine; X or Y might still be reasoning strategically, and they might not form the intention that they say they will.
The preceding does not rise to the level of definitive proof. It is still possible that there is some way of describing conditional intentions such that they can account for collective goals, but the problem is a difficult one; conditional action intentions entail an asymmetry which makes it exceedingly difficult for the conditional intentions to account for the constraints that intentions should supply. If we want to account for collective goals, then, I suggest that we turn to outcome intentions.

5.6 Outcome intentions and collective goals

Unlike action intentions, outcome intentions can account for collective goals. In this section I argue that engaging in cooperative reasoning can be analyzed as rationally holding an outcome intention.

Cooperative reasoning differs from individual reasoning in that cooperative reasoning is on behalf of team goals, whereas individual reasoning is on behalf of individual goals. As we saw above, an individual goal is a set of constraints over an agent’s deliberation, such that the deliberating agent finds it rational to act so as to produce towards one of the goal outcomes. Individual goals constrain the action of the possessing agent; if X holds an individual goal, then that goal imposes constraints on X’s actions—the goal impacts which actions X can consider it rational to perform. As such, individual goals represent actions for just one agent—the possessing agent. Individual goals are action intentions. X’s having a goal of cooperating in a Prisoner’s Dilemma is equivalent to holding an intention to bring about any outcome in which X cooperates; it thus encodes information about X’s action. It does not encode information about the action that Y will or might perform, since the intention is compatible with any action that Y might play.

Collective goals, on the other hand, have to contain information about the actions of everybody in a collective. Whereas individual goals are guides to the deliberation of an individual, collective goals are guides to the deliberation of a collective. Ontologically, collectives just are collections of individuals; our challenge, therefore, is to understand what it means for a collective goal to guide the deliberation of a collective, and how this differs both from agents deliberating upon one individual goal, and from agents deliberating upon a set of distinct, but still individual, goals.

Outcome intentions can account for collective goals because outcome intentions do make reference to the actions of each of the agents in a collective. Take an arbitrary outcome
intention $\phi$, which rationally permits an agent to bring about any of a set of outcomes $O$. Each outcome in the intention set $O$ can be produced only through a set of actions by each agent in the MADP, which means that each outcome makes reference to those actions which produce the intended outcome. Consider the prisoner’s dilemma. One outcome intention would be to intend the mutual cooperation outcome $(cooperate, cooperate)$; this outcome makes reference to the actions of every agent, because the outcome requires that every agent cooperate in order for the outcome to be brought about.

Remember our requirements for something to function as a collective goal. Just as individual goals structure the deliberation of an individual, and make reference to the actions of the intending agent, so too do collective goals structure the deliberation of the collective while making reference to the actions of the collective. Both of these conditions are satisfied by outcome intentions. As mentioned above, outcome intentions make reference to the actions of every agent in the collective, because the intended outcome requires specific contributions by each agent in the collective to be brought about. Moreover, outcome intentions structure the deliberation of every agent in the collective so long as the outcome intentions are shared by every member of the collective. Let us say, for simplicity, that collective action in a MADP requires that every agent in the MADP be a part of the collective. Then in order for some outcome intention to represent the goal of the collective, the outcome intention must be held by every agent in the MADP.

So, then, let us return to the prisoner’s dilemma example. In order for the outcome intention of bringing about the mutual cooperation outcome to be $X$ and $Y$’s collective goal, both $X$ and $Y$ have to hold the outcome intention. The outcome intention then structures the deliberation of both $X$ and $Y$; it structures $X$’s deliberation by making it rational for $X$ to choose to cooperate, and it structures $Y$’s deliberation in the same way. Collective outcomes need not structure agents’ deliberations in the same way, of course; the agents might share an outcome intention of bringing about the outcome in which $X$ cooperates but $Y$ defects, for instance.

An outcome intention thus represents a collective goal by representing those outcomes which the collective aims to bring about. Outcomes are combinations of actions by a set of agents, and so by representing the outcomes that the agents find it rational to bring about, outcome intentions also make reference to the actions of each agent in the collective.

---

16This assumption is generally too strong; it is possible for subset of the agents in a MADP to constitute a collective and bring about collective action. We lose nothing by employing this shortcut way of talking, however.
We have not yet gone into detail about what collective action looks like; all we have said so far is that we can represent a collective goal for the purposes of deliberation as an outcome intention. We will give a more complete account below of what collective action looks like. First, we should address the worry with which we began the chapter: if collective goals (in the form of outcome intentions) represent actions of each agent in the collective, then how can our account of collective goals remain compatible with the own action condition?

5.7 Outcome intentions and the own action condition

Key to the claim that outcome intentions do not violate the own action condition is the fact that intentions, as we have characterized them here, impose negative constraints upon deliberation, and not positive constraints. For a volitional mental state to impose positive constraints on an agent, the mental state must normatively require that the agent perform those actions which are in the content of the mental state. If we interpret \(X\)'s intention to \(\phi\) as imposing positive constraints on \(X\), then \(X\) is normatively required to \(\phi\); if she does not \(\phi\), then she has failed to satisfy the normative commitment of her intention.

On the other hand, volitional mental states which impose negative constraints upon an agent do not normatively require that the agent perform the content of the mental state. Rather, negative constraints impose on the possessing agent a norm forbidding the agent from performing those actions contrary to the satisfaction of the content of the mental state. If \(X\)'s intention to \(\phi\) is interpreted as imposing negative constraints upon \(X\), then her intention will generate upon her a norm to not perform those actions which would be incompatible with her doing \(\phi\).

The model of intentions that we have set forth so far models intentions as imposing negative constraints. Under our model, intentions are characterized by their role in deliberation, and their role in deliberation is to constrain deliberation. This constraint is a negative constraint, and not a positive constraint.

Why are negative constraints over the actions of other agents consistent with the requirements of the own action condition, where positive constraints would not be? Positive constraints are of the form ‘perform action \(\phi\)’. If we were to understand agent \(X\)’s intention to do \(\phi\) as a positive constraint, then her intention would positively require her to do \(\phi\). If \(\phi\) were a collective action, then \(\phi\) would reference actions for each agent in the group. \(X\)’s positive intention to \(\phi\) would then require \(X\) to perform not only her component of \(\phi\), but
5.7. OUTCOME INTENTIONS AND THE OWN ACTION CONDITION

also the action components of every other agent in the group; this would violate the own action condition. As we saw in Chapter 4, this was the problem that Bratman faced; in attempting to characterize the collective intention as a positive intention, he inadvertently required that agents intend the actions of other agents.

Negative constraints, on the other hand, do not positively require that agents perform any specific action. Rather, negative constraints prevent agents from performing any action which is incompatible with the intended outcome. X’s intention to φ imposes on her negative constraints to not perform those actions incompatible with her φ-ing. These constraints are relevant for X only insofar as she is engaged in deliberation about actions which might be incompatible with her φ-ing; the negative constraints do not have any effect on X’s deliberation otherwise.

Consider now collective goals. If X holds the collective goal that some group G bring about outcome o, then X accepts this collective goal as constraining her deliberation. As such, X accepts that any action of hers that is incompatible with the production of o is irrational. She also accepts that any action of any agent in G that is incompatible with the production of o is irrational.

Why would X accept constraints on the rational deliberation of other agents in G? This is the important difference between individual goals and collective goals; in holding a collective goal of producing o, X views herself as a member of a collective, and she views the collective (rather than the individual) as the fundamental unit of agency in producing o. As such, she considers rational restrictions on everyone’s deliberation, and not just her own; in holding the collective goal, she understands the goal to restrict every agent in the group. X’s deliberation, however, is only substantively affected by those restrictions which apply to her own actions; her deliberation is unaffected by the restrictions on other agents in the group, because those restrictions do not restrict any of the actual options that X has to choose between. The restrictions that collective goals impose on other agents in a group are practically irrelevant, but they are still important in distinguishing between individual goals and collective goals; in collective goals the agent understands those restrictions on other agents to exist, whereas in individual goals the agent does not.

We argued above that agents with individual intentions cannot hold intentions which constrain the deliberation of other agents; we suggested there that any intention which constrains the deliberation of other agents fails to satisfy the functional criteria of intentions, namely that they provide guidance for the deliberation of an individual agent. Why, then,
can we do so in the case of collective goals?

Collective goals are distinct from individual goals, in part because their functional role is different: individual goals guide the deliberation of an agent, while collective goals guide the deliberation of a group. This difference accounts for the fact that collective goals make reference to deliberative constraints on agents other than the possessing agent, whereas individual goals do not and cannot. The difficulty lies in understanding how collective goals can guide the deliberation of a group while being held, strictly speaking, by individual agents. As we have seen, collective goals can constrain the deliberation of a group so long as they are in fact held by every agent in a group. If every agent in a group holds the same collective goal, then the deliberation of the group will be constrained in the ways functionally necessary to promote the production of the intended outcome.

Because collective goals are held and deliberated upon by individuals, it is not necessary that every agent in a group hold the same goal in order for that goal to be a collective goal. Indeed, agents might have mistaken beliefs about other agents; X might deliberate on behalf of a collective goal while believing that Y also holds the collective goal, whereas Y might be deliberating strictly on behalf of an individual goal. This fact does not render X’s collective goal no longer a collective goal; rather, it means that X’s deliberating on behalf of the collective goal is likely to be fruitless, as the collective goal will not in fact constrain the deliberation of the relevant group, and so it will fail to perform its function as a collective goal. There is, however, no incongruity in X’s holding and deliberating on behalf of the collective goal; X allows the collective goal to structure and restrict her own deliberation over the actions available to her in the decision problem, which is what the collective goal ought to do for X’s deliberation.

In order to rationally hold a collective goal, an agent must hold the belief that the other agents in the group also reason according to the collective goal. The agent who fails to satisfy this condition cannot reasonably believe that her or his deliberation upon the collective goal will lead to the production of the collective goal, as she or he will believe the other agents in the group unlikely to contribute the actions requisite for the production of the group goal. And, because intentions are functionally characterized by their production of the intended outcome, agents cannot hold a collective goal or intention while at the same time failing to believe that other agents in the group also intend to reason according to the collective goal or intention.

This belief condition applies to individual agents holding a collective goal; it is not a
condition on whether or not some goal is a collective goal or an individual goal. A goal is a collective goal so long as it references constraints on the deliberation and action of a set of agents; the goal is an individual goal so long as it references constraints on only one individual agent’s deliberation and action. Both collective and individual goals are directed at outcomes. The difference, then, lies in whose actions the goal purports to constrain.

5.7.1 Negative and positive constraints

We argued above that our account of collective goals remains compatible with the own action condition because collective goals entail only negative constraints and not positive constraints. Collective goals make reference to the actions of all the agents in a collective, but they do not violate the own action condition because they only entail negative constraints—agents are merely prevented from performing actions incompatible with the success of the collective action. They are not enjoined to actually do actions of other agents, because collective goals do not entail any positive constraints.

The distinction between negative and positive constraints is thus crucial to my claim that my account of collective action does not violate the own action condition. We argued earlier that Bratman’s we-intention theory of collective action violates the own action condition because his intentions entail both positive and negative constraints. The theory that I present here, on the other hand, does not have such a problem, because intentions do not entail positive constraints.

Mental attitudes that entail only negative constraints, rather than both negative and positive constraints, are poorly suited to move an agent from deliberation to action. Negative constraints only tell agents what they ought not do; but something still needs to tell agents what they ought do. This is why, in most theories of action, intentions play both roles: they provide both negative and positive constraints upon agents. Under the usual view, which is endorsed by Bratman, holding an intention to do some action \( x \) generates pressure upon an agent both to not do those actions incompatible with \( x \)—the negative side—but also to actually do \( x \)—the positive side. According to my theory, on the other hand, holding a goal or intention of doing \( x \) only constrains the intending agent to not do those actions which conflict with her or his doing \( x \).

So, then, how do I account for the positive constraints? On my theory, this role is played by the decision problem itself. Any agent who faces a decision problem and deliberates about what to do thereby incurs a positive constraint: one must do whatever action best resolves
the decision problem. If agent $X$ sees herself as facing a decision between performing actions $x_1$, $x_2$, and $x_3$, then the very fact that $X$ recognizes herself as facing this decision problem provides her with the positive constraint of doing whichever action is most justified by her available reasons. Suppose $X$ has a goal of doing $x_1$. Then $X$’s goal makes it irrational for her to do either $x_2$ or $x_3$, but does not itself provide positive constraints; $X$’s positive constraint of doing $x_1$ is generated by the fact that she faces a decision problem in the first place.

The positive constraints that decision problems generate are not constraints based upon the content of the agent’s intention. Agents who face decision problems have a positive constraint of performing that action which is best supported by reasons, but they do not have a positive constraint of doing whatever is in the content of their intention. Thus, agents who hold a collective goal which refers to actions of other agents do not have any positive constraint of performing the other agents’ actions; rather, their positive constraint is only to do that action which is best supported by their reasons.

The account of collective goals presented here is thus importantly different from that presented by Bratman. If positive constraints are generated by the decision problem that agents face, and not by collective goals, then the fact that a collective goal refers to other agents’ actions does not entail that the goal will violate the own action condition. Rather, agents still only face positive constraints of performing their own actions. The negative constraints that collective goals generate do affect the actions of other agents, but these negative constraints have no practical effect—agents are merely prohibited from doing those of other agents’ actions which are incompatible with the collective goal, but as this claim is vacuous, there is no violation of any reasonable condition on intentions.

5.8 Rationally holding and acting upon collective goals

Just as with individual goals, it is possible for a collective goal to be irrationally held. The case is more complicated with collective goals, however, because whether an intention is rationally held depends both on the agent’s beliefs about her or his other mental states, and also about the agent’s beliefs about the mental states of the other agents in the collective. With respect to the rationality of collective goals, there are two things to consider. The first is whether an individual agent is justified in reasoning according to some particular collective goal; the second is whether a collective is justified in possessing a collective goal.
5.8.1 Individual rationality and collective goals

In order for an agent to rationally hold an intention, whether that intention represents an individual goal or a collective goal, the agent must believe it possible for her or his actions to lead to the satisfaction of the goal. Thus, the agent who holds a collective intention must believe that the conditions are right for the agent’s intention to be satisfied. We first present a simplification of the core belief conditions that must be satisfied for a group of agents to be engaged in a collective action; afterwards, we note the various ways in which the core conditions can be softened while still allowing for collective action.

These conditions are all belief conditions; they all speak to the agent’s beliefs about the nature and structure of the MADP. Further, the content of the beliefs must be common beliefs; the agent must believe that every agent in the decision problem holds the belief, and believes the same to hold of the others, and so forth. The first common belief condition is about common agency. In order for an agent \( X \) to rationally hold a collective goal, \( X \) must have beliefs about the number of agents in the decision problem who are collectively reasoning, and must believe the other agents in the decision problem to hold similar beliefs. Note that we are not saying that every agent in the collective must actually hold in common a belief about the number and identity of the agents in the collective in common. Rather, what we are saying is that each agent in the collective must believe that there are common beliefs about the number and identity of agents in the collective. In assessing the rational possession of a collective goal, what matters is that each agent have reason to believe that the collective goal is likely to be achieved, and this requires that each agent hold beliefs about what the other agents in the group believe with respect to the identity of the collective. And while the actual constitution of the collective matters when we consider whether the collective is actually going to be successful, it does not matter when we consider the rationality of agents’ holding collective goals.

In addition to the common agency condition, an agent’s beliefs must satisfy a common goal condition in order for the agent to rationally hold the collective goal. The common goal condition merely states that the agent must believe the possession of the collective goal to be common knowledge among every agent in the collective. Whatever each agent believes the collective goal to be, the agent must also believe that every other agent also takes that goal to be the collective goal. Again, what matters is not whether some goal really is held in common among all the agents in the collective; what matters is merely whether each agent believes it to be held in common, since we are concerned with the rational conditions
for holding the collective goal.

Finally, there is a common problem condition: agents must each believe that the nature and structure of the decision problem that the agents face is common knowledge among all agents in the collective. The structure of the decision problem consists of the actions available to each deliberating agent, the outcomes that are possible to be produced, and the outcome functions which map sets of actions onto outcomes.

To illustrate the above, take two agents, $X$ and $Y$, who face a decision problem. They each have two actions available to them: $x_1$ and $x_2$ for $X$, and $y_1$ and $y_2$ for $Y$. Suppose also that there are four outcomes which might be produced: $o_1$, $o_2$, $o_3$, and $o_4$. $X$ can rationally hold and deliberate according to the collective goal of producing $o_4$ so long as the above conditions are met. The common agency conditions requires that $X$ believes that the fact that $X$ and $Y$ are engaged in collective agency with one another be a commonly held belief; $X$ must believe that she and $Y$ constitute a collective, and $X$ must believe that $Y$ believes that they constitute a collective, and so forth. The common goal condition requires that $X$ believe that $o_4$'s being a collective goal is a common belief—$Y$ takes $o_4$ to be the collective goal, and $Y$ believes $X$ to take $o_4$ to be the collective goal, and so forth. And finally, the common problem condition requires $X$ to believe that she and $Y$ have common beliefs about the structure of the decision problem. If $X$ believes that $o_4$ is produced through her doing $x_2$ and $Y$’s doing $y_2$, then $Y$ believes it also. If $X$ believes that she only has two actions available to her, $x_1$ and $x_2$, then $Y$ believes it also. And so forth.

Now, it is possible for collective action to occur even when these core conditions are violated. it is possible, for instance, for agents to produce an outcome even though one agent has different beliefs about the available actions than another agent does. In this case, it is hard to see how an agent can rationally hold a collective goal while failing to satisfy one of these conditions. If $X$ knows that $Y$ has different beliefs about which actions constitute the decision problem, then $X$ cannot be confident that $Y$’s possessing the collective goal will lead to the production of the intended outcome. If $X$ were to continue to hold such a collective goal in the face of the perceived discrepancy in their beliefs, then $X$ would be irrational. Nevertheless, it is conceivable that agents might be able to engage in collective action in spite of such discrepancies.

There are more clearly problematic challenges to our core conditions. In particular, we might be inclined to believe that collective action is possible in spite of some ambiguity among the participants about who is participating and what their motivations are. A
pyramid is built through the actions of a large number of individuals; must each worker believe that the identity of every one of the other workers is common knowledge? Surely not. Agent X might be doing her part of the collective action—she places the bricks where they should be placed, and is confident that the other agents in the collective action are doing the same—without having any well-defined belief about who the other agents are. Assuming the pyramid to be successfully built, however, it seems clear that the pyramid was built as a result of a collective action.\footnote{Thanks to Don Loeb for the example, and for helping me to see the force of the worry.}

What makes this a case of collective action is the fact that the \textit{constraints} are still believed to be held in common, whoever the agents are. While X is doing her part in the collective action, X is also confident that the other agents are doing their parts as well. In other words, X is confident that the other agents have their actions constrained by the same collective goal as she is constraining her deliberation according to.

Note that X might not even be in a position to understand exactly what the collective goal is. X might have a vague understanding that her actions are contributing to a collective project, without being certain about exactly what the collective project is. Indeed, it is in these cases that our intuitions about who should acceptable moral responsibility for the success of an immoral collective action are the most shaky. Consider, for example, the contributions of IBM executives to the production of the Holocaust. Their data processing machines may have been crucial to the execution of the Holocaust. How detailed do their beliefs about the collective project have to be in order for them to be morally complicit?\footnote{Edwin Black makes a compelling argument that executives of IBM Germany were extremely complicit in Black 2001, in part because they were fully aware of the purposes to which their computing devices were being put. If we suppose that the executives were slightly less aware of Hitler’s plan than Black alleges, however, then it might be an open question as to whether the executives were members of the collective action or not.}

Our intuitions, I think, largely track the extent to which agents constrained their deliberation and their actions according to the collective goal. If the IBM executives understood Hitler’s intended use of their machines, but constrained their deliberation entirely according to their own business interests, they might have been (and continue to be) morally accountable for their (individual) actions but nevertheless not have been members of the collective action. If, however, they were prepared to let Hitler’s goals constrain their own planning, we might think them to have been more fully-fledged participants in the collective action.

What our theory of collective action should track, then, is the extent to which each agent in a collective has her or his deliberation constrained by the collective goal (assuming the
agents to all be rational). Given the model of deliberation over subjective decision problems that we have adopted, we are best accepting the core belief conditions; they suffice to show how agents in a collective action must believe that all other agents in the collective action are bound by the collective goal. Whether these core belief conditions are necessary can be left for another discussion.

5.8.2 Collective rationality and collective goals

When we consider the question of when it is rational for a collective to hold a collective goal, here the answer that we give closely mirrors the answer that we gave in the individual case. Indeed, we can get quite some distance in assuming collectives to be ‘super individuals’ for the purpose of determining whether its goals are rational or irrational. Admittedly, it is possible for there to be, in the final analysis, no collective there at all. Take a case of three agents, X, Y, and Z. X believes that X and Y form a collective (and that this belief is commonly held between X and Y); Y believes that Y and Z form a collective (and that this belief is held in common); and Z believes that Z and X form a collective. In this case, the common agency condition is satisfied, and yet it is also apparent that there is no real collective in effect—just a set of individuals each reasoning (with at least some justification) on behalf of goals that they take to be collective goals.

Such cases are the rare exception, however. These cases do not help us to understand when a collective goal is rationally held, since there is no real collective to speak of. Accordingly, let us instead focus for now on those cases in which agents hold true beliefs about who constitutes the collective; this will let us focus on the issue of when collective goals are rationally possessed. In such cases, there are, as with individual agents, two ways in which we can fruitfully consider whether a collective intention is rationally formed and rationally held. The first is whether the collective goal is preference rational; the second is whether the collective goal is intention rational.

Whether a collective goal is, for the collective, preference rational depends on our possessing a way of measuring the preferences of the collective. There are Paretian considerations that we can use here; if a collective goal leads to the production of a Pareto optimal outcome according to the agents’ individual preferences, then we can say that the collective goal is preference rational. In Hi-Lo games, i.e. games of pure coordination with two equilibria, one Pareto superior to the other, a collective goal of producing the Hi outcome is preference rational in this sense. In general, however, determining the preferences of a
5.9. COLLECTIVE GOALS AND COORDINATION

5.9. COLLECTIVE GOALS AND COORDINATION

5.9. COLLECTIVE GOALS AND COORDINATION

While we have frequently appealed to prisoner’s dilemmas as examples, there is a danger in over-reliance upon the example; we run the risk of making it seem as though all collective actions take the form of responses to prisoner’s dilemmas. In fact, prisoner’s dilemma situations are relatively simple, in that each agent has only two actions; if the collective goal proscribes one action for each of them, then they each have only one remaining action available. There are choice situations, however, in which agents have more than two choices, or for which agents’ intentions might not suffice to determine which action they can play. Suppose X and Y each have three actions available to them: \( a, b, \) and \( c \). Their actions combine to produce the outcomes ranging from \((a, a)\) to \((c, c)\). X and Y might have a collective goal of coordinating their behavior on either the \(a\) action or the \(c\) action—that is, they might intend to bring about either of the outcomes \((a, a)\) or \((c, c)\). In this case,
both $X$ and $Y$ will realize that it is irrational to perform action $b$; there is no intention-compatible outcome which is produced by either agent’s performing $b$. Actions $a$ and $c$ are both rationally permissible for $X$ and $Y$ to play, insofar as they both potentially contribute to the production of an intention-compatible outcome. However, the collective goal of bringing about either $(a,a)$ or $(c,c)$ does not suffice for the agents to determine whether they should each do action $a$ or action $c$; whether they successfully coordinate depends on their selecting the same action as the other agent, and their collective goal does not suffice to determine which of the two permissible outcomes they ought to coordinate on.

More information is needed, then, to determine which action an agent reasoning on behalf of a collective goal should play. In this respect, collective goals are no different than individual goals. An agent’s individual intention of producing one of a set of outcomes does not by itself suffice to determine which of the permissible outcomes to bring about; the agent also needs some way of evaluating the desirability of the various permissible outcomes. In the individual case, as we have seen, the agent is able to choose one outcome to bring about over others because of her or his individual value function, which ranks the outcomes according to their overall desirability. In the collective action case, a similar situation holds; in order to determine which action to perform, the deliberating agent needs to have a value function with which to rank the various outcomes which are permissible to be brought about. Unlike in the individual action case, however, the agent’s value function does not suffice to determine what action the agent should perform. The agent’s value function suffices to determine which outcome the agent individually judges to be the best to bring about (consistent with the agent’s collective intentions); it does not suffice, however, to determine which action the agent should perform. The problem is that, in multi-agent decision problems, the agent cannot bring about a desired outcome unilaterally; the agent must coordinate her or his behavior with the other agents in the decision problem. The agent’s value function cannot enforce coordination on an outcome by itself; more must be done in order to ensure that the agents coordinate properly on some outcome.

What more need be done? Unfortunately, the fact that a set of agents are each engaged in deliberation upon a collective goal does not guarantee that the agents will successfully coordinate upon one outcome. If one outcome is not already sufficiently salient to ensure that the agents coordinate upon that outcome, then the mere fact that the agents intend to deliberate upon collective goals will not be enough to overcome the lack of salience.\textsuperscript{19}

\textsuperscript{19}Compare Schelling 1960.
Return again to the example above, in which agents X and Y are each deliberating upon their collective goal of producing either outcome \((a, a)\) or \((c, c)\). Regardless of the value systems that they each use to rank one outcome over another, it is clear that the agents will not have produced a outcome compatible with their collective goal unless they can coordinate their actions, either to produce \((a, a)\) or to produce \((c, c)\). If the agents cannot discuss their actions ahead of time, then they must attempt to coordinate their actions on the basis of their beliefs about the likely behavior of the other agent. For this, their own value functions ranking the outcomes are irrelevant, except insofar as they provide evidence for which action the agents might play, and therefore for which outcome they might coordinate upon.\(^{20}\) What matters is not what outcome the agents want to bring about, but rather what outcome the agents can bring about—and what determines the latter is what action the other agent is going to play. If X believes that Y is going to play \(a\) (with the aim of bringing about \((a, a)\)), then X has reason to play \(a\). At the same time, Y might believe that X is going to play \(c\) (with the aim of bringing about the \((c, c)\) outcome); Y would then be justified in playing \(c\). Though both X and Y are deliberating on a collective goal, their (justified) beliefs about the likely behavior of the other agent lead them to fail to coordinate their actions upon an outcome which falls within the permissible scope of their mutual collective goal.

Though this result is unfortunate from the standpoint of bringing about more cooperation in the world, it is nevertheless as it should be. That two agents are both engaged in deliberation upon a collective goal does not guarantee that they will successfully coordinate their behavior on that goal, unless the goal is already so specific as to eliminate any possible failed coordination. The platoon of army soldiers who are out of communication with one another, but who are each deliberating upon their collective goal of taking control of the island, might not successfully coordinate their actions with one another; there might be more than one way of seizing the island, and they might misjudge which actions are the best way to do it. The couple who intend to meet either at the ballet or at the sporting event might still fail to coordinate, not because of a lack of sufficient deliberation upon a collective goal, but rather because their goal (viz. attending the same event together) is vague enough to allow for more than one way of successfully satisfying it, and the agents

\(^{20}\)Even in the face of one or both agents’ possessing strong preferences in favor of one outcome over another, the agents still face a problem of confirming their coordination; this situation is similar to the problem faced by agents facing Hi-Lo games. See Sugden 2000, Sugden 2003, Bacharach 2006, Hollis 1998, Colman 1997, et cetera for a discussion of the problems of choice in asymmetric coordination games such as Hi-Lo.
do not possess the information necessary to distinguish between them. Successful cooperation requires that the agents coordinate their behavior on one goal which is permissible according to the common collective intention or goal, and yet such successful coordination is not guaranteed by the mere presence of a common collective intention. We will leave aside a detailed discussion of the various strategies that agents might employ to improve their chances of successfully coordinating upon a collectively permissible outcome; we need only note here that pre-choice communication will be the most common and effective way of resolving such difficulties.

5.10 Collective goals versus common goals

What about common goals? We can consider common goals to be goals which are individual, in that agents deliberate towards them using individual reasoning, but the intended outcome is known to be held in common by a set of agents. An example of a common goal is people distributing themselves in a subway car: everyone intends an outcome in which people are evenly distributed throughout the car, with no two people clustered together. However, it is typically the case that the agents in the subway car are all engaged in individual reasoning; people in the subway car do not consult with one another, and they do not consider themselves to be a part of a common unit of agency with the other agents in the car. In spite of the fact that they are individually reasoning, they can quite accurately predict the behavior of the other agents, and the resulting outcome is one which is intended by all. It would seem, then, as though a set of agents individually reasoning towards a common goal could be described as deliberating towards a collective goal.

In spite of the fact that common goals are common, they are not collective as we defined the term. Agents deliberating towards common goals are engaged in individual deliberation, so that in representing their goal outcome they represent only the restrictions upon their own actions that the common goal imposes. In the case of the agents in the subway car, each agent considers only the ways in which they have to move themselves in order to produce the desired result. Their deliberation takes into account the likely behavior of the other agents in the car, but each agent understands the other agents in the subway car to be constrained only by their own goals and attitudes.

Contrast this with a case of genuine collective deliberation by every agent in the subway: if each agent is engaged in collective reasoning, then they reason towards a goal which is
5.10. **COLLECTIVE GOALS VERSUS COMMON GOALS**

held in common by all—but the goal itself indicates which actions agents can perform and cannot perform. Each agent then reasons according to a goal which indicates not only their own permissible actions, but the permissible actions of other agents as well; and the agents reason with the belief that the other agents have adopted this collective goal. In the individual case, the agents each have a desired outcome, and they each reason so as to bring about the desired outcome, but they have less confidence about whether the other agents of the group are rationally restricted in any way—the other agents might be acting merely in order to maximize their preferences, or they might be acting in accordance with their individual goal of having separation from the other agents. If we were to characterize the content of the (individual) common goal and the collective goal, respectively, it might be as follows: each agent’s individual goal is to maintain separation from every other agent, and so the common goal is to have the spacing of the subway car be such that each agent is separated from every other agent; the collective goal, on the other hand, is that each agent does their part in bringing about an outcome in which every agent is separated from every other.

So much, then, for our account of collective goals. In the next chapter, we will provide a summary of our account of collective action, and discuss how the account we have provided here differs from the accounts we have criticized in Chapter 2.
Chapter 6

Searle’s challenge answered

6.1 Introduction

In the previous chapters, I have presented the outlines of a satisfactory theory of collective action. The theory has been described in its essentials; it remains here to show how the theory is applied.

Accordingly, the present chapter will proceed as follows. First, I summarize the account of collective action that we have been developing up to this point, including a consideration of examples of its use. We will first look at typical examples in the literature, to see how the theory handles those cases. We will then pay particular attention to examples which have posed problems for previous theories, and show that the proposed account handles those cases equally well.

We will then return to Searle’s challenge. As we will see, we can draw a distinction between two sets of actions: the first consists entirely of individual actions individually motivated, and the second consists of actions performed in pursuit of a collective goal. My claim is that being able to account for the difference in these two cases lies at the heart of understanding collective action and the justification of actions taken in pursuit of collective goals.

Finally, in order to get clearer on what the proposed theory of collective action is (and what it isn’t), we will examine the differences between the theory of collective action developed here and the theories that we have looked at in previous chapters. As we will see, the proposed view has a number of similarities with many other theories of collective action. In each case, however, the differences are important, and by carefully considering
the differences, we can both bring the proposed theory into sharper relief, and also more clearly see the problems that face other theories.

6.2 A theory of collective action summarized

Let us, then, consider the theory of collective action that we have developed in the previous chapters. Agents who face a decision problem can engage in *cooperative reasoning* or *strategic reasoning*. In order to rationally engage in cooperative reasoning, an agent must rationally hold a collective goal, and must reason about which action to perform in light of the collective goal that she or he possesses.

Goals—whether those goals are action intentions or outcome intentions—constrain an agent’s deliberation; they make it so that deliberating agents must find it intention-irrational to chose those action which lead to outcomes which are not compatible with the agent’s goal. Collective goals are similar, but they constrain the deliberation of *every* agent in the collective, and not just some. Collective goals, as we saw, are equivalent to outcome intentions: they are intentions to bring about some particular outcome (or one of a set of outcomes).

It is sufficient to rationally hold a collective goal, then, that the agent hold those core beliefs which are conducive to believing that the agent’s actions might contribute to the successful production of the collective goal. In particular, it is sufficient that the agent who holds the collective goal believes that there is a common belief among the agents in the collective about who the collective consists of; that the agent believes that there are common beliefs about the structure of the decision problem (including what actions are available to agents, what outcomes are producible, which actions lead to which outcomes, and perhaps information about the preferences of each agent in the collective); that the agent believes that there is a collective belief about what the collective goal is, and that it is held by each in the collective. If these conditions are satisfied, then the agent is rationally permitted to deliberate according to the collective goal.1

Deliberating in the light of a collective goal requires that an agent accept the constraints that the collective goal imposes on her or his deliberation. In particular, it means

---

1See the discussion at the end of Chapter 5; these conditions may not be necessary for an agent to rationally engage in collective reasoning, but they are sufficient.
that certain outcomes—those outcomes in the agent’s intention set—are permissible outcomes, and that agents must choose actions to play which are believed to bring about one of the permissible outcomes. In single-agent decision problems, the deliberation is simpler; the deliberating agent need only evaluate which actions lead to which outcomes, and consequently judge as irrational those actions which lead to outcomes incompatible with the agent’s intention. In multi-agent decision problems, the deliberating agent needs to choose an action based upon her or his beliefs about the likely action of the other agents in the decision problem; the agent must then judge as irrational those actions which, in conjunction with the expected actions of other agents, will lead to the production of outcomes incompatible with the agent’s intention.

The fact that there could be multiple outcomes which are compatible with collective goals means that agents sometimes need to engage in further deliberation about which of the intention-compatible actions the agent ought to perform. This decision, I suggested, is rationalized according to the agent’s preferences. Intentions remove some outcomes from rational consideration, but preferences ultimately rationalize the agent’s choice.

That an agent’s intentions could be satisfied by multiple outcomes also means that multiple agents reasoning according to a common collective goal are not guaranteed successful coordination. Agents not only need to share a collective goal, but they also need to correctly reason about the best way to coordinate their actions so as to produce one of the collectively intended outcomes. The standards of correctness for such reasoning are the same as in game theory: actions taken in pursuit of equilibria are rationalizable.

6.3 Collective action applied

Let us, then, consider how our account of collective action handles the cases that have been prevalent in the literature. In doing so, we can reassure ourselves that our theory of collective action is at least as strong as the other theories that we have been criticizing up to this point.

6.3.1 The footballers’ problem and Hi-Lo games

Sugden uses the footballers’ problem—what is also sometimes called a Hi-Lo game—to argue that traditional rational choice theories cannot adequately account for collective action. We suggested in Chapter 2 that Hi-Lo games by themselves might not lead to the result that
Sugden wants; traditional rational choice theories might have the resources needed to explain
the rational selection of Hi choices in Hi-Lo games. Nevertheless, the footballers’ problem
that Sugden describes can be considered a case of collective action, and so any adequate
theory of collective action ought to be able to account for it.

Let us consider Sugden’s own presentation of the example, from Sugden 2000:

Suppose that A and B are two attacking players in a football team. A has the
ball, but a defender is converging on him. B has more space, so A wants to pass
the ball to him. There are two directions in which B could run so as to intercept
a pass from A: call these left and right. Correspondingly, there are two points
on the field, left and right, to which A could pass the ball to be picked up by
B. There is no time for communication, or for one player to wait to see what
the other player does: each must simultaneously choose left or right. Suppose
that the move to the right puts B into a slightly better position. Say that the
probability that the pass will result in a goal is 10 per cent if both choose left
and 11 per cent if both choose right. If one chooses right and the other left, the
probability is zero. What should each player do?²

Sugden assumes this to be an instance of collective action. He suggests that the foot-
ballers’ problem is one of a broad class of decision problems in which

there is some objective (in this case, scoring a goal) which each of a set of
individuals wants to achieve, and which can most effectively be achieved if those
individuals coordinate their actions. Thus, we might say, it represents a situation
in which social cooperation would be useful or valuable.³

Sugden goes on to suggest other activities in the same class of problems, including ‘coop-
eration between adults to raise children, cooperation between hunters to kill game, [and]
cooperation between fighters in defence or predation.’

Sugden seems to be suggesting here that all that is required for collective action is that
a set of agents produce a collectively beneficial outcome. We rejected this approach in
Chapter 1; as we noted, agents can produce a collectively beneficial outcome even though
they are not reasoning collectively at all. In other words, agents might coordinate upon a
collectively beneficial outcome while engaged in purely strategic, individual reasoning.

As we have seen, the difference between cooperative reasoning and strategic reasoning
lies in the acceptance of common constraints. When all the agents in a collective accept
the existence of common constraints on their reasoning, then the agents are engaged in

²Sugden 2000, p. 179.
cooperative reasoning; when the agents do not accept the existence of common constraints on their reasoning, then they are engaged in individual reasoning. Individual reasoning can lead to the production of a collectively beneficial outcome, but we should be cautious in necessarily defining such cases of collective behavior as collective action properly speaking.

So, is the footballers’ problem a case of genuine collective action? I think that we can indeed understand the footballers’ problem to be a case of collective action, so long as we understand the footballers to be bound by the common pursuit of a collective goal. Sugden suggests that this is the case, when he suggests that neither A nor B have any preference to bring about one of the non-coordinative outcomes.

Still, we can make the condition more explicit. Let us say that both footballers A and B have the intention of scoring a goal. Sugden models this decision problem as a $2 \times 2$ matrix, and we can do this for convenience; if we are being precise, however, we should describe the decision problem in fuller detail. We can say that there are two outcomes in the decision problem: goal and nogoal. Each agent ‘deliberates’ between two actions; for A, the actions are passleft and passright, while for B the actions are runleft and runright. Given that A and B both intend to bring about the outcome goal, their deliberation then centers on the question of how to best bring about that outcome. They both know that the action pair (passleft, runleft) has a 10 percent chance of producing the outcome goal, while the action pair (passright, runright) has an 11 percent chance of producing the outcome. They also know that failing to coordinate is incompatible with their producing the goal outcome.

From this, A and B can both conclude that failing to coordinate their actions is incompatible with the satisfaction of their collective intention. The only actions which they can perform which might satisfy their collective intention is to coordinate their actions.

Now, at this point Sugden might object that our theory of collective action does no better than traditional rational choice theory. Traditional rational choice theory fails, Sugden argued, because it cannot make it uniquely rational for each of the agents to choose right; the decision problem has two equilibria, and so traditional rational choice theory cannot uniquely recommend the selection of one equilibrium over another, even when the one equilibrium is clearly superior.

Insofar as this is true, however, so much the worse for the traditional rational choice action.

---

4It is, of course, a bit of a stretch to suggest that A and B are deliberating about what to do. It is more likely that, as highly trained athletes, A and B are using non-deliberative means to resolve this decision problem—they intuit the best action to perform in light of their goals. We can follow Sugden, however, in taking this to be a case of deliberative collective action.
theory that Sugden objects to. It is true that there are multiple equilibrium games for which it is not clear how a rational agent—even a cooperatively disposed rational agent—ought to choose; Battle of the Sexes games are one notable class. We can be skeptical, then, that elaborate programs of equilibrium refinement such as that proposed by Harsanyi and Selten could ever represent genuine principles of rational choice (as opposed, say, to pragmatic principles of choice).\footnote{Harsanyi and Selten 1988.} We need not depend on the success of such programs, however, in order for it to be rational for both $A$ and $B$ to choose to pass or run to the right. It does seem to be a principle of rationality that, when a game has a set of equilibria of which one is Pareto optimal, then rational agents in the game will choose the Pareto optimal outcome. Indeed, we do not need a theory of collective rationality in order to rationalize such choices; a group of purely self-serving, self-interested agents will also find it rational to bring about the Pareto optimal equilibrium outcome.

So, then, if $A$ and $B$ are engaged in collective deliberation, then they must hold some collective goal in common, and that collective goal must constrain each of their deliberations equally. We characterized their collective goal earlier as being a goal of scoring a goal; the agents’ task, then, was to deliberate about the best means of bringing about the goal given their beliefs about the probable actions of the other agents. Importantly, however, this is not the only way to describe the situation. There is another way that we might characterize the collective goal. We could individuate the outcomes more finely, so that instead of there being two possible outcomes, $\text{goal}$ and $\text{nogoal}$, there are at least three possible outcomes: $\text{goal-by-moving-right}$, $\text{goal-by-moving-left}$, and $\text{nogoal}$. If we are to characterize the outcomes in this way, then we also have greater flexibility in characterizing the collective goals of $A$ and $B$. In particular, we can describe the collective goal as being e.g. $\text{goal-by-moving-right}$. In such a case, then $A$ and $B$ do not need to engage in further deliberation about which action to perform (assuming that they are both intention rational); they only have one action compatible with their intention—going right—even though the their performing their respective actions do not guarantee the production of the goal outcome (since they still have only an 11 percent chance of bringing about the outcome).

The point that we are making here is that there are two ways of accommodating the footballers’ example that Sugden appeals to. The first way is by characterizing $A$ and $B$ as sharing a collective intention of scoring a goal, through whatever means are best; the second way is by characterizing $A$ and $B$ as sharing a collective intention of scoring a goal.
through some particular means. Neither way is necessarily the ‘proper’ way of analyzing the footballers’ problem, and neither way is clearly suggested by Sugden’s presentation of the problem. Either way is acceptable as an analysis of the problem, however, as in both interpretations we can clearly state the collective goal of the agents, and the way in which the collective goal constraints the deliberation of the agents in the way typical of their sharing a collective goal.

Moreover, we should not be concerned about the fact that there exist more than one way of describing the decision problem that the agents in the footballers’ problem face. There are (at least) two ways in which we might understand what the footballers’ goal is, such that their actions might constitute a collective action. We might be uncertain as to which way of modeling their decision problem is correct; we might need additional information from the participants to know what the actual goal is. We might be skeptical, then, that we have modeled the decision problem correctly; we should not be skeptical, however, that one way of modeling the problem is the way which is most faithful to the participating agents’ actual goals and mental states.

What can we conclude from this? Sugden identifies a situation which is problematic for traditional rational choice theories, namely the footballers’ problem. One of the justifications for Sugden’s team preference theory of collective action is that his theory is both able to solve the problems that face traditional rational choice theories, and it provides a satisfactory account of collective action.

We argued above that the team preference approach is not as successful in accounting for collective action as its adherents claim. It is also important to note that traditional rational choice theories may not face the problems in accounting for footballers’ problems that Sugden implies. Now, whether a team preference theory of collective action is successful is ultimately independent of whether traditional rational choice theories can account for agents’ behaviors in footballers’ problems. But it is worth pointing out here that the footballers’ problem does not pose any problems for the theory of collective action that we have developed here. There are, as we saw, different ways of describing the collective action that Sugden references in his example. Our theory can account for the collective action regardless of how we describe it, however. Thus, if the footballers’ problem provides support

---

6 Andrew Colman borrows Sugden’s example in Colman 2003. For a sampling of defenses of the claim that traditional rational choice theory has the resources necessary to account for footballers’ problems, see the responses to Colman’s article, in particular Gintis 2003, Hausman 2003, and Monterosso and Ainslie 2003.
for Sugden, then it also provides support for us. Likewise, even if the footballers’ problem poses problems for traditional rational choice theories, it does not pose any problems for the theory we have presented here.

### 6.3.2 Going to New York together

Bratman favors a broadly constructivist theory of collective action, and his favored examples are intended to bolster support for his claims that a set of conditions for collective action are, in fact, sufficient for collective action. With respect to each of the members of the set, then, we have a couple of options: we can either accept that the element of the set is also necessary for collective action, in which case we need to show how our proposed theory of collective action can also account for that condition; or, we can reject the condition as being unnecessary for collective action, in which case we must show how we can modify the example so as to have a case of collective action without the questionable condition.\(^7\)

Let us remind ourselves of Bratman’s conditions for shared intention.\(^8\) It is sufficient for agents \(X\) and \(Y\) to be engaged in a collective action, Bratman argues, so long as they satisfy the we-intention condition: that each agent intends that they do the joint activity; the meshing subplans condition: that they intend their joint action-relevant subplans to mesh with those of the other; and the common knowledge condition: that the satisfaction of the other two conditions be common knowledge between the agents.

One of Bratman’s examples involves two agents going to New York together.\(^9\) In order for us to discuss this example properly, however, we should make a distinction that we did not need to make before. When we discussed Bratman’s theory earlier, we left the meshing subplans condition at a certain level of generality, in large part because our concerns were primarily with the adequacy of the we-intention condition. We should expand on the meshing subplans condition a bit here, however, as it is important for Bratman, in intending that their subplans mesh, the agents in a collective action also intend that the joint action be performed in part by way of the intention of the other agent. Each agent, in other words,

---

\(^7\)We should note that Bratman does not claim that his analysis of collective action aims at providing the necessary conditions for collective action. Bratman allows, for example, that the phenomenon of sharing an intention might be multiply realizable—that there might be many different ways of sharing an intention. See, in particular, Bratman 1999c.

\(^8\)Collective action, as I use the term, is based on shared intentions. Collective action is not, however, equivalent to what Bratman calls a ‘shared cooperative activity; as we noted in Chapter 2, shared cooperative activities are collective actions which also meet conditions of mutual support which are stronger than are needed for collective action.

\(^9\)See Bratman 1993.
must respect the agency of the other in a collective action. Bratman is careful to rule out what we can call ‘Mafia examples’: examples in which each agent intends a collective action, say going to New York together, but one agent intends to coerce or otherwise bypass the intentional agency of the other agent (as, for example, by stuffing the other agent in the trunk of the car).10

The meshing subplans condition forbids Mafia examples, because two Mafia agents who each intend that they go to New York together by way of stuffing the other in the trunk do not intend that their action-relevant subplans mesh; rather, they are engaged in strategic reasoning, viewing the actions and intentions of the other agent as things to work against, rather than things to work with. And when we interpret the meshing subplans condition correctly, and rule out Mafia examples as legitimate instances of collective action, we are left with more reasonable instances of collective action: two agents are engaged in a collective action of going to New York together so long as they each intend that they go to New York together, and they intend to do so by way of, and because of, their meshing subplans regarding their trip to New York, and all of the preceding is common knowledge.

Does our theory of collective action successfully account for X and Y’s collective trip to New York? It does. As we have pointed out in Chapter 3, we need to be careful about how we interpret the we-intention condition. I argued above that what Bratman calls ‘other-agent conditional mediation’ does not provide adequate grounds for accounting for an individual agent’s intention that a group do some joint action. The problem, in brief, is that an intention which is conditional upon another agent’s acting in some way does not suffice to account for an intention that they perform some descriptively neutral joint action.11

Nevertheless, I do not think that we need to reject the we-intention condition entirely. There must be some intention that agents in a collective action each have which is directed at the collective action (as opposed to mere individual actions). It is very misleading to call it an individual intention that a group do some collective action; such a label suggests that the agents’ intentions impose positive constraints upon the agents to bring about the collective action. It is better, but still imprecise, to say that the agents each have individual intentions that the group produce some particular outcome. (In our theory, agents do hold

---

10 We are assuming, of course, that being stuffed in the trunk of a car is incompatible with the claim that the agents are engaged in a collective action. The Mafia members are not, as Don Loeb suggested, teenagers stuffing themselves in the trunk of a car in order to sneak into a drive-in!

11 See Bratman 1999c for other-agent conditional mediation; see Chapter 3 for my argument against it.
outcome intentions; it is, however, imprecise to suggest that the content of the agents’ individual intention is that the group bring about some outcome.) Best yet is that we say that agents each individually allow their deliberation about what to do to be constrained by the group’s goal—that they each intend their actions to be constrained by the group’s goal.

If we interpret the we-intention condition in this way, then our theory of collective action can account for this condition of collective action. It is characteristic of collective action, we argued, that agents engaged in collective action are collectively reasoning towards a collective goal, and collective reasoning requires the agents to be constrained according to outcome intentions and not just to action intentions.

We can also accept and accommodate the meshing subplans condition on collective action. The first part of the condition—that the agents’ plans mesh with one another—is satisfied by the fact that outcomes are functions of actions performed by the set of agents. Accordingly, we can draw a clear distinction between two outcomes, the first of which is brought about by one agent locking the other in a trunk and driving to New York, and the second of which is brought about by two agents each choosing freely to take a train to New York. The first outcome is surely not one that the agents would each agree to be permissible according to their collective intention; they would not each agree to allow such an outcome to constrain their deliberation. The latter, however, is compatible with a collective intention to go to New York together, and so is an outcome that collectively reasoning agents could agree to constrain their mutual deliberation.

There are two things to note here. The first thing to note is that we can imagine the Mafia case to be compatible with their cooperatively reasoning about ways to get to New York together. Imagine $X$ and $Y$ to be playing a game, a spy-versus-spy sort of game in which they are attempting to get to New York together. Any outcome in which they arrive in New York together is a successful satisfaction of their collective intention. However, the two agents do not value every successful satisfaction of their collective intention equally; $X$ would prefer to be the one driving rather than the one in the trunk, as would $Y$.

If this is the case, then $X$ and $Y$ are engaged in a sort of strategic reasoning within a broader context of cooperation. This happens all the time, as for example when two agents play games of strategic reasoning with one another such as chess or cards. The games are cooperative; when playing chess, neither player views it acceptable to cheat and move contrary to the rules, or to distract the other player while she or he is thinking so as
to force her or him into making sub-optimal moves. The players accept certain common constraints on their game playing, and every outcome which is properly constrained counts as a successful satisfaction of their collective goal. Nevertheless, they might still reason strategically against one another, because they each value the outcomes which might be produced differently.

The second thing to note is that there might be many ways in which their collective outcome can be satisfied. \(X\) and \(Y\) might take a train to New York together, or take a cab to New York together, or fly together, or so forth. What matters is that \(X\) and \(Y\) are in agreement about the acceptable ways for them to reach New York together; so long as they are in agreement, and so long as they are each constraining their deliberation accordingly, then they are engaged in collective reasoning. In the Mafia examples, the outcome which is produced—one agent putting the other in the trunk of the car—is not one of the acceptable outcomes according to the collective goal, and so our theory of collective action rules out Mafia examples.

We do not need, then, to make explicit Bratman’s condition that the agents’ subplans mesh; this condition is already built into our description of outcomes as being functions of the actions performed by agents.

As for the second aspect of the meshing subplans condition, namely that the collective action be produced through the agency of each of the participants, we can accommodate this condition as well. The reason is that our theory of collective action requires that each agent be an agent—that they freely choose the action that they are going to perform. This requires that each agent respect the agency of the other. If we were to violate this condition, then agents would view the action performed by the other agents as being causally determined by the agent’s own choice—which strips them of agency and, accordingly, changes the decision problem from a multi-agent decision problem into a single-agent decision problem. And whatever we might say about single-agent decision problems, it should be clear that single-agent decision problems can never yield collective action.

Finally, then, the common knowledge condition. This, again, is something that we can agree with and accommodate—so long as we change the common knowledge condition into a belief of common knowledge condition. That is, each agent must believe the satisfaction of the conditions to be common knowledge. If this condition is satisfied, then each agent is justified in reasoning and acting according to the collective goal. If not, then not. It is too strong, however, to require that the satisfaction of the conditions actually be common
knowledge; agents are justified in acting on behalf of a collective goal with less than full-blown common knowledge.

We can thus deal with Bratman’s example just as easily as we can with Sugden’s. Bratman argues that any acceptable theory of collective action must be able to account for cooperative trips to New York together, and he also argues that any such theory must be able to explain why Mafia-style trips to New York together are ruled out as collective actions. As we just saw, the theory that we have presented here can do both: it can account for typical cases of collective action such as trips to New York together, and it can also explain why the Mafia members are not engaged in a collective action.

6.3.3 Walking and driving together

There is a set of examples that our theory of collective action cannot easily account for. This is the set of examples used by Gilbert and by Roth to argue in favor of the necessity of interpersonal normative commitments holding between the participants in a collective action.

The examples themselves are not problematic, and indeed can be easily accommodated in our theory of collective action. Gilbert favors talking about two agents taking a walk together. In this example, the participating agents need to successfully coordinate their behavior (as, for example, by neither walking too quickly nor too slowly). Roth talks about shared walks, but he also talks about more open-ended collective activities such as shared car rides, in which the intentions that the agents form impose rational constraints on the other agents in the collective action.

We cannot easily accommodate these examples in our theory of collective action, because we have already rejected the claim that interpersonal normative commitments necessarily inhere to collective action. We reject, then, Roth’s claim that $X$ and $Y$ are not actually engaged in a collective action if $Y$ does not accept $X$’s intention to drive first on the trip to Vegas as necessarily authoritative for him. Similarly, we reject Gilbert’s claim that insofar as $X$ and $Y$ are engaged in a collective walk with one another, it is necessarily true that each owes the other the performance of those actions necessary for the production of the collective action.

Nevertheless, we can explain where Gilbert and Roth’s examples go awry. Roth’s problem, we suggested, was that he has no principled way for his agents to know which revisions to the original collective plan are authorized revisions, and which are illicit. $X$’s deciding
to drive first might be an acceptable intention for both; revising the collective action to go to Reno instead of Vegas might be acceptable, though it is less likely; and revising the collective action so as to drive to Mexico together is quite likely an unacceptable revision of the original collective intention of driving to Vegas. How do we know which? On our theory of collective action, we can explain this. The original collective goal specifies which outcomes are acceptable to produce and which are not; the agents then have to reason about the optimal way to produce one of the acceptable outcomes. Is it acceptable for X to drive first? Almost certainly. It is also possibly the case that both X and Y have different preferences in this respect; both X and Y might want to drive first, and so they might be engaged in strategic reasoning with one another to see who gets to do it. Perhaps X runs to take the keys first, or perhaps X secretly replaces Y’s coffee with decaf so as to make Y more tired in the morning. The outcome—X’s driving first and Y’s driving second—is an acceptable outcome from the standpoint of the collective goal, and so both agents are willing to produce it. (Contrast this with an instance in which Y is emphatically not willing to drive second; in such a case, X’s driving first would not be part of the collective goal.)

This would also let us see why it might be compatible with X and Y’s collective intention that they go to Reno instead of Vegas. Suppose their collective intention were to gamble at a location within reason. Vegas might be a more preferable outcome for each of them, and so they might aim to coordinate their behavior on bringing about that outcome. Should the costs of going to Vegas prove too high, as for instance when the route to Vegas is blocked, then the agents might be willing to coordinate on going to Reno without reforming or reissuing their collective intention. This is not the default, however! If X and Y have a collective goal of gambling together in Vegas, it is far more likely that a trip to Reno will not be compatible with that original intention—and so any intention that X or Y form about going to Reno instead of Vegas will be an illicit intention.

Likewise, it is unlikely that the scope of X and Y’s original collective intention is broad enough to accommodate a trip to Mexico instead of Vegas. But, again, it is hard to see how Roth can rule out such a possibility. With the theory that we have proposed, it is easy to see how: even if the trip to Mexico is part of the original subjective decision problem, it is not a member of the set of outcomes which represents the collective goal.

The same situation applies to Gilbert. Gilbert wants it to be the case that there are normative commitments that necessarily hold between the participants of a collective action, but she also wants it to be the case that all such normative commitments are generated by
the fact that the agents are engaged in a collective action. The obligation that agents have to contributing to a collective goal, then, must arise from the fact that they are engaged in a collective action, and nothing else.

Contrary to this, however, I think that we can account for certain obligations that agents have to perform their part of a collective action. Our way of doing so appeals to the ordinary obligations generated by intentions, however, and does not appeal to any special properties of collective actions or collective goals. If X and Y share a collective goal of walking together, then both X and Y are constraining their deliberation so as to produce one of the acceptable outcomes. Both X and Y know that if X were to walk too quickly, the outcome produced would not be compatible with their collective goal. Because of this, X has reason to slow down; her reason is generated by the fact that she has a goal of producing a certain outcome, and that outcome cannot be produced if she walks quickly.

Now, on the one hand, this commitment is not a directed commitment. It is not a special commitment that X owes to Y, or that Y owes to X. Although the other participant in the collective activity might have a special basis from which to criticize non-performance, as for instance if that person is uniquely harmed by the other agent’s non-performance, it is nevertheless true that there is no general basis upon which to ground special obligations from one to another. X has a commitment to not walk too quickly because she intends to produce the collective goal; as such, her commitment is akin to any other intention-based commitment. If Gilbert is right that there is necessarily a special directed obligation that holds between X and Y, then our story of collective action cannot account for it.

This is not a worry, however. First, we have already argued that such special directed obligations do not necessarily exist. As such, we can leave open the question of whether such special directed obligations ever exist. It is clear that normative obligations do sometimes hold between participants in a collective action; the two crucial questions are whether they always necessarily hold, and whether the normative obligations can be accounted for without making an appeal to plural subjects. We argued above that these normative commitments do not always necessarily hold. Further, we argued above that when these normative obligations do obtain, they can be accounted for without appealing to plural subjects.

But second, and equally important, it is not clear what of importance we get from special directed commitments that we do not already get from ordinary intention-based commitments. If X and Y share a collective goal of walking together, then X has a belief
about what her contribution is to the collective action—walking at the right speed, not abandoning the walk partway through, and so forth. X has a commitment to do those things, insofar as X intends to participate in the collective action.

So what happens if X abandons her intention? Her commitment to performing those actions does indeed go away. This is the one major difference between our account of X’s commitments and Gilbert’s account: for us, commitments to act stem from the intentions that one holds, and not from the mere fact that one is engaged in a collective action. Abandoning the intention thus has the effect of eliminating the attendant commitment. If we think, then, that X is under some pressure to perform her part of the collective action, then that pressure might come from social obligations that X faces to not change her intention—as, for instance, if she led Y to believe that she would not change her intention. In this way, we can account both for each agent’s commitments in a collective action to do their part of the collective action, and we can also explain, to a limited degree, the extent to which agents face pressure to not unilaterally change or remove those intentions.

Does this strategy give us everything that the plural subject theorist can? No, it does not. For the plural subject theorist, one’s commitment to the collective action and to the other agents in the collective action is not conditional upon adopting the collective goal, and so it binds the agents regardless of their practical attitudes. The plural subject theorist locates the ‘social glue’ in the wrong place, however. It is not participation in a collective action which regulates whether certain actions or behaviors are socially appropriate. Participation in a collective action only determines whether an action is practically appropriate. If we want to know whether an action is socially appropriate, we must turn away from action theory. It might not be socially appropriate to be arbitrary with respect to one’s collective goals; likewise, it might not be socially appropriate to lead others to believe that one holds a collective goal that one does not, in fact, hold. These are not failures of collective action, however—at least not in any way other than the trivial way in which my abandoning my intention to take a walk is a failure of action.

Moreover, the account I give of collective action will prove not entirely satisfactory for either Gilbert or Roth because it suggests that the commitments that agents have towards one another in a collective action are not necessary features of collective action. Nevertheless, I argue, this is a better way of understanding collective action. Directed commitments simply are not necessary features of collective action, as I hope to have successfully argued above.
If, then, we strip Gilbert’s example and Roth’s example of the strong normative obligations that are not intrinsic features of collective action, our theory can account for the examples. I am not suggesting that the examples that Gilbert and Roth describe are impossible; agents might engage in a collective action with the norms that Gilbert argues in favor of, or with the strong practical intersubjectivity that Roth supports. These features are not necessary features of collective action, however; assuming that we can account for them in some other way—and I suggest that we can through the usual ways of accounting for normative features of social life—we needn’t build them into our generic example of a collective action.

### 6.3.4 Searle’s cooperative businessmen

Searle’s example was the one that started us off. Searle was interested in the difference between two cases: in the first case, the agents perform a set of actions which produce an outcome, but the actions do not jointly constitute a collective action, whereas in the second case, the agents perform the same set of actions but they do constitute of collective action. Searle’s question was, in virtue of what does the difference lie?

We are now in a position to answer this question a bit more properly. The difference between the self-interested businessmen who are not engaged in a collective action, and the cooperative businessmen who are, lies in the nature of the constraints that they adopt for themselves and for their deliberation. The self-interested businessmen have individual intentions; their intentions constrain only their own actions. Moreover, although they might have beliefs about the other agents’ intentions, they do not think that their goals are also the goals of the other agents.

In the story as Searle describes it, it is hard to make a clear distinction between the two cases. Still, we can draw some important distinctions. Each selfish businessman intends to act selfishly, with the understanding that humanity will thereby benefit. The priority seems here to be first to act selfishly, and second to benefit humanity. That humanity benefits is an expected consequence of their behavior, and not the intended outcome itself. In the case of the cooperative businessmen, they intend first to benefit humanity, and second to act selfishly. For the cooperative businessmen, acting selfishly is a means to the end, where the end is humanity’s benefit; for the selfish businessmen, the end is acting selfishly, and humanity’s benefit is an expected consequence.

The case becomes clearer when we consider counterfactual scenarios. In the case of
the selfish businessmen, suppose it should happen that being selfish does not lead to the betterment of humanity. What would the agents do? Well, their intention ‘tracks’ being selfish, and not bettering humanity; another way to express this same point is that the agent’s deliberation is constrained by the goal of being selfish, and is not constrained by the goal of bettering humanity. Thus, in such a case, the agent would find it rational to act selfishly; the agent would not find it rational to act so as to better humanity.

The situation is different in the case of the cooperative businessmen, however. In this case the shared goal is bettering humanity rather than acting selfishly, and so the agents’ deliberation must be constrained by the goal of bettering humanity. Should it happen that the agents can better humanity better by not acting selfishly—or should it happen that their original plan of bettering humanity by acting selfishly cannot succeed for whatever reason—then the agents will reason so as to coordinate their behavior properly to benefit humanity, rather than acting selfishly.

The question, then, is which outcomes are constraining the agents’ deliberation. Now, Searle might object that we are unfairly assuming that the agents are actually aiming at bettering humanity. He might object that we can just as easily define the collective action to be each agent’s acting in accordance with her or his individual intentions. And, Searle might argue, if we define the collective action in this way, then we no longer have the resources necessary to draw a distinction between the collective outcome (the bettering of humanity) and a set of individual actions (acting selfishly).

But this objection would miss the mark. Searle needs to draw a distinction between a collective outcome and a set of individual actions. Imagine what would happen if we were to try to do away with this distinction: we would have to define the collective action to be each agent’s acting in accordance with her or his individual selfish intentions. So far, so good; this is what Searle seems to suggest. But, the question would be, what makes this a collective action at all? In order for a set of actions to be a collective action, the agents must be doing something together. If we define the collective action as nothing more than the agents’ each acting selfishly, without regard to the outcome produced, then the agents are not really doing anything together at all—they are merely acting individually.

Now, I am not trying to argue that in order for a set of actions to be a collective action, they must have some benefit for the group, or for humanity. The question is not who benefits, but is rather whether or not there is some goal to the collective action beyond merely the performance of some actions by some agents. In order for a set of actions to
be a collective action, they must aim at some collective goal. Now, this collective action might be merely a description of the actions to be performed by a set of agents. Such cases are surely on the margins of collective action, however. It would not be surprising for such an agent to be unsure as to whether or not she or he was actually engaged in a collective action, or whether she or he was engaged in an individual action. If \( X \) and \( Y \) are engaged in a collective action where the goal is merely for \( X \) to do action \( x \) and, independently, for \( Y \) to do action \( y \), then it would not be surprising if either \( X \) or \( Y \) were unsure as to whether they were really engaged in a collective action while doing \( x \) or \( y \).

### 6.4 Contrastive examples

Let us, then, summarize where we stand. Searle’s challenge is to describe the difference between a set of agents bringing about an outcome through individual reasoning, and the same set of agents bringing about the (superficially) same outcome through collective reasoning. After all, Searle argued, the same outcome seems capable of being produced either through a set of individual actions, or through genuinely collective action.

Our answer is that the difference between an outcome produced through collective action and the same outcome produced through individual actions lies in the kind of reasoning employed to produce the outcome—when the agents’ actions are justified by individual reasoning, then the outcome is produced through individual actions, whereas when the agents’ actions are justified by collective reasoning, then it is a collective action.

An intuitive way of defining collective action is of a set of agents each doing their part in a collective activity. As we saw in Chapter 2, this is Searle’s strategy, and Tuomela’s strategy; it is also, in a less direct way, the strategy employed by Bacharach, Sugden, and Gold. The approach we offer here is similar in spirit; agents engaged in collective action are ‘doing their part’ of the collective action, because the collective action is circumscribed by the collective goal. The collective goal then ‘picks out’ each individual agent’s action by constraining her or his deliberation to just those actions which might possibly promote one of the outcomes compatible with the collective goal. Agents who are collectively reasoning constrain themselves according to the collective goal and understand the other agents to be likewise constrained, because the collective constraints constitute the collective ‘plan’.

In Searle’s example, the selfish businessmen only constrain themselves according to their
individual goal of maximizing their self-interest; not only is the outcome of everyone benefit-
ting an expected consequence of the agent’s intention rather than the content of the agent’s intention, but the agent’s intention is an individual intention, in that it does not constrain any of the other selfish businessmen. Each businessman knows that her or his intention to maximize self-interest has no authority to constrain the other businessmen, and thus that any constraints on their behavior will come from their own goals and mental attitudes. The cooperative businessmen, on the other hand, are engaged in a collective action; they have a collective goal, namely to benefit everyone, and their actions are thus constrained so as to produce an outcome compatible with that collective goal. Each cooperative businessman accepts such constraints on her or his reasoning, and each cooperative businessman also understands the other cooperative businessmen to be likewise constrained. The outcome that the cooperative businessmen produce might well be the same as that produced by the selfish businessmen, down to identical actions performed. The difference, however, is that the selfish businessmen act in accordance with their individual intentions, whereas the cooperative businessmen act in accordance with the collective goal.

In what way does my response differ from the approaches considered already? The team preference approach of Bacharach, Sugden, and Gold depended on two central features: team preference functions, which captured both the values of the team and the team’s goal, and team reasoning, which enabled agents to arrive through practical reasoning at an individual action to perform in pursuit of the team goal. Like the team preference approach, I argue that agents face a subjective decision problem and use a value function to determine which action they should perform. One significant difference between the approach I advocate here and the team preference approach is that I argue that collective goals ought be determined independently of each agent’s value function. This allows for each agent to have her or his own individual judgment about the value of various outcomes under consideration, while still enabling the team to hold a common goal independent of those individual value functions. Just as an individual agent’s goals are not determined by her or his value function or preferences, neither are the goals of a collective determined by the value functions of the agents in the collective; in the account presented here, goals float independently of individual value functions.

In order to make this difference clearer, consider the example of two agents, X and Y, engaged in a collective action of painting a house together. X and Y have different preferences for the group; X would prefer that the house be painted blue, while Y prefers
that the house be painted red. Can they engage in collective action in this circumstance? According to the Bacharach, Sugden, and Gold approach, they cannot; the fact that they have different team preferences means that they cannot engage in collective action. On our theory, however, the agents can engage in collective action, so long as they agree that they have reason to bring about some outcomes, and no reason to bring about other outcomes. Suppose, then, that X and Y adopt as a collective goal painting the house yellow. They can do this, I argue, even though neither agent has as a preference painting the house yellow; they can do this because team goals are not reducible to preferences, whether those preferences are individual or team. Bacharach, Sugden, and Gold, by contrast, do not have the resources to do this.

Another significant difference between my proposed account and team preference functions is that my team goal does not entail one unique action for each participant in the collective action. For the team preference approach, team reasoning requires that agents play that action which is their part of the team goal; this requires that there be one unique outcome which is the team’s goal. On the account I propose, there might be a set of outcomes which equally satisfy the team’s goal, just as an individual agent’s intention might be satisfied by a range of outcomes. When there are multiple outcomes compatible with a team’s goal, and when the agents of the team are not able to communicate with one another, failures of collective agency are possible. On the team preference approach, it is difficult to see how agents might fail to successfully engage in collective action once they possess a team goal; this result is undesirable.

To see this, imagine that X and Y have a collective goal of seeing an action movie together. Suppose further that there are three action movies playing in the movie theater. Can we make sense of this? On my theory of collective action, we can—the collective goal is to see an action movie, and so the agents’ deliberation is rationally restricted to bringing about one of those three outcomes. By itself, this collective goal does not guarantee successful coordination; the agents must still find a way to coordinate on the right action movie. (This is likely to be based on their individual preferences, though it might also be based on other information, such as, for instance, the salience of the movies.)

As for we-action theories of collective action, such as proposed by Bratman, my proposed account agrees that individually held intentions with content about the group activity are crucial to understanding collective action. Bratman’s account of those intentions was a positive account, and so it violates compositionality; for Bratman, each agent’s intention is
an intention that the group (positively) perform some collective action $J$. This collective action is a composition of individual actions, however, and so Bratman’s theory amounts to the claim that each agent intend a set of individual actions. The proposed theory avoids this difficulty by characterizing intentions negatively and not positively. Intentions, or goals, are constraints on future deliberation; they have no positive role in causing or controlling present-based behavior. As such, an agent’s intention to produce some outcome might restrict that agent’s future deliberation, and it might reference actions of other agents, but intentions which reference actions of other agents need not play any role in the agent’s individual deliberation. We thus have no violation of any plausible own-action condition, which holds that agents can only deliberate over their own actions and outcomes which can be produced by their actions.

Our disagreements with Bratman’s theory, then, are not of the kind that are made clearer through appeal to examples. This is understandable; the theory we presented here is closest in spirit to Bratman’s, in that we both present a we-action theory of collective action. It is not that Bratman cannot account for certain instances of collective action, then; rather, it is that Bratman’s theoretical account of intentions with distinct we-actions as their content is untenable as it stands.

We-intention theories of collective action, such as those proposed by Searle or by Tuomela, argue that collective action is best analyzed in terms of intentions to participate in the collective action, but they argue that collective intentions are distinct from individual intentions. This much my theory agrees with; in one sense, collective goals are different from, and irreducible to, individual goals. For us, however, collective goals are not reducible to individual goals because the content is necessarily different: collective goals are necessarily outcome-intentions, while individual goals are not. Thus, collective goals make reference to the actions of a group of agents, whereas individual goals make reference only to the actions of an individual. Also, intentions are defined by their role in constraining deliberation, and collective goals are phenomenologically indistinct from individual goals in that they both only constrain the deliberation of the possessing agent. There is thus a phenomenological continuum from deliberation on individual goals to deliberation on collective goals, which accounts for the fact that the boundaries between the two might be vague—a fact that we-intention theories of collective action have difficulty in accounting for.

A further difficulty for we-intention theories of collective action is in accounting for the derivation of individual actions from collective intentions. Both individual and collective
intentions constrain an agent’s deliberation and action in the same way; the puzzle for we-intention theories is why ontologically distinct mental attitudes such as I-intentions and we-intentions should function identically in deliberation and action. On the account I propose, there is no puzzle; collective goals restrain an agent’s deliberation in the exact same way as individual goals, because they are both goals—they both constrain an agent’s deliberation such that the deliberation is compatible with a set of goal outcomes. The difference between individual goals and collective goals lies solely in the set of actions they restrict, but, as we have seen, this difference is imperceptible from the standpoint of the individual deliberating agent. Individual goals and collective goals are two faces of the same coin; while there are differences, they are at bottom the same mental phenomenon with the same functional role, and it is therefore not surprising that they function the same way and that they are, in certain cases, phenomenologically similar to one another.

To see these two difficulties, consider again a case in which two agents X and Y are engaged in collective action of attacking an enemy fortification. They have two route they can take to try to attack the enemy; they can take the desert route, or they can take the jungle route. They are separated from one another, however, and so they cannot directly communicate with each other or coordinate their actions. Both X and Y engage in deliberation about what to do so as to best satisfy the collective goal of attacking the enemy fortification; they are deliberating about how to further their collective action. Now, the collective intention that they each have does not yet specify the action that they are to perform; they each need to engage in further deliberation to acquire that content. According to the we-intention theories, however, this further deliberation is difficult to account for. Both Searle and Tuomela require that the entire collective action be specified, down to the individual contributions of each of the participants, in order for the collective action to occur. For Searle, this is because of the part-whole relationship that individual intentions have to collective intentions; there can be no collective intention without the accompanying individual intention. For Tuomela, this is because collective intentions just are plans which specify parts for each of the participants.

Finally, my proposed theory is different from plural subject theories in that it is fundamentally a reductive theory; it appeals solely to individual agents engaged in deliberation about which actions they are going to individually perform. Sometimes these individual actions are taken in pursuit of an individual goal, whereas for other times the actions are taken in pursuit of collective goals; this nevertheless, the theory still appeals only to individual
agents and individual processes of deliberation. Understanding collective action and collective deliberation, on my account, amounts to understanding the behavior of individuals. Plural subject theories appealed to non-reductive plural subjects to account for collective action; as we argued above, such theories are only plausible as accounts of collective action if we do not have any more reductive theories on hand. I have attempted here to give such a plausible theory.
Appendix A

A formal model of intentions in rational choice

A.1 Games with and without intentions

Let us first draw a distinction between strategic games without intentions—the kind traditionally modeled in game theory—and strategic games with intentions. Strategic games with intentions are the model of deliberation we developed in Chapter 4 and Chapter 5. We present first a model of strategic games without intentions, in order to better highlight the differences.

A game $\Gamma_{a-\Phi}$ is an strategic game without intentions when $\Gamma_{a-\Phi}$ is an $n$-tuple:

$$\Gamma_{a-\Phi} = \langle N, (A_i), O, \pi, (\succeq_i) \rangle$$

where:

- $N$ is a finite set of $n$ agents;
- $A_i$ is a finite set of actions $\{\alpha_{i1}, \ldots, \alpha_{ik}\}$ for each agent $i \in N$ (we will speak generically of agent $i$’s action as $\alpha_i$);
- $O$ is a set of outcomes;
- $\pi : (A_i)^n \rightarrow O$ is an outcome function which maps vectors of actions onto the set of outcomes; and
• $\succeq_i$ is a preference relation which is reflexive, transitive, and well-defined. In other words, for all outcomes $o_1, o_2, o_3 \in O$ and for every agent $i \in N$,

- (reflexivity) $o_1 \succeq_i o_1$;
- (transitivity) if $o_1 \succeq_i o_2$ and $o_2 \succeq_i o_3$ then $o_1 \succeq_i o_3$; and
- (well-definedness) either $o_1 \succeq_i o_2$ or $o_2 \succeq_i o_1$.

($i$ can be indifferent between two outcomes, such that $o_1 \succeq_i o_2$ and $o_2 \succeq_i o_1$.)

Our strategic game without intentions is a traditional rational choice theory; outcomes are generated by the actions performed by agents in the decision problem, and each agent has an individual preference function according to which she or he evaluates the desirability of the outcome produced.

We can distinguish strategic games without intentions from strategic games with intentions. A game $\Gamma_a$ is a strategic game with intentions when $\Gamma_a$ is an $n$-tuple:

$$\Gamma_a = \langle N, (A_i), O, \pi, (\phi_i), (\succeq_i) \rangle$$

where recurring elements are as above, and:

- $\phi_i = \{\phi_{i_1}, \ldots, \phi_{i_k}\}$, where $\phi_{i_j} \subseteq O$ for $1 \leq j \leq k$;
- $\phi_i \subseteq \mathcal{P}(O)$ is $i$’s intention set over the power set of outcomes $O$.

Each element $\phi_{i_j}$ of agent $i$’s intention set $\phi_i$ can be thought of as a single intention to bring about either some unique outcome or some member of a set of outcomes. Suppose $X$ faces a decision problem in which she is deliberating between watching Scary Movie, watching Funny Movie, watching Dramatic Movie, staying home and reading a Funny Book, or staying home and Cooking. $X$ knows she can only do one thing. The set of outcomes for $X$’s decision problem is $O = \{\text{SM, FM, DM, FB, C}\}$. (In this decision problem, the agent’s available actions are directly and transparently connected to their respective outcomes.) We can distinguish between two different intention sets that $X$ might hold. For the first, let us say that $X$ holds two intentions during her deliberation: she intends to stay home, and she intends to do something fun. The intention to stay home would be represented as $\phi_{X_1} = \{\text{FB, C}\}$, while the intention to do something fun would be represented as $\phi_{X_2} = \{\text{FM, FB}\}$. $X$’s overall intention set, then, would be $\phi_X = \{\phi_{X_1}, \phi_{X_2}\}$. 
The second intention set that $X$ might hold takes $X$’s intentions to be more specified. Instead of holding two intentions (i.e. having an intention set with two members), $X$ might hold only one intention: the intention to read a funny book. This intention would be represented as $\phi'_{X_1} = \{FB\}$. $X$’s intention set here would be $\phi'_{X} = \{\phi'_{X_1}\}$. As it turns out, the two intention sets $\phi_{X}$ and $\phi'_{X}$ are practically equivalent; I will explain and justify this claim below. Here, we note merely that an agent’s outcome intention can be more or less specified, and that an agent can hold any number of intentions in her or his intention set.

### A.2 Outcome intentions and action intentions

For simplicity, we define intentions over the set of outcomes; thus, we can speak of agent $X$’s intention to bring about outcome $o$. We might equally have allowed intentions to also be defined over actions, to facilitate our speaking of $X$’s intention to perform some action $\alpha$. Doing so is not necessary, however. If we want to refer to $X$’s intention to perform some action $\alpha_X$, we can instead refer to $X$’s intention to bring about any outcome which is produced by her doing $\alpha_X$. Let us indicate by

$$\phi^\alpha_i(\alpha_i_z)$$

agent $i$’s intention to perform action $\alpha_{i_z}$. The following two claims are then equivalent:

- $\phi^\alpha_i(\alpha_i_z)$; and
- $\phi_i = \{o : \pi(\alpha_i_z, \alpha_{-i}) \text{ for all } \alpha_{-i} \in (A_{-i})\}$.

Thus, although we could build action intentions directly into the formal model, we need not do so; for our purposes here, we will take outcome intentions to be primitive, and action intentions to be derivative.

### A.3 Overall intentions and practical equivalence

Speaking of intention sets will prove unwieldy. First, the elements of an intention set are themselves sets of outcomes (i.e. individual intentions), rather than outcomes. And second, as we saw above, talking of intention sets requires us to draw a distinction between having

---

1Some theorists do take action intentions to be primitive; see, for example, Roy 2008.
two intentions, one to stay home and one to do something fun, and having only one intention, namely the intention do something fun at home.

To address this problem, let us define an overall intention. We will use the capital letter $\Phi$ to indicate overall intentions, and the lowercase letter $\phi$ to indicate intention sets. If agent $i$ has an intention set $\phi_i = \{\phi_{i1}, \ldots, \phi_{in}\}$, then we can say that $i$ has an overall intention $\Phi_i$ such that:

$$\Phi_i = \bigcap_{j=1}^{n} \phi_{ij}.$$ 

Overall intentions pick out those outcomes which are compatible with each of an agent’s intentions; if $\Phi_i = \emptyset$, then $i$ has no actions available which are consistent with her or his intention set.

We can now define practical equivalence. We can say that two intention sets $\phi_i$ and $\phi_i'$ are practically equivalent so long as their corresponding overall intentions are the same—that is, $\Phi_i = \Phi_i'$.

The other benefit to using overall intentions is that we can easily distinguish which outcomes belong to the content of an agent’s intention set. If some outcome $o_1$ belongs to agent $X$’s overall intention, then bringing about $o_1$ is compatible with $X$’s intention set. (This is not to say that $X$ intends to bring about $o_1$; this is a stronger claim, and is true only if $o_1$ is the only element of $X$’s overall intention.)

For these reasons, we will generally refer to an agent’s overall intention, rather than to her or his intention set. If we have particular reason to talk about an agent’s intention set or the individual intentions that constitute it—if, for example, we want to detail the various structural norms that adhere to an agent’s intentions and beliefs—then we will refer back to the agent’s intention set.

A.4 Intention rationality and preference rationality

Given the definition of our games with intentions above, we can give a more precise definition of the two types of rationality mentioned earlier: intention rationality and preference rationality.

We will say that an agent is intention rational if and only if the agent decides, in the face of a decision problem, to perform an action which is compatible with her or his intention set; an action is compatible with an agent’s intention set so long as the agent believes it will
(or can) lead to an outcome which belongs to her or his overall intention. Let us say that an agent’s action is strongly intention rational when playing that action will definitely lead to an outcome which belongs to the agent’s overall intention. That is, an agent $i$’s action $\alpha_i$ is strongly intention rational so long as:

$$\pi(\alpha_i, \alpha_{-i}) \in \Phi_i$$

for all $\alpha_{-i} \in (A_{-i})$.\(^2\)

An agent’s action is weakly intention rational if and only if the agent decides, in the face of a decision problem, to perform an action which can contribute to an outcome which belongs to her or his overall intention, but which might fail to bring about such an outcome according to the actions of other agents in the decision problem.\(^3\) Or, to put it another way, an agent’s action is weakly intention rational so long as it is not strongly intention-irrational. Thus, an agent $i$’s action $\alpha_i$ is weakly intention rational so long as:

$$\pi(\alpha_i, \alpha_{-i}) \in \Phi_i$$

for some $\alpha_{-i} \in (A_{-i})$.

Actions can be more or less weakly intention rational, depending on such factors as the performing agent’s beliefs about the probable actions of other agents, or on the number of available actions for the other agents which produce intention-compatible outcomes.

Accordingly, an action which is not intention rational is one which is neither strongly intention rational nor weakly intention rational. Such an action is guaranteed to produce an outcome which falls outside of the agent’s intention set.

Intention rationality, whether strong or weak, is distinguishable from preference rationality. Intuitively, preference rationality is to be equivalent to the notion of rationality employed by traditional rational choice theories: an agent is rational when she or he chooses that action which leads to the most preferable outcome possible given her or his beliefs about the likely actions of other agents in the game, given her or his knowledge about the structure of the game, and so forth. The difficulty in formalizing preference rationality stems

\(^2\)We will speak interchangeably of an action’s being strongly or weakly intention rational, or of an agent’s being strongly or weakly intention rational. An intention rational agent is one who decides, in the face of a decision problem, to play an intention rational action (whether strong or weak, accordingly).

\(^3\)We also might want to model situations in which an agent’s action is not guaranteed to produce an intention-compatible outcome because of chance effects or uncertainty in the action itself, as for example with throwing darts at a bulls-eye or playing a lottery. We might want to say that such actions are weakly intention rational, in that they are actions which are compatible with producing an intended outcome, even though they do not guarantee such an outcome. In the formalism presented, we do not explicitly model aleatoric or uncertain connections between actions and outcomes.
from the fact that there is no one clear way to formalize this notion. For single-agent choice
situations, there is no problem; the deliberative agent uses her or his beliefs about which
actions lead to which outcomes in order to choose which action to perform. In multi-agent
choice situations, however, the problem becomes more complicated; each agent must choose
an action based not only on which outcomes that action might contribute to, but also on
the basis of which actions the other agents are likely to perform.

Game theorists employ various *solution concepts* in order to define rationality (for us,
preference rationality) in multi-agent decision problems. The most well-known solution
concept is that of the Nash equilibrium, which defines a rational outcome to be one in
which every agent’s action is a best response to every other agent’s action in the choice
situation—that is, an outcome in which each agent’s action is preference-maximizing against
the actions of the others. A rational action, accordingly, is an action played towards an
equilibrium. Formally, in a strategic game with intentions $\Gamma = \langle N, (A_i), O, \pi, (\phi_i), (\succ_i) \rangle$, an
action profile $\alpha^* \in A$ is a (pure-strategy) equilibrium so long as:

$$\pi(\alpha_i^*, \alpha_{-i}^*) \succ_i \pi(\alpha_i', \alpha_{-i}^*)$$

for all $\alpha_i' \in A_i$ and for all $i \in N$.

If there is only one equilibrium $o^* = \pi(\alpha^*)$, then agent $i$ is preference rational if she or he
chooses to play her or his component action $\alpha_i^*$ of the action profile. If there is more than
one equilibrium, determining an agent’s rational action is more difficult; each agent needs
to appeal to an equilibrium-selection program which picks out one unique equilibrium as
the one to work towards. This is not a trivial project; more will be said on this later.

There are other solution concepts besides that of the equilibrium, however; we might
take an action to be rational so long as it is *perceived* to be preference-maximizing given
an agent’s *beliefs* about the likely actions of the other agents in the game. (According to
Osborne and Rubinstein, such an action is called *rationalizable*; we will call such actions
*belief-maximizing*.\footnote{Osborne and Rubinstein 1994, see chapter 4.}) Or, we might take an action to be rational so long as it survives culling
from repeated iterations of the elimination of strongly- or weakly-dominated strategies.
(One strategy is *strongly dominated* by another when it always yields an inferior outcome
regardless of the actions of the other agents in the game; one strategy is *weakly dominated*
by another when it never provides a superior outcome regardless of the actions of the other
agents in the game—although it might generate an outcome about which, when compared
to the weakly-dominating strategy, the agent is indifferent.)

All of these solution concepts are meant to capture the concept of rational behavior in a multi-agent decision problem, and ‘rational behavior’ here is meant to track purposeful actions towards an optimal outcome. Each of these solution concepts has its advantages and disadvantages. We will generally use the equilibrium as our solution concept; it has the advantage of not requiring that agents have detailed beliefs about the likely behavior of other agents prior to deliberation, and it is well-suited to tracking preference-maximizing behavior over repeated trials of a game, or in large populations. We could also employ belief-maximization as our solution concept; this would be appropriate especially when agents explicitly have beliefs about the likely behavior of other agents prior to deliberation. And, though we should eschew the kind of elaborate equilibrium-selection program espoused by e.g. Harsanyi and Selten, we can accept that there are principles of equilibrium-selection which are clearly rational for agents to employ; these might involve, for instance, dominance reasoning.\footnote{Harsanyi and Selten 1988.}

The preference-rational agent, then, acts so as to bring about her or his most highly preferred outcome possible, while the intention-rational agent acts in accordance with her or his intentions. These are two differing notions of rationality; the preference-rational agent is not necessarily an intention-rational agent, nor vice-versa. When we introduce intentions into our game, however, it is possible for an agent who is both intention-rational and preference-rational to bring about a more preferred outcome than can the agent who is merely preference-rational.\footnote{We will skip over a detailed discussion of such games here. Generally, however, the games tend to fall into one of three categories: they might be games with multiple equilibria, in which one agent can maximize the value of the outcome produced by forming an intention to act towards one equilibrium rather than another; they might be games with one equilibrium but which feature Pareto-superior subgames, such that if the agent intends to not act towards the equilibrium, the agents are all guaranteed of a better outcome (though there might not be any remaining equilibria in the game); and they might be games where deliberation is costly, such that forming an intention which reduces the number of outcomes under consideration has the effect of ensuring that a more desirable outcome is brought about.}
References


REFERENCES


REFERENCES


